Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

On Recursive Bayesian Predictive Distributions

P. Richard Hahn, Ryan Martin & Stephen G. Walker

To cite this article: P. Richard Hahn, Ryan Martin & Stephen G. Walker (2018) On Recursive Bayesian Predictive Distributions, Journal of the American Statistical Association, 113:523, 1085-1093, DOI: 10.1080/01621459.2017.1304219

To link to this article: https://doi.org/10.1080/01621459.2017.1304219

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2018, VOL. 113, NO. 523, 1085–1093, Theory and Methods
https://doi.org/10.1080/01621459.2017.1304219

Taylor & Francis
Taylor & Francis Group

On Recursive Bayesian Predictive Distributions

P. Richard Hahn[a], Ryan Martin[b], and Stephen G. Walker[c]

[a]Booth School of Business, The University of Chicago, Chicago, IL; [b]Department of Statistics, North Carolina State University, NC; [c]Department of Mathematics, University of Texas at Austin, TX

ABSTRACT

A Bayesian framework is attractive in the context of prediction, but a fast recursive update of the predictive distribution has apparently been out of reach, in part because Monte Carlo methods are generally used to compute the predictive. This article shows that online Bayesian prediction is possible by characterizing the Bayesian predictive update in terms of a bivariate copula, making it unnecessary to pass through the posterior to update the predictive. In standard models, the Bayesian predictive update corresponds to familiar choices of copula but, in nonparametric problems, the appropriate copula may not have a closed-form expression. In such cases, our new perspective suggests a fast recursive approximation to the predictive density, in the spirit of Newton's predictive recursion algorithm, but without requiring evaluation of normalizing constants. Consistency of the new algorithm is shown, and numerical examples demonstrate its quality performance in finite-samples compared to fully Bayesian and kernel methods. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received December 2015
Revised December 2016

KEYWORDS
Copula; Density estimation; Nonparametric Bayes; Prediction; Recursive estimation

1. Introduction

Predictive distributions play a prominent role in Bayesian theory; in fact, sequences of predictive densities fully characterize a Bayesian model via the well-known de Finetti representation theorem, as discussed in de Finetti (1937) and Hewitt and Savage (1955). The Bayesian predictive density is obtained by updating the prior to the posterior and then marginalizing over the model parameters. In particular, if $f(y \mid \theta)$ is the statistical model for iid real-valued data $Y_1, \ldots, Y_n$ and π is the prior distribution for the parameter θ, then the predictive density for Y_{n+1}, given $(y_1, \ldots, y_n)$, is given by

$$p_n(y) = \int f(y \mid \theta)\, \pi_n(\mathrm{d}\theta), \tag{1}$$

where the posterior distribution π_n for θ, given $(y_1, \ldots, y_n)$, is

$$\pi_n(\mathrm{d}\theta) = \frac{\prod_{i=1}^n f(y_i \mid \theta)\, \pi(\mathrm{d}\theta)}{\int \prod_{i=1}^n f(y_i \mid \theta')\, \pi(\mathrm{d}\theta')}.$$

Intuitively, this Bayesian approach should be ideally suited to the accurate and coherent updating of information. However, by examining (1), one can see that there is no obvious route to quickly update the predictive directly: when a new observation is received, one first updates the posterior and then computes the integral to obtain the predictive. This can be especially prohibitive when Monte Carlo methods are needed to compute the posterior. The goal of this article is to show that the Bayesian predictive distribution update can, indeed, be expressed in a recursive form, making fast online Bayesian prediction possible, even in complex nonparametric models.

To show that the Bayesian predictive p_n in (1) can be updated without directly passing through the posterior, alleviating the need for Monte Carlo methods in Bayesian prediction, our starting point is a new observation that the predictive updates can be expressed in terms of a sequence of bivariate copula densities (e.g., Nelsen 1999). This observation is interesting for at least three reasons:

- according to de Finetti's representation theorem, this sequence of copula densities provides an alternative characterization of the Bayesian model itself;
- in cases where this sequence of copula densities can be identified analytically, this representation provides fast recursive updates to the Bayesian predictive;
- and, even in cases where the sequence of copula densities cannot be written down analytically, the copula representation provides new insights on how to approximate the recursive updates.

The latter point above leads to the main contribution of the article. Many applications require both flexible nonparametric modeling and fast online estimation (Caudle and Wegman 2009). Such applications include color modeling and tracking of objects (e.g., Elgammal et al. 2003; Han et al. 2008), finance (Lambert et al. 1999), network security, remote sensing. A particularly notable application area is the use of the Twitter data stream to make real-time predictions (Gerber 2014). However, the challenges in updating the Bayesian predictive are most acute in nonparametric problems, so kernel-based densities estimates (e.g., Raykar, Duraiswami, Zhao 2010; Nakamura and Hasegawa 2013) are often preferred over Bayesian methods in these applications. While Bayesian methods have been used in parametric online prediction problems (e.g., the dynamic state-space models of West et al. 1985), their adoption in analogous nonparametric settings has been limited by extreme

CONTACT Ryan Martin rgmarti3@ncsu.edu Department of Statistics, North Carolina State University, NC 27607.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

computational demands (Creal 2012). Bayesian computational methods, even those geared toward sequential analysis (e.g., Drovandi et al. 2013), do not focus on the *predictive distribution* directly, and therefore devote considerable resources to the computation of a posterior distribution over parameters. The approach in this article will be to compute the predictive distribution directly.

Perhaps the most commonly used Bayesian nonparametric model is the mixture of Dirichlet processes (e.g., Escobar 1988; Escobar and West 1995), but the need for Markov chain Monte Carlo methods to compute the posterior motivated Newton and Zhang (1999) and Newton (2002) to propose a *predictive recursion* algorithm for estimating the posterior; see, also, Martin and Ghosh (2008), Tokdar et al. (2009), and Martin and Tokdar (2009, 2011). Despite its name, the predictive recursion algorithm is not fully satisfactory for estimating the predictive distribution: it targets the posterior instead of the predictive, so integration is needed to compute normalizing constants, etc. Our copula characterization of the predictive update remains valid in nonparametric problems, but it may not be possible to derive the sequence of copula densities in closed-form. It does, however, suggest a new version of the predictive recursion algorithm that targets the predictive density directly, avoiding the difficult problem of computing normalizing constants. Besides being intuitively clear and fast to compute, we show both theoretically and numerically the accuracy of our proposed recursive predictive density estimate.

The layout of the article is as follows. In Section 2, we provide the details of our representation of the predictive via copula models and identify the particular sequence of copula densities for some common Bayesian models. Our investigation of the mixture of Dirichlet processes model lays the foundation for our recursive algorithm that directly targets the predictive densities presented in Section 3. Numerical examples given in Section 4 demonstrate that the recursive copula approach is competitive with other common density estimation methods, that is, Bayesian Gaussian mixture models and kernel density estimation methods, on prediction tasks. In Section 5, we establish Kullback–Leibler consistency of the predictive distribution sequence. Section 6 provides some concluding remarks and the online supplementary materials provide some technical proofs and other details about the recursive algorithm.

2. A New Look at Bayesian Predictive Updates

2.1. Characterizing the Updates via Copula Densities

To characterize the Bayesian predictive updates, we take a sequential point of view. That is, if p_{n-1} is the predictive density for Y_n based on observations $(y_1, \ldots, y_{n-1})$, then we want an update $(p_{n-1}, y_n) \mapsto p_n$ for the predictive density for Y_{n+1} based on observations $(y_1, \ldots, y_n)$. Consider the bivariate function $k(y, y')$ that satisfies

$$p_n(y) = p_{n-1}(y)\, k(y, y_n). \tag{2}$$

Therefore,

$$k(y, y_n) = \frac{p_n(y)}{p_{n-1}(y)}$$

which is symmetric in (y, y_n), since

$$k(y, y_n) = \frac{\int f(y \mid \theta)\, f(y_n \mid \theta)\, \pi_{n-1}(\mathrm{d}\theta)}{\int f(y \mid \theta)\, \pi_{n-1}(\mathrm{d}\theta) \int f(y_n \mid \theta)\, \pi_{n-1}(\mathrm{d}\theta)}. \tag{3}$$

The function $k(y, y_n)$ in (3) is easily seen to be a bivariate copula density function; that is, for some symmetric copula density c_n, which depends only on the sample through the sample size, we have

$$k(y, y_n) = c_n(P_{n-1}(y), P_{n-1}(y_n)) \tag{4}$$

where $c_n(u, v) = c_n(v, u)$ is a symmetric copula density, and P_{n-1} is the distribution function corresponding to the predictive density p_{n-1}.

We can now write the update $(p_{n-1}, y_n) \mapsto p_n$ as

$$p_n(y) = c_n(P_{n-1}(y), P_{n-1}(y_n))\, p_{n-1}(y) \tag{5}$$

and for each Bayesian model there is a unique sequence c_n. Now (5) allows for the direct update of the predictive and moreover it can be seen that all one needs to direct a sequence of predictive densities is to define a sequence of copula functions c_n, the key to which is that $c_n \to 1$ as $n \to \infty$, that is, the sequence of copula converges to the independent copula as the sample size increases.

To put this all into context, the de Finetti characterization of a Bayesian model is in terms of a (dependent) joint distribution over all future observables $p(y_1, y_2, y_3, \ldots)$ and such a joint distribution can always be expressed in compositional form $p(y_1)p(y_2 \mid y_1)p(y_3 \mid y_1, y_2) \ldots$. Additionally, Sklar's theorem (Sklar 1959) tells us that any joint distribution can be represented in copula form. These elements are familiar. This article focuses on the computational properties of a copula representation for the bivariate conditional distribution $p(y_n, y_{n+1} \mid y_{n-1}, \ldots y_1)$, as given in (4), which will lead to a novel approximation of the predictive update in (5).

2.2. Parametric Model Examples

In this section, we consider some standard Bayesian models, focusing on identifying the corresponding sequence c_n of copula densities that characterizes the predictive updates.

Example 1 (Exponential model). Here, we consider the model and prior as $f(y \mid \theta) = \theta e^{-\theta y}$ and $\pi(\theta) = e^{-\theta}$, respectively. Then standard calculations give

$$p_{n-1}(y) = n\, \frac{T_{n-1}^n}{(T_{n-1} + y)^{n+1}},$$

where $T_{n-1} = 1 + y_1 + \cdots + y_{n-1}$, and

$$p_n(y) = (n+1)\, \frac{(T_{n-1} + y_n)^{n+1}}{(T_{n-1} + y_n + y)^{n+2}}.$$

Therefore,

$$k(y, y_n) = \frac{(n+1)\,(T_{n-1} + y_n)^{n+1}\,(T_{n-1} + y)^{n+1}}{n\, T_{n-1}^n\,(T_{n-1} + y_n + y)^{n+2}},$$

which can be seen to be symmetric in (y, y_n). Now

$$1 - P_{n-1}(y) = \left(\frac{T_{n-1}}{T_{n-1} + y}\right)^n$$

and so $y = T_{n-1}\left[(1 - P_{n-1}(y))^{-1/n} - 1\right]$. Therefore,

$$k(y, y_n) = \frac{n+1}{n} \frac{\{1 - P_{n-1}(y)\}^{-(n+1)/n} \{1 - P_{n-1}(y_n)\}^{-(n+1)/n}}{[\{1 - P_{n-1}(y)\}^{-1/n} + \{1 - P_{n-1}(y_n)\}^{-1/n} - 1]^{n+2}}$$

and so we have the Clayton copula (Clayton 1978), that is,

$$c_n(u, v) = \frac{n+1}{n} \frac{(1-u)^{-1-1/n} (1-v)^{-1-1/n}}{\{(1-u)^{-1/n} + (1-v)^{-1/n} - 1\}^{n+2}},$$

with parameter n^{-1}, describing the sequence of predictive distributions. Note that, as $n \to \infty$, c_n converges to the independence copula.

The calculations in Example 1 can be generalized to cover an exponential family model with conjugate prior, that is, $f(y \mid \theta) = \xi(y)\, e^{y\theta - b(\theta)}$ and $\pi(\theta) \propto e^{\lambda\theta - \tau b(\theta)}$. A by-product of this argument is the identification of a new and general class of copula that contains the Archimedean class. Details are provided in the online supplementary material.

Example 2 (Normal model). Here, we consider a normal model $f(y \mid \theta) = \mathrm{N}(\mathrm{y} \mid \theta, 1)$ and a conjugate prior $\pi(\theta) = \mathrm{N}(\theta \mid 0, \tau^{-1})$. We claim that the predictive updates are characterized by a Gaussian copula with correlation parameter $\rho_n = (n + \tau)^{-1}$. In particular, we claim that the c_n in (5) is the Gaussian copula density c_{ρ_n}, where

$$c_\rho(u, v) = \frac{\mathrm{N}_2(\Phi^{-1}(\mathrm{u}), \Phi^{-1}(\mathrm{v}) \mid 0, 1, \rho)}{\mathrm{N}(\Phi^{-1}(\mathrm{u}) \mid 0, 1)\mathrm{N}(\Phi^{-1}(\mathrm{v}) \mid 0, 1)}, \qquad (6)$$

with $\mathrm{N}_2(\cdot \mid 0, 1, \rho)$ the standard bivariate normal density, with correlation ρ, and Φ the $\mathrm{N}(0, 1)$ distribution function. To see this, start with the known form for the predictive,

$$p_{n-1}(y) = \mathrm{N}\left(\mathrm{y} \;\middle|\; \frac{\mathrm{T_{n-1}}}{\mathrm{n-1+\tau}}, \frac{\mathrm{n+\tau}}{\mathrm{n-1+\tau}}\right),$$

where $T_{n-1} = y_1 + \cdots + y_{n-1}$. If we set $\mu_n = T_{n-1}/n$ and $\sigma_n^2 = (n + \tau)/(n - 1 + \tau)$, then we have

$$P_{n-1}(y) = \Phi\left(\frac{y - \mu_n}{\sigma_n}\right).$$

Then, the ratio $p_n(y)/p_{n-1}(y)$ is exponential and the key term in the exponent is

$$\left(y - \frac{y_n + T_{n-1}}{n+\tau}\right)^2 \frac{n+\tau}{n+1+\tau} - \left(y - \frac{T_{n-1}}{n-1+\tau}\right)^2 \frac{n-1+\tau}{n+\tau}.$$

Next, using the fact that $\Phi^{-1}(P_{n-1}(y)) = (y - \mu_n)/\sigma_n$, the key term in the exponent of $c_{\rho_n}(P_{n-1}(y), P_{n-1}(y_n))$ is

$$\frac{\rho_n^2}{1-\rho_n^2}\left[\left(\frac{y-\mu_n}{\sigma_n}\right)^2 + \left(\frac{y_n-\mu_n}{\sigma_n}\right)^2\right] - \frac{2\rho_n}{1-\rho_n^2}\left(\frac{y-\mu_n}{\sigma_n}\right)\left(\frac{y_n-\mu_n}{\sigma_n}\right).$$

The expressions in the two previous displays are equal up to constant terms when $\rho_n = (n + \tau)^{-1}$, which proves the claim. Note that if the model were $f(y \mid \theta) = \mathrm{N}(\mathrm{y} \mid \mu, \sigma^2)$, with $\theta = (\mu, \sigma^2)$, and we put a standard conjugate prior on the variance parameter σ^2, then we would recover the Student-t copula for the update.

Example 3 (Multinomial model). Consider a multinomial model where there are M categories and $f(y \mid \theta) = \theta_y$, where $\theta = (\theta_1, \ldots, \theta_M)$ is a probability vector. Take a conjugate prior $\theta \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_M)$, where each α_y is nonnegative. For data $y_1, \ldots, y_n$, let T^n be the frequency table, with T_y^n denoting the number of observations equal to y, $y \in \{1, \ldots, M\}$. Using the standard theory for the multinomial-Dirichlet model, the predictive distribution p_n is given by

$$p_n(y) = \frac{T_y^n + \alpha_y}{n + \beta}, \quad y \in \{1, \ldots, M\},$$

where $\beta = \sum_{j=1}^M \alpha_j$. From here, we can easily recover the predictive density ratio in (3):

$$k(y, y_n) = \frac{p_n(y)}{p_{n-1}(y)} = \frac{n - 1 + \beta}{n + \beta}\left\{1 + \frac{1(y = y_n)}{T_y^{n-1} + \alpha_y}\right\}.$$

To see what copula the predictive update corresponds to, we need to convert to the distribution function scale to find the function C_n such that

$$\begin{aligned} &C_n(P_{n-1}(y), P_{n-1}(y_n)) \\ &\quad = \sum_{z \le y,\, z' \le y_n} k(z, z') p_{n-1}(z) p_{n-1}(z') \\ &\quad = (1 - w_n)\, P_{n-1}(y) P_{n-1}(y_n) + w_n\, P_{n-1}(y) \wedge P_{n-1}(y_n), \end{aligned}$$

where $w_n = (n + \beta)^{-1}$ and $x \wedge y = \min\{x, y\}$. Thus, C_n is a mixture of the Frechet–Hoeffding copula, $C_M(u, v) = u \wedge v$, and the independence copula, $C_I(u, v) = uv$. Note that the $1 - w_n$ weight assigned to independence copula converges to 1 as $n \to \infty$.

2.3. A Nonparametric Model Example

Here, we consider a nonparametric model, namely, a mixture of Dirichlet processes model as considered by Escobar (1988) and Escobar and West (1995), given by

$$f(y, G) = \int K(y \mid \theta)\, \mathrm{d}G(\theta),$$

where $K(y \mid \theta)$ is a given kernel and the prior assigned to G is a Dirichlet process prior $\mathrm{DP}(\mathrm{c}, \mathrm{G_0})$, where G_0 is the base measure and $c > 0$ is the precision parameter (Ferguson 1973). This model was first introduced by Lo (1984) and the constructive definition of the Dirichlet process, see Sethuraman (1994), means we can write

$$f(y, G) = \sum_{j=1}^{\infty} w_j\, K(y \mid \theta_j),$$

where the (θ_j) are iid G_0 and the weights (w_j) follow a stick-breaking construction, that is, $w_1 = v_1$ and, for $j > 1$, $w_j = v_j \prod_{\ell < j}(1 - v_\ell)$, with (v_j) iid $\mathsf{Beta}(1, c)$. Hjort et al. (2010) give details on this model and inference procedures using Markov chain Monte Carlo (MCMC).

Let us assume that $K(y \mid \theta) = \mathrm{N}(\mathrm{y} \mid \theta, 1)$ and G_0 is $\mathrm{N}(0, \tau^{-1})$, as in Example 2. We can extend this to include a prior on the variance and we will recover the Student-t copula instead of the Gaussian copula. Now, for the first update, we can compute the copula density; it is given by

$$\frac{\mathrm{E}\{\mathrm{f(y, G)\, f(y_1, G)}\}}{p_0(y)\, p_0(y_1)}, \qquad (7)$$

where $p_0(y) = \int K(y \mid \theta)\, \mathrm{d}G_0(\theta)$ is a $\mathrm{N}(0, 1 + \tau^{-1})$ density, and

$$\mathrm{E}\{\mathrm{f}(\mathrm{y}, \mathrm{G})\, \mathrm{f}(\mathrm{y}_1, \mathrm{G})\} = \alpha \int K(y \mid \theta)\, K(y_1 \mid \theta)\, \mathrm{d}G_0(\theta) + (1 - \alpha)\, p_0(y)\, p_0(y_1), \tag{8}$$

and $\alpha = \sum_{j=1}^{\infty} \mathrm{E}(\mathrm{w}_j^2)$. Hence, the copula is a mixture of the Gaussian copula, c_{ρ_0}, in (6) with ρ_0 as in Example 2, and the independence copula. Rewriting to explicitly highlight the copula representation yields

$$p_1(y) = (1 - \alpha)\, p_0(y) + \alpha\, p_0(y)\, c_{\rho_0}(P_0(y), P_0(y_1)). \tag{9}$$

Note that when written in this form, $p_0(y)$ need not be Gaussian to define a valid update; the assumption of the Gaussian kernel is reflected in the form of c_{ρ_0}, and $p_0(y)$ can be any choice of the density function. One can think of this as first transforming ones data to standard normal and then applying the Bayesian update corresponding to the Dirchlet process model.

While it is not straightforward to extend the above derivation to a general update from $p_{n-1} \to p_n$ our strategy will be to iteratively apply (9) at each step, analogous to the approach of Newton for recursively approximating the posterior.

3. Nonparametric Recursive Predictive Distribution

Motivated by the calculations for the mixture of Dirichlet processes model in Section 2.3, we propose the following recursive algorithm for directly updating the predictive, completely avoiding the posterior. In particular, fix an initial guess P_0, with density p_0, and a sequence of weights $(\alpha_n) \subset (0, 1)$. Then, for $n \geq 1$ sequentially compute

$$p_n(y) = (1 - \alpha_n)\, p_{n-1}(y) + \alpha_n\, p_{n-1}(y)\, c_\rho(P_{n-1}(y), P_{n-1}(y_n)), \tag{10}$$

where c_ρ is the Gaussian copula density in (6). The sequence (α_n) is based on stick breaks which are iid $\mathrm{Beta}(1, c + n - 1)$. Therefore, they look like roughly n^{-1}, which is effectively what Newton took them to be; see (16) below. Note that $\sqrt{1 - \rho^2}$ is analogous to a kernel density bandwidth setting; after prescaling the data, we find that $\rho = 0.95$ works well in practice. Also note that the copula formulation amounts to applying a Gaussian transformation at each step, before carrying out the $n = 1$ Dirichlet update.

Here, we make three remarks. First, in the Gaussian copula model in Example 2, the sample size was captured by ρ_n but, in (10), the ρ is held fixed and the sample size is carried by α_n. Indeed, it is α_n going to 0 that takes us to the independence copula. Second, the coherence property enjoyed by the "correct" Bayesian update, that is,

$$\int \cdots \int \prod_{j=0}^{k} p_{n-j}(y_{n+1-j})\, \mathrm{y}_n \cdots \mathrm{y}_{n-k+1} = p_{n-k}(y_{n+1}), \quad k = 1, \ldots, n,$$

comes at a price—it cannot be satisfied, for $k > 1$, with a recursive estimate of the predictive. On the other hand, by sacrificing this property for $k > 1$, we can get a fast update which is still theoretically and numerically accurate. To be clear, the update p_0 to p_1 is the exact Bayesian update and, therefore, must be good; our proposal is to replicate this "good" update for all n. We lose the coherence property above, but gain computational efficiency; simulations (reported later) suggest that the resulting approximation of the predictive is satisfactory at various values of $n > 1$. Third, although Newton's original algorithm can be used to compute an approximation to the predictive, there are difficulties due to the need to evaluate intractable normalizing constants. Indeed, the ith step of Newton's original algorithm, requires evaluation of a normalizing constant $\int K(y_i \mid \theta)\, \mathrm{d}G_{i-1}(\theta)$, which cannot be computed analytically since G_{i-1} is not of any standard form. By working directly with the predictive, as we do here, there is no need to evaluate such normalizing constants.

A few words should also be said about the implementation. It is actually simpler to work on the distribution function scale, where the algorithm looks like

$$P_n(y) = (1 - \alpha_n)\, P_{n-1}(y) + \alpha_n\, H_\rho(P_{n-1}(y), P_{n-1}(y_n)), \tag{11}$$

where

$$H_\rho(u, v) = \Phi\Big(\frac{\Phi^{-1}(u) - \rho\, \Phi^{-1}(v)}{\sqrt{1 - \rho^2}}\Big). \tag{12}$$

In this formulation, it is evident that P_n in (11) is a weighted average of P_{n-1} and a suitable transformation of a normal distribution with variance $1 - \rho^2$ and centered at $\rho\Phi^{-1}(P_{n-1}(y_n))$. As ρ nears 1, this second term becomes a step distribution with single jump at $\Phi^{-1}(P_{n-1}(y_n))$. Intuitively, the method is similar to kernel density estimation, with two differences: iteratively applied adaptive transformations based on the current distribution estimate $\Phi^{-1}(P_{n-1}(\cdot))$, and shrinkage toward P_{n-1}.

Computationally, we take a fixed grid of points, $\{\bar{y}_m : m = 1, \ldots, M\}$, in $\mathbb{R}$ and compute the sequence $P_n(\bar{y}_m)$ for each m. Then the distribution function $P_n(y)$ can be plotted by interpolation. From this, the density $p_n(y)$ can be obtained by approximating the derivative by a difference ratio. Given the distribution function or density evaluated on a fine grid of points, features of the predictive distribution, such as the mean or quantiles, can be readily obtained.

We conclude this section by giving an illustration of the recursive predictive distribution estimator for univariate data; a bivariate data example is presented in the online supplementary material. We compared to a Dirichlet process mixture of normals as well as a mixture of Pólya trees. The example is taken from the R package `DPpackage` (Jara et al. 2011). The data consists of $n = 82$ velocity measurements (in km/second) of galaxies obtained from an astronomical survey of the Corona Borealis region. Figure 1 shows three density estimates: a mixture of Pólya trees, a Dirichlet process mixture of normals, a kernel density estimate, and the new recursive approximation. For this fit, we use an empirical Bayes selection of the hyperparameters with p_0 a normal density with variance 9 and mean set to the mean of the data; we also take $\rho = 0.95$ and $\alpha_i = (i + 1)^{-1}$. The priors of the Pólya tree and Dirichlet process model are set according to the demonstration code from the `DPpackage`. Note that the Dirichlet process fit is quite close to the recursive approximation (the dashed versus the solid densities).

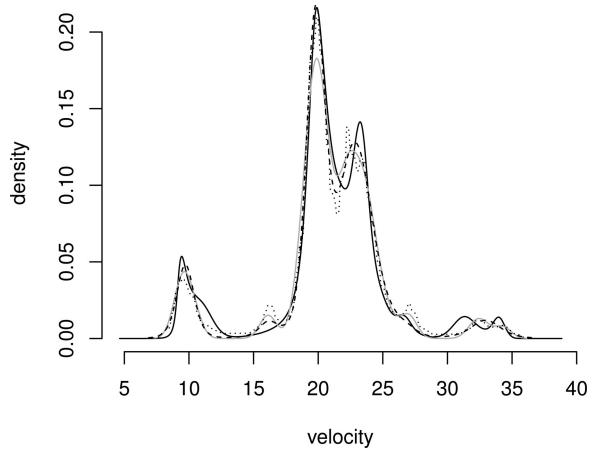

Figure 1. With $n = 82$ observations from Roeder (1990), the fits from a Pólya tree mixture model (dotted), a Dirichlet process mixture of normals (dashed), the new recursive method (solid) give visually similar fits. A kernel density estimate is also shown in gray.

4. Simulation Studies

4.1. Setup

Keeping with our focus on predictive distributions, this section evaluates the recursive estimator (11), with $\rho = 0.95$ and $\alpha_i = (i+1)^{-1}$, in terms of a predictive loss, measuring the difference between a prediction and a future realization of an observable variable. Specifically, we consider a finite vector of quantiles, defining a vector valued check-loss function. The check loss function is a piecewise linear loss function which can be expressed as

$$\ell_q(y, a) = (1-q)(a-y)\mathbb{1}(y < a) + q\mathbb{1}(y > a)(y-a), \quad q \in (0,1). \quad (13)$$

Check loss gets its name from the check-shaped graph of the function. Check loss can be justified intuitively in terms of asymmetric costs. To take a simple example, consider a restaurant: too much inventory leads to waste via spoilage at some cost per unit (purchase price), while too little inventory leads to foregone sales due to unfulfillable orders at a distinct cost per unit (because orders for multiple items are canceled in their entirety). Check loss is intimately related to quantile estimation as follows: it is straightforward to show that for any density function $f(y)$ with distribution function $F(y)$, the integrated (expected) check loss is minimized at $F^{-1}(q)$. In our simulation study, we use a vector-valued check loss function defined by a vector parameter q; specifically, we consider $q = (0.001, 0.01, 0.1, 0.25, 0.5, 0.75, 0.9, 0.99, 0.999)$.

4.2. Batch Mode Simulation Study

For our first simulation study, we compare the performance of our recursive approximation of the predictive density to that arising from the posterior of a Bayesian Dirichlet process mixture model, fit using the function `DPdensity` as well as a Pólya tree mixture model using the function `PTdensity`, both from the R package `DPpackage` (Jara et al. 2011). Because the goal of our simulation is to compare the closeness of the approximation, all model hyperparameters were calibrated to replications of the data before the simulation study was started, to ensure that the model fits were not grossly inappropriate. Details of the model fitting are available in the authors' R script.

Table 1. Summary statistics of the distribution of Δ_q defined relative to the Dirichlet process mixture of Gaussians across 500 simulations for $n = 50$ observations.

q	Mean	Median	St. Dev.
0.001	−10%	0%	56%
0.01	−26%	0%	50%
0.10	0%	0%	2%
0.25	0%	0%	1%
0.50	0%	0%	1%
0.75	0%	0%	1%
0.90	−0%	0%	3%
0.99	−3%	0%	10%
0.999	−8%	0%	90%

We generate the data, Y, according to a two component mixture of t-distributions with 5 degrees of freedom. One of these components is fixed to have location parameter 1 and scale parameter 1. The second component has mean μ and scale $s + 1$. We simulate 500 independent samples from this distribution of size $n = 50$. For each sample, the values of μ, s and the mixing proportion w, are drawn at random according to $w \sim \text{Beta}(2, 2)$, $s \sim \text{Gam}(1, 1)$ and $\mu \sim \text{N}(0, 4)$.

To evaluate each method, we compute the mean check loss on a Monte Carlo sample of size 100,000 from the true distribution, using the optimal action according to the inferred predictive distribution using each method, which we denote $a_{\text{recursive}}$ and a_{bayes}, respectively. We also compute a_{truth} which is the check loss minimizer according to the true data-generating distribution. Finally, we consider the scaled difference of integrated check loss:

$$\Delta_q = \frac{\text{E}\{\ell_q(Y, a_{\text{recursive}})\} - \text{E}\{\ell_q(Y, a_{\text{bayes}})\}}{\text{E}\{\ell_q(Y, a_{\text{truth}})\}}. \quad (14)$$

We evaluate $\Delta_q^{(j)}$ for $j = 1, \ldots, 500$ trials and a range of q. The upshot is that for the "easier" quantiles, the three methods all agree nicely. There is greater discrepancy for very high and very low quantiles; it is notable, however, that the recursive update method gives better average loss on these quantiles, although the reason why is not clear. The comparisons to the Dirichlet process mixture of normals are given in Table 1; the comparison to the Pólya tree mixture is given in Table 2. This simulation study was conducted for various samples sizes, from 10 to 100, with qualitatively similar results.

Table 2. Summary statistics of the distribution of Δ_q defined relative to the Pólya tree mixture across 500 simulations for $n = 50$ observations.

q	Mean	Median	St. Dev.
0.001	−3%	0%	67%
0.01	−22%	0%	63%
0.10	0%	0%	3%
0.25	0%	0%	1%
0.50	0%	0%	1%
0.75	0%	0%	2%
0.90	0%	0%	3%
0.99	−3%	0%	27%
0.999	−70%	−32%	159%

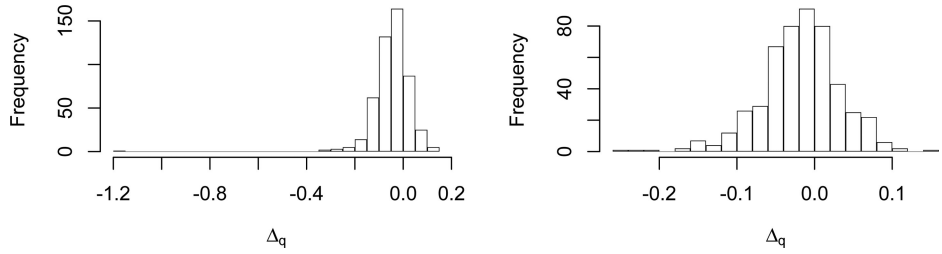

Figure 2. Distribution of scaled difference in summed check loss ($q = 0.10$) over 500 simulations, for $n = 50$ sequential observations. Negative numbers mean the recursive method outperformed the competing method. Units are in percentage of the theoretical optimal check loss. Left panel compares to Bayesian particle learning mixture model, right panel compares to kernel density estimate with adaptive bandwidth.

4.3. Sequential Simulation Study

Next, we consider online prediction according to the check loss function. That is, as individual observations arrive, we want to make an optimal action to be evaluated upon the subsequent observation. In this scenario, the extreme slowness of an MCMC approach precludes the use of the routine Gibbs sampled Gaussian mixture model, as this setting would demand rerunning the full sampling chain each time a new observation arrived. As such, our comparison method for this exercise is the Dirichlet process Gaussian mixture model particle learning method described by Carvalho, et al. (2010), which is, by construction, more computationally suited to the online setting. We do not provide the details of this method here. Additionally, we compare to a kernel density estimator with bandwidth selected by the method of Sheather and Jones (1991).

Note that the recursive bivariate copula approach is approximately as fast as the kernel density approach, with minor differences due to implementation specifics, such as what language the code is written in. The particle filter approach, while much faster than MCMC, requires storing a great deal of additional information (the "state vectors" of the filter) and, as a result, takes longer to compute. It should be mentioned that this additional overhead comes with a benefit, which is that the particle method gives full posteriors over model parameters in an online fashion; our approach bypasses those elements to directly compute the predictive and is faster as a result.

For this simulation, we consider a sample of size $n = 50$ with data generated according to the same recipe as described in the previous section. An initial four observations are used to "prime" the predictive distributions; observations are then introduced one-by-one and a check-loss-optimal prediction is made based on the posterior predictive at each time point, which is then evaluated at the subsequent observation, $j = 5, \ldots, 50$. The aggregate check-loss over this period is then computed and stored. This process is repeated for 500 simulations. For this study, we consider the tenth percentile, $q = 0.1$.

For the recursive copula method, we take p_0 a standard Cauchy distribution. We implement the particle learning algorithm using utility functions provided in the R package `Bmix` (Taddy 2010). We use default parameter values as given in provided one-dimensional density estimation demo in that package, with 200 particles. For the kernel density method, we reestimate the bandwidth with every new observation.

As before, we consider the standardized difference:

$$\Delta_q = \frac{\sum_{j=5}^{50} \ell_q\left(y_j, a_{\text{recursive}}^{(j-1)}\right) - \sum_{j=5}^{50} \ell_q\left(y_j, a_{\text{particle}}^{(j-1)}\right)}{\sum_{j=5}^{50} \ell_q\left(y_j, a_{\text{truth}}\right)}, \qquad (15)$$

where $a^{(k)}$ denotes the inferred optimal action after observing k data points. Figure 2 and Table 3 summarize the distribution of Δ_q for $q = 0.10$.

As with the batch simulation study, our claim is not that the bivariate recursive method is outright superior to these alternatives. However, these simulations highlight certain virtues of the approach—speed and ease-of-implementation—while demonstrating that the performance is broadly competitive. Concretely: our method is ten times faster than the `Bmix` package at computing the posterior predictive (with 200 particles) and our main function is 15 lines long, while the functions underlying `Bmix` are many hundreds of lines long.

It is worth emphasizing that simulation studies such as those reported here are inherently sensitive to prior specification: after all, attempting to infer the 10th percentile based on only 50 observations is a difficult task that will benefit from wise choices of prior. That said, we argue that the recursive bivariate copula approach has an advantage in terms of being relatively transparent in terms of its prior specification (the initial distribution function p_0 can be a convenient parametric form) and its hyper-parameter ρ. Mixture models of any kind do not boast this advantage.

Table 3. Summary statistics of scaled difference in summed check loss ($q = 0.10$) for two competing methods over 500 simulated datasets.

Comparison method	Mean	Median	$\Pr(\Delta_{0.1} < 0)$
Particle learning GMM	−4.6%	−4.3%	0.77
Kernel density estimate	−1.9%	−1.5%	0.64

5. Asymptotic Convergence Theory

The recursive algorithm is designed to approximate the posterior predictive under the Dirichlet process mixture model. When the sample size is large, the posterior predictive agrees with the true data-generating distribution, so it makes sense to investigate the asymptotic convergence properties of the recursive estimator P_n in (11) to the true distribution function $P^\star$. Recall that the proposed algorithm is based on a Gaussian copula via the function H_ρ in (12), and throughout we take the copula correlation parameter $\rho \in (0, 1)$ to be fixed. We will also require that the weight sequence (α_n) satisfies

$$\alpha_n = a(n+1)^{-1}, \quad n \geq 1, \qquad (16)$$

for some sufficiently small $a > 0$; see Lemma 2 below. This implies that

$$\sum_{i=1}^{\infty} \alpha_i = \infty \quad \text{and} \quad \sum_{i=1}^{\infty} \alpha_i^2 < \infty,$$

which is standard in the stochastic approximation literature (e.g., Kushner and Yin 2003).

We prove that the recursive predictive distribution sequence (P_n) converges to the true distribution $P^\star$ in the Kullback–Leibler sense, with probability 1 as $n \to \infty$. Toward this, consider the algorithm for the predictive density $p_n(y)$, given by

$$\begin{aligned} p_n(y) &= (1-\alpha_n)p_{n-1}(y) + \alpha_n p_{n-1}(y)\, c_\rho(P_{n-1}(y), P_{n-1}(Y_n)) \\ &= p_{n-1}(y)[1 + \alpha_n\{c_\rho(P_{n-1}(y), P_{n-1}(Y_n)) - 1\}], \end{aligned}$$

where $c_\rho(u, v)$ is the bivariate Gaussian copula density (6) with correlation parameter $\rho > 0$ and P_0 is an initial guess. Let K denote the Kullback–Leibler divergence, and $p^\star$ the true data-generating density; the goal is to show that $K(p^\star, p_n) \to 0$ $P^\star$-almost surely. Our analysis here is based on that in Martin and Tokdar (2009) for proving consistency of Newton's original predictive recursion algorithm. However, since there is no natural mixture model structure, some new ideas are needed. The main ingredient is a representation (18) of the Gaussian copula density as a sort of mixture.

To start, write

$$\begin{aligned} &K(p^\star, p_n) - K(p^\star, p_{n-1}) \\ &\quad = -\int \log \frac{p_n(y)}{p_{n-1}(y)} p^\star(y)\, dy \\ &\quad = -\int \log[1 + \alpha_n\{c_\rho(P_{n-1}(y), P_{n-1}(Y_n)) - 1\}] p^\star(y)\, dy. \end{aligned}$$

For x away from -1, that is, $x \approx 0$, the following inequality holds:

$$\log(1+x) \geq x - 2x^2, \quad x \approx 0.$$

This inequality can be applied in our case, since $c_\rho \geq 0$ and $\alpha_n \to 0$, and it gives

$$\begin{aligned} &K(p^\star, p_n) - K(p^\star, p_{n-1}) \\ &\quad \leq -\alpha_n \int \{c_\rho(P_{n-1}(y), P_{n-1}(Y_n)) - 1\} p^\star(y)\, dy + R_n, \end{aligned}$$

where the "remainder term" R_n is given by

$$R_n = 2\alpha_n^2 \int \{c_\rho(P_{n-1}(y), P_{n-1}(Y_n)) - 1\}^2 p^\star(y)\, dy.$$

Taking conditional expectation with respect to $\mathscr{A}_{n-1} = \sigma(Y_1, \ldots, Y_{n-1})$, we get

$$\begin{aligned} &\mathsf{E}\{K(p^\star, p_n) \mid \mathscr{A}_{n-1}\} - K(p^\star, p_{n-1}) \\ &\leq -\alpha_n \int\int \{c_\rho(P_{n-1}(y), P_{n-1}(y')) - 1\} p^\star(y) p^\star(y')\, dy\, dy' \\ &\quad + \mathsf{E}(R_n \mid \mathscr{A}_{n-1}). \end{aligned} \tag{17}$$

If the double integral above is positive, and the remainder term is negligible, then $K_n := K(p^\star, p_n)$ is an "almost supermartingale" (Robbins and Siegmund 1971) and converges to an almost sure limit, say, K_∞. To handle the double integral in (17), and to show that the limit is almost surely zero, some manipulation of the copula density c_ρ is needed.

Traditionally, the copula density is written as in Equation (6). However, for our theoretical analysis, it will be convenient to rewrite the copula density as

$$c_\rho(u, v) = \int \psi_\theta(u)\psi_\theta(v)\mathsf{N}(\theta \mid 0, \rho)\, d\theta, \tag{18}$$

where ψ_θ is a ratio of normal densities,

$$\psi_\theta(u) = \frac{\mathrm{N}(\Phi^{-1}(u) \mid \theta, 1-\rho)}{\mathrm{N}(\Phi^{-1}(u) \mid 0, 1)}.$$

This follows from routine calculations using normal convolutions. The point is that the Gaussian copula has a type of mixture or "conditionally iid" representation.

Consistency requires two preliminary results; see the online supplementary material for the proofs. For the first, write $T(p_n)$ for that double integral on the right-hand side of (17), that is,

$$T(p) = \int\int \{c_\rho(P(y), P(y')) - 1\} p^\star(y) p^\star(y')\, dy\, dy',$$

where p is a generic density with distribution function P. If we plug in the alternative representation (18) of the copula density into the formula for $T(p)$ and interchange the order of integration, we get

$$\begin{aligned} T(p) &= \int \Big[\Big\{\int \psi_\theta(P(y)) p^\star(y)\, dy\Big\}^2 - 1\Big] \mathsf{N}(\theta \mid 0, \rho)\, d\theta \\ &= \int \Big\{\int \psi_\theta(P(y)) p^\star(y)\, dy - 1\Big\}^2 \mathsf{N}(\theta \mid 0, \rho)\, d\theta, \end{aligned}$$

where the last expression follows from the formula $\mathrm{E}(\mathrm{X}^2) - \mathrm{E}^2(\mathrm{X}) = \mathrm{E}\{\mathrm{X} - \mathrm{E}(\mathrm{X})\}^2$ and the fact that $\int \psi_\theta(u)\mathrm{N}(\theta \mid 0, \rho)\, d\theta = 1$ for all u.

Lemma 1. Consider a density p whose support contains that of $p^\star$. Then, $T(p) \geq 0$ with equality if and only if $p = p^\star$ Lebesgue-almost everywhere.

Our second preliminary result demonstrates that the remainder term R_n is negligible, that is, it vanishes sufficiently fast that it does not disrupt the supermartingale-like dynamics of Kullback–Leibler sequence $K_n = K(p^\star, p_n)$.

Lemma 2. Write $\bar{P}_0 = 1 - P_0$. Suppose that P_0 and $p^\star$ satisfy

$$\int \{P_0(y) \wedge \bar{P}_0(y)\}^{-2\rho/(1+\rho)} p^\star(y)\, dy < \infty. \tag{19}$$

Furthermore, assume that a in (16) satisfies

$$0 < a < \frac{2\rho + 2}{7\rho + 1}.$$

Then, $\sum_n \mathsf{E}(R_n \mid \mathscr{A}_{n-1}) < \infty$ $P^\star$-almost surely.

The integrability condition (19) can be understood as a requirement that the recursive algorithm's initialization cannot be too light tailed compared to $p^\star$; this is consistent with our choice in Section 4 to use a heavy-tailed P_0.

Theorem 1. Let p_n be the predictive density for Y_{n+1}, given $Y_1, \ldots, Y_n$ defined above, with correlation parameter $\rho \in (0, 1)$ and with weight sequence (α_n) that satisfies (16). If the true

density $p^\star$ is continuous and satisfies (19) for the given P_0, then $K(p^\star, p_n) \to 0$ $P^\star$-almost surely.

Proof. From the expression for $\mathrm{E}(\mathrm{K_n} \mid \mathscr{A}_{\mathrm{n}-1}) - \mathrm{K_{n-1}}$, and Lemmas 1 and 2, it follows from Robbins and Siegmund (1971) that

$$K_n \to K_\infty \quad \text{and} \quad \sum_n \alpha_n T(p_n) < \infty, \quad P^\star\text{-almost surely.}$$

It remains to show that $K_\infty = 0$ $P^\star$-almost surely. Suppose, to the contrary, that $K_\infty > 0$ with positive probability. Then, p_n is away from $p^\star$ (in the Kullback–Leibler sense) for all but finitely many n with positive probability. More precisely, there is a set of positive Lebesgue measure on which $p_n \neq p^\star$. By Lemma 1, this implies $T(p_n) > 0$ for all but finitely many n. Since $T(p_n)$ is bounded away from zero, we get $\sum_n \alpha_n T(p_n) = \infty$ with positive probability, which contradicts the second conclusion in the above display. Therefore, we must have $K_\infty = 0$ almost surely, completing the proof. □

6. Conclusion

In this article, we have identified an interesting new connection between Bayesian predictive updates and well-known bivariate copulas. Besides the new light cast on this previously unknown connection between Bayesian inference and copulas, which can provide further and deeper insights and understanding about both, this development makes clear that Bayesian predictive updates do not require posterior computations. This opens the door for online Bayesian prediction, as well as for Bayesian predictive analysis for researchers who are uncomfortable with the implementation and/or slow speeds of MCMC methods.

The new recursive algorithm developed here is important because it provides a direct attack on the predictive density, which can simplify both the modeling and the computational aspects in applications. First, if the predictive is the goal, then needing to specify a mixture model, especially, a support for the mixing distribution, is undesirable, and the new algorithm circumvents this. Second, Newton's original algorithm requires computation of a normalizing constant at each iteration, and these are never available in the closed form. For mixing distributions supported on one- or two-dimensional spaces, this can easily be handled with quadrature but, to date, there is no efficient strategy for computing these normalizing constants for higher-dimensional spaces. Again, the new version that directly attacks the predictive distribution avoids all of these difficulties.

Supplementary Materials

The online supplementary material contains an extension of the proposed recursive algorithm to the bivariate case, a generalization of the result derived in Example 1, and the proofs of Lemmas 1 and 2.

Acknowledgments

The authors thank the Editor, Associate Editor, and referees for their helpful comments on the previous version of this article.

Funding

This work is partially supported by the U. S. National Science Foundation, grants DMS -1507073 and DMS -1506879, and by the U. S. Army Research Offices, Award #W911NF-15-1-0154.

References

Carvalho, C. M., Lopes, H. F., Polson, N. G., and Taddy, M. A. (2010), "Particle Learning for General Mixtures," *Bayesian Analysis*, 5, 709–740. [1090]

Caudle, K. A., and Wegman, E. (2009), "Nonparametric Density Estimation of Streaming Data using Orthogonal Series," *Computational Statistics & Data Analysis*, 53, 3980–3986. [1085]

Clayton, D. G. (1978), "A Model for Association in Bivariate Life Tables and its Application in Epidemiological Studies of Familial Tendency in Chronic Disease Incidence," *Biometrika*, 65, 141–151. [1087]

Creal, D. (2012), "A Survey of Sequential Monte Carlo methods for Economics and Finance," *Econometric Reviews*, 31, 245–296. [1086]

de Finetti, B. (1937), "La prévision: Ses lois Logiques, ses Sources Subjectives," *Annals of the Institute of Henri Poincaré*, 7, 1–68. [1085]

Drovandi, C. C., McGree, J. M., and Pettitt, A. N. (2013), "Sequential Monte Carlo for Bayesian Sequentially Designed Experiments for Discrete Data," *Computational Statistics & Data Analysis*, 57, 320–335. [1086]

Elgammal, A., Duraiswami, R., and Davis, L. S. (2003), "Efficient Kernel Density Estimation using the Fast Gauss Transform with Applications to Color Modeling and Tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25, 1499–1504. [1085]

Escobar, M. D. (1988), *Estimating the means of several normal populations by nonparametric estimation of the distribution of the means*. Unpublished PhD dissertation, Department of Statistics, Yale University. [1086,1087]

Escobar, M. D., West, M. (1995), "Bayesian Density Estimation and Inference using Mixtures," *Journal of the American Statistical Association*, 90, 577—588. [1086,1087]

Ferguson, T. S. (1973), "A Bayesian analysis of Some Nonparametric Problems," *Annals of Statistics*, 1, 209–230. [1087]

Gerber, M. S. (2014), "Predicting Crime using Twitter and Kernel Density Estimation," *Decision Support Systems*, 61, 115–125. [1085]

Han, B., Comaniciu, D., Zhu, Y., and Davis, L. S. (2008), "Sequential Kernel Density Approximation and its Application to Real Time Visual Tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7, 1186–1197. [1085]

Hewitt, E., and Savage, L. J. (1955), "Symmetric Measures on Cartesian Products," *Transactions of the American Mathematical Society*, 80, 470–501. [1085]

Hjort, N. L., Holmes, C. C., Müller, P., and Walker, S. G. (2010), *Bayesian Nonparametrics*. Cambridge University Press. [1087]

Jara, A., Hanson, T., Quintana, F., Müller, P., and Rosner, G. (2011), "DPpackage: Bayesian Semi- and Nonparametric Modeling in R," *Journal of Statistical Software*, 40, 1–30. [1088,1089]

Kushner, H. J., and Yin, G. G. (2003), *Stochastic Approximation and Recursive Algorithms and Applications*. (2nd Ed.), New York: Springer. [1091]

Lambert, C., Harrington, S., Harvey, C., and Glodjo, A. (1999), "Efficient online Nonparametric Kernel Density Estimation," *Algorithmica*, 25, 37–57. [1085]

Lo, A. Y. (1984), "On a Class of Bayesian Nonparametric Estimates. I. Density Estimates," *Annals of Statistics*, 12, 351–357. [1087]

Martin, R., and Ghosh, J. K. (2008), "Stochastic Approximation and Newton's Estimate of a Mixing Distribution," *Statistical Science*, 23, 365–382. [1086]

Martin, R., and Tokdar, S. T. (2009), "Asymptotic Properties of Predictive Recursion: Robustness and Rate of Convergence," *Electronic Journal of Statistics*, 3, 1455–1472. [1086,1091]

Martin, R., and Tokdar, S. T. (2011), "Semiparametric Inference in Mixture Models with Predictive Recursion Marginal Likelihood," *Biometrika*, 98, 567–582. [1086]

Nakamura, Y., and Hasegawa, O. (2013), "Robust Fast Online Multivariate Nonparametric Density Estimator," *Lecture Notes in Computer Science*, 8227, 180–187. [1085]

Nelsen, R. B. (1999), *An Introduction to Copulas*, New York: Springer. [1085]

Newton, M. A., and Zhang, Y. (1999), "A Recursive Algorithm for Nonparametric Analysis with Missing Data," *Biometrika*, 86, 15–26. [1086]

Newton, M. A. (2002), "On a Nonparametric Recursive Estimator of the Mixing Distribution," *Sankhya*, 64, 306–322. [1086]

Raykar, V. C., Duraiswami, R., and Zhao, L. H. (2010), "Fast Computation of Kernel Estimators," *Journal of Computational and Graphical Statistics*, 19, 205–220. [1085]

Robbins, H., and Siegmund, D. (1971), "A Convergence Theorem for Non negative Almost Supermartingales and Some Applications." in *Optimizing Methods in Statistics* (*Proc. Sympos., Ohio State Univ., Columbus*), New York: Academic Press, pp. 233–257. [1091,1092]

Sethuraman, J. (1994), "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, 4, 639–650. [1087]

Sheather, S. J., and Jones, M. C. (1991), "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation," *Journal of the Royal Statistical Society*, Series B, 53, 683–690. [1090]

Sklar, M. (1959), Fonctions de répartition á *n* dimensions et leurs marges. Université Paris 8. [1086]

Taddy, M. A. (2010), "Autoregressive Mixture Models for Dynamic Spatial Poisson Processes: Application to Tracking Intensity of Violent Crime," *Journal of the American Statistical Association*, 105, 1403–1417. [1090]

Tokdar, S. T., Martin, R., and Ghosh, J. K. (2009), "Consistency of a Recursive Estimate of Mixing Distributions," *Annals of Statistics*, 37, 2502–2522. [1086]