Predictivist Bayes density estimation

P. Richard Hahn

Abstract: This paper develops a novel computational approach for Bayesian density estimation, using a kernel density representation of the Bayesian predictive distribution. The new method is based directly on posterior predictive distributions, bypassing the usual likelihood-prior representation. The predictivist approach permits parallel posterior computation and leads to tractable analysis of posterior convergence properties.

1. Introduction

This paper develops a new computational approach for Bayesian density estimation as a first step towards developing similar techniques for other, less tractable, nonparametric Bayesian models. The new method is naively parallelizable, relying on Monte Carlo techniques, but not Markov chains. An analytic approximation to the posterior mean density is available (e.g., no Monte Carlo is necessary) and bounds on the rate of posterior concentration are derived as a function of sample size.

The theory and computation of nonparametric Bayesian density estimation with Dirichlet processes are by now well established, see Ferguson [1973], Sethuraman [1994], Escobar and West [1995], and Walker et al. [1999]. Neal [2000] summarizes various algorithms for sampling from Dirichlet process mixtures. Other approaches to nonparametric Bayesian density estimation have been studied, for example those based on Polya trees [Lavine, 1992, 1994, Paddock et al., 2003] or Gaussian processes [Murray et al., 2009], but Dirichlet processes remain the most popular model for density estimation in applied settings.

In this paper, a fundamentally different approach will be taken, based on the idea of predictive inference; key references include de Finetti [1974, 1975] and Lauritzen [1988]. Exchangeability and contractability are key concepts in this approach; a comprehensive abstract treatment can be found in Kallen-
berg [2005]. In the specific density estimation model proposed here, predictive
distributions will be represented using kernel density estimates; this idea was
previously explored in West [1991]. We are able to extend the idea beyond West
[1991] by focusing on finite rather than infinite exchangeability; in this there are
correspondences between this work and work on Bayesian methods for fi-
nite population sampling [Ghosh and Meeden, 1997], although here the finite
population is simply a computational device.

2. Predictivist Bayes

This section of the paper considers how posterior inference can be conducted
natively via predictive distributions, from model specification through posterior
sampling. First, a correspondence is established between posterior predictive
distributions and posterior distributions. This relationship provides a way to
sample from the posterior by sampling directly from a joint posterior predictive
distribution. Second, a correspondence is established between likelihood func-
tions and limiting posterior predictive distributions. Both derivations follow
from the definition of the prior predictive distribution:

\[ p(x_{1:N}) = p(x_1, \ldots, x_N) = \int \prod_{i=1}^{N} f(x_i | \theta) \pi(\theta) d\theta. \]

Through these definitions, real-valued parameters with measures admitting density functions
with respect to the Lebesgue measure will be assumed.

The purpose of this section is to emphasize, for readers familiar with the
“likelihood-prior” specification of Bayesian models, the equivalence between this
now-standard approach and exchangeable joint distributions. In more detail, the
foundational work of de Finetti de Finetti [1974, 1975] established that any in-
finitely exchangeable distribution can be expressed as a mixture of independent
and identically distributed random variables; see also Hewitt and Savage [1955].
That is, if the joint distribution over observables (the left hand side) is infinitely
exchangeable, then there exists a likelihood-prior representation (the right-hand side). These days, Bayesian applied statistics typically starts from a direct specification of the likelihood-prior form, from which infinite exchangeability follows. Although the likelihood-prior representation (the right-hand side) is often convenient from a model specification point of view, establishing a constructive relationship between the likelihood $f(x \mid \theta)$ and prior $\pi(\theta)$ and the predictive distribution (left-hand side) will form the basis of the computational approach developed in the following sections.

2.1. **Predictivist posterior sampling**

Let $\hat{\theta}_N := \hat{\theta}_N(X_{1:N})$ be a consistent estimator when applied to i.i.d. samples from $f(x \mid \theta)$. Then, for any $\epsilon > 0$, iterated expectation gives the expression

$$
\Pr(\left| \hat{\theta}_N - \theta \right| < \epsilon) = \int \Pr(\left| \hat{\theta}_N - \theta \right| < \epsilon \mid \theta) \pi(\theta) d\theta.
$$

(2)

where both $\hat{\theta}_N$ and $\theta$ are random variables. By the assumption of consistency and the dominated convergence theorem, the right-hand side converges to 1 as $N \to \infty$, showing that the random variable $\hat{\theta}_N$ converges in probability to a random variable with distribution $\pi$. In other words, the prior distribution may be conceptualized as the limiting distribution of $\hat{\theta}_N$ under a specified prior predictive distribution.

An analogous interpretation of the posterior distribution is available, by fixing the first $n$ elements of $X_{1:N}$. Specifically, for any $\epsilon > 0$, iterated expectation gives

$$
\Pr(\left| \hat{\theta}_N - \theta \right| < \epsilon \mid x_{1:n}) = \int \Pr(\left| \hat{\theta}_N - \theta \right| < \epsilon \mid \theta, x_{1:n}) \pi(\theta \mid x_{1:n}) d\theta.
$$

(3)

Accordingly, the posterior distribution may be conceptualized as the limiting distribution of $\hat{\theta}_N$ under the posterior predictive. This correspondence motivates us to develop a posterior sampling procedure based on drawing samples directly from $p(x_{n+1}, \ldots, x_N \mid x_{1:n})$. 
Algorithm 1: Posterior sampling via predictive distributions

For $M$ and $N$ large integers.
For $j = 1, \ldots, M$,
1. Draw $x^{(j)}_{n+1}, \ldots, x^{(j)}_N$ from $p(x_{n+1}, \ldots, x_N \mid x_{1:n})$.
2. Define $x_{1:N}^{(j)} = (x_1, \ldots, x_n, x_{n+1}^{(j)}, \ldots, x_N^{(j)})$ and set $\theta^{(j)} = \hat{\theta}_N^{(j)}$.

Note that each sample $\theta^{(j)}$ depends on a Monte Carlo approximation, based on a sample of size $N$. In practice, it is often convenient to sample from $p(x_{n+1}, \ldots, x_N \mid x_{1:n})$ using the compositional form

$$p(x_{n+1} \mid x_{1:n})p(x_{n+2} \mid x_{1:(n+1)}) \cdots p(x_N \mid x_{1:(N-1)}). \quad (4)$$

The following two examples illustrate the simplicity and viability of this approach in cases where the sequence of predictives (4) is known.

Example: Bernoulli likelihood. Suppose $X_i \sim \text{Bernoulli}(\theta)$ with prior $\theta \sim \text{Beta}(\alpha, \beta)$. Integrating over this prior yields the following (conjugate) predictive updates

$$p_t(x_{t+1} \mid x_{1:t}) = \text{Bernoulli}\left(\frac{\alpha_t}{\alpha_t + \beta_t}\right), \quad \alpha_t = \alpha_{t-1} + x_t, \quad \beta_t = \beta_{t-1} + 1 - x_t. \quad (5)$$

Let $\alpha_0 = \beta_0 = 1$ and suppose $n = 10$ observations are observed, and that seven of them are ones: $\sum_{i=1}^n x_i = 7$. The left panel of Figure 1 shows simulated predictive sequences 1000 steps into the “future” from the prior and from the posterior; the right panel shows that repeating this exercise 5000 times nicely recapitulates the known Beta(8, 4) posterior distribution.

\footnote{A Julia notebook is available covering all examples: \url{https://nbviewer.jupyter.org/github/danmerl/PFBDE.jl/tree/master/}.}
Fig. 1. Left panel: Gray lines depict 20 simulated data sequences from the prior predictive; they terminate 1000 steps in the future at points that are uniformly distributed in the interval. Solid lines show 20 simulated data sequences beginning from the point \( n = 10 \) with an observed sample average of 0.7; restricting to sequences that run through the point (10, 0.7) yields sample paths that terminate in a more concentrated region. Right panel: The histogram at \( t = 1000 \) for 5000 simulated posterior predictive data sequences for \( n = 10, \bar{x}_n = 0.7 \); it nicely recapitulates the known \( \text{Beta}(8, 4) \) posterior distribution for \( \theta \), which is shown overlaid in black.

Example: Gaussian likelihood with known variance. Suppose \( X_i \sim N(\theta, 1) \) with prior \( \theta \sim N(\mu_0, \phi_0^{-1}) \). Integrating over this prior yields the following (conjugate) predictive updates

\[
p_t(x_{t+1} | x_{1:t}) = N(\mu_t, 1 + 1/\phi_t), \quad \mu_t = \frac{x_t + \mu_{t-1}\phi_{t-1}}{1 + \phi_{t-1}}, \quad \phi_t = 1 + \phi_{t-1}. \tag{6}
\]

Forward simulation yields (approximate) posterior distributions over \( \theta := \hat{\theta}_N = \bar{x}_N \), as in the Bernoulli example above and similarly recapitulates, as expected, the usual Bayesian posterior.

2.2. Model specification by enforced sequential coherence

For many models, the prior predictive is not available in a convenient form. In this section, a novel procedure is presented, which allows specifying the prior predictive distribution without reference to a prior (posterior) over parameters. Consider dividing expression (1) by \( p(x_{2:N}) \):
\[ p(x_1, \ldots, x_N) \]  
\[ p(x_1 | x_2: N) = \int f(x_1 | \theta) \pi(\theta | x_2: N) d\theta, \]

(7)

where the second line follows from the definition of Bayes rule. Assuming (weak) posterior consistency about a fixed \( \theta^* \), it follows that

\[ \lim_{N \to \infty} p(x_1 | x_2: N) = \lim_{N \to \infty} \int f(x_1 | \theta) \pi(\theta | x_2: N) d\theta = f(x_1 | \theta^*), \]

(8)

showing that the likelihood function may be conceptualized as the limiting predictive distribution of the learning (conditioning) process. More suggestively (writing \( x \) for \( x_1 \) and \( \theta \) for \( \theta^* \)):

\[ \lim_{N \to \infty} p(x | x_1: N) = f(x | \theta). \]

(9)

Focusing on the left-hand side motivates us to think about model specification in terms of density estimators [West, 1991], approximating (9) by taking \( N \) to be large, but finite. That is, we can specify a nonparametric model simply by letting the left-hand side be a nonparametric density estimator. Practically, this will entail allowing \( \hat{\theta}_N \) to be \( N \)-dimensional.

While specifying (9) establishes a “likelihood”, the model is yet incomplete. As was seen in section 2.1, it is the full prior (posterior) predictive over \((X_1, X_2, \ldots)\) that yields samples from the prior (posterior). Equivalently, the full sequence of predictives, as in (4), needs to be specified. In the following section it is shown that the full sequence can be derived by imposing an additional, intuitive symmetry condition.

### 2.2.1. Sequential coherence

Any independent and identically distributed Bayesian model — e.g., models of the form given by the right-hand side of (1) — has an *exchangeable* predictive
distribution, meaning that the distribution is invariant to permutation (shuffling) of the indices:

\[(X_{k_1}, X_{k_2}, \ldots, X_{k_N}) \overset{d}{=} (X_1, X_2, \ldots, X_N)\]  

(10)

for any permutation \((k_1, k_2, \ldots, k_N)\) of integers \(1, \ldots, N\). Thus, to specify the statistical model directly in terms of the predictive, it is necessary that exchangeability hold, while preserving the chosen form of (9). Here, our strategy for doing so will be based on a weaker notion called sequential coherence. Sequential coherence stipulates a certain relationship between subsequent and previous predictive distributions — informally, one’s expected predictive density tomorrow is one’s predictive density today: for all \(t < j \leq N\),

\[(X_{j+1} \mid X_{1:t}) \overset{d}{=} (X_j \mid X_{1:t}).\]  

(11)

In fact, it is straightforward to show that this condition implies (the seemingly stronger condition) that all subsequences of the same length have the same distribution, a condition called contractability:

\[(X_{k_1}, X_{k_2}, \ldots, X_{k_t}) \overset{d}{=} (X_1, X_2, \ldots, X_t)\]  

(12)

for any \(k_1 < k_2 < \cdots < k_t\) and \(t \leq N\).

Contractability and exchangeability are equivalent in the case of infinite sequences (see Theorem 1.1 of Kallenberg [2005]). For finite sequences, exchangeability implies contractability, but not vice versa. However, any finite contractable sequence can be used to generate an exchangeable sequence with equivalent subsequence distributions simply by independently shuffling its elements (lemma 1.11 of Kallenberg [2005]). In particular, the full sequence \((X_1, \ldots, X_{N+1})\) and the subsequence \((X_1, \ldots, X_N)\) will maintain their distributions, implying that \(X_{N+1} \mid X_{1:N}\) will have the desired distribution. Moreover, because for exchangeable sequences, any subset of the same size has the same distribution, it follows that the shuffled contractable distribution satisfies

\[X_j \mid X_{1:(N+1)\setminus j} \overset{d}{=} X_{N+1} \mid X_{1:N}\]  

(13)
for any \( j \leq N+1 \), as required. Therefore, it is sufficient to define an exchangeable model by 1) specifying \( p(x \mid x_{1:N}) \) and 2) generating the (reverse) sequence of predictive distributions to satisfy sequential coherence.

A sequence of predictives obtained in this way will not correspond to an exchangeable random sequence without explicit random permuting; i.e., a joint density evaluation will not be order-independent unless one performs a costly averaging over the permutations (cf. West [1991]). However, to simulate from the posterior according to the algorithm of section 2.1, one must simply (uniformly) randomly permute the draws before evaluating \( \hat{\theta}_N \). Consequently, in the remainder of the paper, no mention will be made of the exchangeability-inducing permutation step; the focus will be on the more critical imposition of sequential coherence (equivalently, contractability).

2.2.2. Model specification via enforced coherence

Operationally, sequential coherence implies an integral equation that one can try to solve. Define \( p_{t-1}(\cdot) := p(x_t \mid x_{1:(t-1)}) \) and \( p_{t-1}(x \mid y) := p(x_t \mid x_{1:(t-2)}, x_{t-1} = y) \). In this notation, sequential coherence stipulates that

\[
p_{t-1}(x) = \int p_t(x \mid y)p_{t-1}(y)dy.
\]

Therefore, our strategy for model determination will be to specify \( p_N(x) \) and then solve (14) iteratively until reaching \( p_{n+1}(x) \).

**Example: Bernoulli likelihood.** For \( \bar{x} = N^{-1} \sum_{i=1}^{N} x_i \), assume that \( X_{1:N} \sim \text{Bernoulli}(\bar{x}) \). Write \( \pi_t = \Pr(X_{t+1} = 1 \mid X_{1:t}), \pi_t(z) = \Pr(X_{t+1} = 1 \mid X_{1:(t-1)}, X_t = z) \) and \( \bar{x}_t = t^{-1} \sum_{i=1}^{t} x_i \). Plugging these definitions directly into (14) gives

\[
\pi_{N-1} = \pi_N(1)\pi_{N-1} + \pi_N(0)(1 - \pi_{N-1})
\]

\[
= \left( \frac{N - 1}{N} \bar{x}_{N-1} + \frac{1}{N} \right) \pi_{N-1} + \frac{(N - 1)}{N} \bar{x}_{N-1}(1 - \pi_{N-1}),
\]

\[
= \bar{x}_{N-1}.
\]
Iterating shows that the coherent predictive sequences use the current sample average at time $t$ as the prediction probability for observation $t+1$. Simulation from this sequence, as described in Section 2.1, yields a posterior distribution over $\theta_N := \bar{x}_N$. A similar derivation holds for the Gaussian example from the previous section.

Note that to exactly duplicate the Bayesian solution demonstrated in the previous section, one can “seed” the backward induction procedure with two pseudo-observations, one of which is a one and the other a zero. Indeed, observe that Algorithm 1 is still well-defined in the event that the observed data is augmented with pseudo-data, providing an intuitive mechanism for incorporating prior knowledge.

3. Predictivist Bayes density estimation

3.1. Enforcing coherence of a kernel density predictive sequence

To specify a Bayesian predictivist model for nonparametric density estimation, let $p_N(x) := p(X_{N+1} = x \mid x_{1:N})$ be given by a kernel density estimator \cite{Rosenblatt et al., 1956, Parzen, 1962, Silverman, 1986} of the form

$$K^\tau_N(x) = N^{-1} \sum_{i=1}^{N} \phi(x \mid x_i, \tau),$$

where $\phi(\cdot \mid \mu, \tau)$ is a normal density function with center $\mu$ and variance (bandwidth) $\tau$. This choice implies that the coherence equation can be written as

$$p_{N-1}(x) = \int K^\tau_N(x)p_{N-1}(y)dy. \tag{16}$$

We recognize this as an inhomogeneous Fredholm equation of the second kind; applying the method of successive substitution \cite{Arfken, 2013} gives a series expanded representation of $p_{N-1}(x)$ as

$$p_{N-1}(x) = \sum_{j=1}^{\infty} \frac{N-1}{N^j} K^\tau_N(x). \tag{17}$$
See the appendix for a step-by-step derivation.

To establish that this series converges we can leverage insights from the statistical context, by expressing (31) as an expectation

\[ p_{N-1}(x) = \mathbb{E} K_{N-1}^{\tau}(x), \]

where \( Z \sim \text{Geometric}(\rho) \) for \( \rho = \frac{N-1}{N} \). Because \( K_{N-1}^{\tau}(x) \) is bounded as a function of \( Z \) and \( Z \) is supported on the positive integers (which is closed), this expectation exists. Because each term in (31) is itself a kernel density estimator and this representation involves only summation and convolution, we can apply the same process to obtain a nested sum expression for each predictive distribution at any number of steps back from \( N \), \((N-2, N-3, \text{etc.})\) simply by applying the mappings \( N \rightarrow N-1 \) and \( \tau \rightarrow 2\tau \). Substitution and iteration yields

\[ p_{N-t}(x) = \sum_{j_1=1}^{\infty} \cdots \sum_{j_t=1}^{\infty} \frac{N-1}{N} \frac{N-2}{(N-1)} \cdots \frac{N-t}{(N-t)} K_{N-t}^{\tau\{\Pi_{j=1}^{\tau} j_h\}}(x). \] (18)

Again, this can be seen as a nested expectation of independent geometric random variables \( Z_h \) with parameters \( \rho_h = \frac{N-h}{N-h+1} \) for \( h = 1 \ldots t \):

\[ p_{N-t}(x) = \mathbb{E}_{1} \mathbb{E}_{2} \mathbb{E}_{3} \cdots \mathbb{E}_{t} K_{N-t}^{\tau\{\Pi_{h=1}^{\tau} Z_h\}}(x). \] (19)

Observe that \( K_{N-t}^{\tau\{\Pi_{h} Z_h\}}(x) \) depends on the \( Z_h \) variables only via their product. Defining

\[ \chi_t = Z_1 \times Z_2 \times \cdots Z_t, \] (20)

gives

\[ p_{N-t}(y) = \mathbb{E} K_{N-t}^{\tau\chi_t}(y) \] (21)

where the expectation is now over \( \chi_t \) for \( t \) between 1 and \( N-n \). This expectation exists by a similar argument as before. Therefore, the predictive density at time \( N-t \) is a scale mixture of normals, with mixing distribution given by a scaled (by \( \tau \)) product of independent geometric random variables (with parameters depending on \( N \) and \( t \)).

Note that in this model \( N \) and \( \tau \) are free parameters that determine the model. Fortunately, frequentist reasoning can help guide the choice of these pa-
Fig 2. Left panel: for $N = 1000$, $n = 50$ and $\tau = 0.04$, the implied kernel, marginally over $\chi_t$, is shown for $t = 1$ (dashed), $t = 400$ (dotted) and $t = 950$. At $t = N - n = 950$, the kernel is visually indistinguishable from a Gaussian kernel with variance 0.04 (solid). Right panel: the shifted log-normal mixing distribution becomes sharper as $t$ approaches $N - n$, collapsing to a near point-mass at $\tau = 0.04$ (shown in solid black). The dashed line shows the $t = 1$ one-step-ahead predictive diffuse mixing density for $n = 50$, $N = 1000$. The gray lines represent values of $t$ between 10 and 950 in increments of 50.

3.2. Computation

As a product of independent (but not identically distributed) geometric random variables, $\chi_t$ has no readily available closed form. However, a central limit theorem (in the log domain) suggests a log-normal approximation. First, note that because the $Z_h$ geometric variables are independent, the product of their expectations gives the expectation of their product. Accordingly, $\eta_t := \mathbb{E} \chi_t = \prod_{h=1}^{t} \rho_h^{-1}$ with $\rho_h = \frac{N-h}{N-h+1}$. Similarly, $VZ_h = \frac{(1-\rho_h)}{\rho_h^2}$, so $\mathbb{E} Z_h^2 = \frac{(2-\rho_h)}{\rho_h^2}$ and $\nu_t := \mathbb{V} \chi_t = \prod_{h=1}^{t} \frac{(2-\rho_h)}{\rho_h^2} - \prod_{h=1}^{t} \rho_h^{-2}$ by properties of variance. The log-normal approximation is improved by respecting the fact that $\chi_t \geq 1$. To that end,
consider a log-normal random variable $S_t$ with mean $\eta_t - 1$ and variance $\nu_t$, which has parameters

$$\mu_t = 2 \log (\eta_t - 1) - \frac{1}{2} \log ((\eta_t - 1)^2), \quad \sigma_t = \sqrt{\log (1 + \frac{\nu_t}{(\eta_t - 1)^2})},$$

and set $\chi_t = S_t + 1$. (Because the log-normal approximation of $\chi_t$ is inaccurate when $t$ is small, one may define the backward induction starting at $N + a$ for $a$ large enough that the log-normal central limit approximation safely holds, defining $N$ as the termination point for the forward simulation. Intuitively, this works because if $N$ was thought to be large enough, then $N + a$ also suffices as a Monte Carlo approximation of $\theta := \lim_{N \to \infty} \hat{\theta}_N$.) Figures 2 illustrates the impact on the implied kernel for various values of $t$; the marginal kernel densities shown in the left panel were computed by numerical integration.

To generate posterior samples (density functions) one must simulate from a kernel density distribution with a log-normal mixture of normal kernels, which can be done straightforwardly as follows.

---

**Algorithm 2: Sampling the coherent Gaussian KDE predictives**

At step $t$,

1. Select a location parameter $u$ at random among the previous $n + t - 1$ data points (of which $t - 1$ are simulated): $(x_1, \ldots, x_n, x_{n+1}^*, \ldots, x_{n+t-1}^*)$.

2. Next, draw a scale parameter $s_t$ from the log-normal distribution with parameters as in (22).

3. Finally, draw (pseudo-)observation $x_{n+t}^*$ from $N(u, \tau(s_t + 1))$.

---

Note that this forward simulation process yields independent samples from $p_N(x \mid x_{1:n}, x_{(n+1):N}^*)$, which may be obtained in parallel. Also, observe again that this algorithm is readily augmented with initial pseudo-data as a way to incorporate prior information.
3.3. Posterior concentration

Expressing the sequential coherence condition in terms of stochastic processes facilitates theoretical analysis. Consider the predictive density at time \( t > n \) as the stochastic process induced by the distribution over \( X_{(n+1):t} \) defined by the algorithm of section 3.2: \( \xi_t = p_t(x \mid x_{1:n}, X_{(n+1):t}) \). In this notation, sequential coherence stipulates that

\[
E(\xi_{t+1} \mid \xi_{n+1}, \ldots, \xi_t) = \xi_t,
\]

which is the definition of a martingale (indexing from \( n + 1 \)). One immediate implication of this definition is that the unconditional expectation is \( E(\xi_N \mid x_{1:n}) = \xi_n \). In other words, the posterior mean density is precisely the one-step-ahead predictive density \( p_n(x) \). In turn, this implies that the posterior mean density \( E\{p_N(x) \mid x_{1:n}\} \) can be expressed approximately as

\[
p_n(x) = n^{-1} \sum_{i=1}^{n} \int \phi(x \mid x_i, \tau(s+1)) \text{lognormal}(s \mid \mu_n, \sigma_n) ds,
\]

where \( \text{lognormal}(s \mid \mu_n, \sigma_n) \) denotes the density of a log-normal distribution with parameters as in (22). Therefore, the posterior mean is consistent as long as this kernel density estimate is consistent for the chosen \( \tau \).

To study the fluctuation of the posterior about this mean, one can apply the Hoeffding inequality Hoeffding [1963]. In particular, for any \( x \),

\[
|p_t(x) - p_{t+1}(x)| \leq \frac{\phi(0 \mid 0, \tau)}{t + 1},
\]

where \( \phi(\cdot \mid 0, \tau) \) denotes the normal density function, with mean 0 and variance \( \tau \), at a point. This bound is derived as follows. First, the highest a kernel density estimate based on \( t \) observations can be at a point occurs when all \( t \) observations are equal, in which case the density has height equal to the mode of the kernel. Second, a kernel density estimate (for any number of observations, at any point) is always greater than zero. Therefore, the most a kernel density estimate can
change in adding one additional data point, is to add an observation to a spot
that was infinitely far away from the previous identical $t$ observations, at which
point the density will go from zero to the maximum height of the kernel, divided
by $t + 1$. In the case of the scale mixtures of normal kernels derived above, this
mode is, at most, the height of the kernel at $t + 1 = N$, which is $\phi(0 \mid 0, \tau)$ by
construction.

Therefore, the Hoeffding inequality gives
\[
\Pr\{|p_N(x) - p_n(x)| \geq \epsilon\} \leq 2 \exp \left( \frac{-\epsilon^2}{2\epsilon^2 \sum_{j=n+1}^{N} (j + 1)^{-2}} \right).
\] (26)
Because the tail of a convergence series vanishes, the right-hand side goes to
zero as $n \to \infty$.

3.4. Illustrations

Synthetic example. For this demonstration, $n = 50$ observations are drawn from
a mixture of two Gaussians with equal weights:
\[
p(x) = \frac{1}{2} \phi(x \mid 2, 4) + \frac{1}{2} \phi(x \mid 10, 1).
\] (27)
The data is fit to a predictivist Bayes kernel density model with $N = 1000$ and
$\tau = 0.08$; the resulting point estimate and uncertainty bands are depicted in
the left panel of Figure 3. For comparison, the $R$ kernel density estimate with
bandwidth selection method $SJ$, as described in Sheather and Jones [1991], is
also shown.

Empirical example: the galaxy data. The “galaxy data” have been widely
used to exemplify Bayesian and non-Bayesian density estimation techniques.
The data are 82 velocity measurements (in km/second) of galaxies obtained
from an astronomical survey of the Corona Borealis region Roeder [1990]. No-
table Bayesian papers using this data include Carlin and Chib [1995], Escobar
and West [1995] and Bernardo [1999]. The right panel of Figure 3 depicts the
Fig 3. Top left: data are drawn from a mixture of two Gaussians, with $n = 50$. Three density estimates overlay the data histogram. Solid is the backward induced KDE with $N = 1000$ and $\tau = 0.08$; dashed is the true density; dotted is the R KDE with bandwidth select method S3. One-thousand draws from the posterior density are shown in gray. Top right: This plot is the same as the left panel except that now $n = 500$; the gray uncertainty bands are notably narrower. Bottom: the galaxy data of Roeder [1990] consists of $n = 82$ astronomical measurements. The posterior mean density is shown for the $N = 1000$ and $\tau = 0.08$ backward induced model (solid line). The dashed line depicts the default KDE in R.
posterior mean for the $N = 1000, \tau = 0.04$ model, along with one-thousand posterior draws to provide visual uncertainty bands.

4. Summary

This paper develops a Bayesian density estimation method by deriving a coherent sequence of kernel density estimators. Relative to traditional MCMC-based Bayesian methods, the main benefits are that

- posterior sampling is naively parallelizable,
- statistical properties (e.g., consistency) are inherited directly from the underlying kernel density estimator,
- pseudo-observations can be used to incorporate prior information in an intuitive way.

More generally, model specification and posterior computation in a predictivist mode is a promising avenue for further development. The most common way to specify a Bayesian model is to specify (a) a family of distributions and (b) a prior. This is equivalent to specifying (c) an exchangeable joint distribution for $X_1, X_2, \ldots$ This paper explores a distinct option: specifying (d) the predictive distribution $X_N|x_1, \ldots, x_{N-1}$ for just a single (large) value of $N$ and (e) declaring that sequential coherence holds. Using (d) and (e) rather than (a) and (b) has two potential benefits: 1) better control over posterior concentration and 2) a novel posterior sampling strategy that is naively parallelizable. One way to think about this new approach is that we are trading infinite coherence for finite coherence (for some large $N$) and by doing so are able to avoid Markov chain methods.

Ongoing work extends this approach to regression models for which existing Markov chain samplers are computationally inefficient.
References


Appendix A: Appendix

The key equation for establishing sequential coherence is

\[ p_{N-1}(x) = \int K_N^\tau(x)p_{N-1}(y)dy. \]  \hspace{1cm} (28)

To solve this (inhomogeneous) Fredholm equation (of the second kind), we employ a technique of successive substitution (see Arfken [2013] for details on other solution techniques and references to additional theory). Begin by “peeling off” the \(N\)th observation \(y := x_N\), obtaining

\[ p_{N-1}(x) = \frac{N-1}{N} K_N^\tau(x) + \frac{1}{N} \int \phi(x \mid y, \tau) p_{N-1}(y)dy. \]  \hspace{1cm} (29)

Next, substitute (29) into itself:

\[ \frac{N-1}{N} K_N^\tau(x) + \frac{1}{N} \int \phi(x \mid y, \tau) \left[ \frac{N-1}{N} K_N^\tau(y) + \frac{1}{N} \int \phi(y \mid y', \tau) p_{N-1}(y')dy' \right] dy \]

which simplifies to

\[ \frac{N-1}{N} K_N^\tau(x) + \frac{N-1}{N^2} K_N^{2\tau}(x) + \frac{1}{N^2} \int \int \phi(x \mid y, \tau) \phi(y \mid y', \tau) p_{N-1}(y')dy'dy. \]

Exchanging the order of integration (and switching the names of \(x\) and \(x'\) for notational consistency), yields

\[ \frac{N-1}{N} K_N^\tau(x) + \frac{N-1}{N^2} K_N^{2\tau}(x) + \frac{1}{N^2} \int \phi(x \mid y, 2\tau) p_{N-1}(y)dy. \]  \hspace{1cm} (30)

Note that the third term in this expression is like the second term in expression (29), with \(N^2\) in place of \(N\) and \(2\tau\) in place of \(\tau\). Therefore, repeated substitution of (29) into the recursion gives \(p_{N-1}(x)\) as

\[ p_{N-1}(x) = \sum_{j=1}^{\infty} \frac{N-1}{N^j} K_N^{j\tau}(x). \]  \hspace{1cm} (31)

as desired.