Bayesian Analysis (2008) **3**, Number 2, pp. 339–366

Bayesian Dynamic Density Estimation

Abel Rodriguez* and Enrique ter Horst†

Abstract. Empirical distributions in finance and economics might show heavy tails, volatility clustering, varying mean returns and multimodality as part of their features. However, most statistical models available in the literature assume some kind of parametric form (clearly neglecting important characteristics of the data) or focus on modeling extreme events (therefore, providing no information about the rest of the distribution). In this paper we develop a Bayesian nonparametric prior for a collection of distributions evolving in discrete time. The prior is constructed by defining the distribution at any time point as a Dirichlet process mixture of Gaussian distributions, and inducing dependence through the atoms of their stick-breaking decomposition. A general construction, which allows for trends, periodicities and regressors is described. The resulting model is applied to the estimation of the time-varying travel expense distribution of employees from a major development bank comparable to the IDB, IMF and World Bank.

Keywords: Dependent Dirichlet process; Nonparametric Bayes; Random probability measure; Travel Costs; Insurance Claim Distributions

1 Introduction

One of the main constraints associated with classical time series analysis is the parametric assumptions involved in the analysis. Even if the evolution process is modeled in a flexible or non-parametric way, observational and evolution noise are typically assumed to follow some parametric distribution. This means that inferences end up being restricted to the moments of the assumed distributions, and changes not captured by those moments are overlooked by the model. Besides, in many applications, the natural measurement object is the distribution itself, which can potentially present skewness and multimodality as part of its features. For example, in genetic epidemiology studies the main object of interest is the evolution of the distribution of DNA damage across time; in insurance and risk management applications, the focus is the distribution of claims and how it changes in time.

As motivating example, consider the problem of estimating the distribution of the claims submitted to an insurance company during one month. These distributions can be very non-normal, and getting accurate density estimates for the distribution of claims is relevant for a number of reasons. On one hand, accurate estimation of the tails of the distribution is important since they determine the size of the reserves for contingencies. Since reserves must be maintained in very liquid but low-yield instruments, large reserves can reduce profits; however, small reserves can create liquidity crisis that can lead

*University of California at Santa Cruz, Santa Cruz, CA, mailto:abel@soe.ucsc.edu
†Instituto de Estudios Superiores en Administración, Venezuela, mailto:enrique.terhorst@iesa.edu.ve

 DOI:10.1214/08-BA313

to bankruptcy. On the other hand, density estimates can be used to reconstruct the distribution of any functional of interest, like the size of the largest claim to be processed in a typical month.

In the context of actuarial science applications, Bayesian parametric and nonparametric methods have proven useful in modeling insurance claim distributions (Reiss and Thomas 1999; Smyth and Jorgensen 2002; Fellingham and Kottas 2007). In the analysis of credit rating scores, Huang et al. (2004) compared credit rating prediction performance between backpropagation neural network (BNN) and support vector machines (SVM), obtaining around 80% of accuracy for both methods in the context of corporate credit rating analysis. In the context of corporate bond credit rating, Chaveesuk et al. (1999) explore three of the most well known supervised neural network paradigms-backpropagation, radial basis function and learning vector quantization-for the task of rating US corporate bonds.

In this paper we develop statistical methods appropriate to estimate and predict densities that evolve in discrete time. We are particularly interested in models where computationally efficient algorithms can be developed. Our models use countably infinite mixtures of Gaussian distributions to represent the unknown density at each time point. These methods can be conceived as an extension of the Dirichlet process Mixture model (Lo 1984; Escobar and West 1995) to collections of distributions that evolve in discrete time. Dependence is built into the mixing distribution by allowing the atoms to evolve dynamically as linear state-space models. Indeed, the models we present can also be regarded as an extension of the Gaussian Dynamic Linear Models (DLMs) of West and Harrison (1997), which are Bayesian versions of the popular Kalman Filter (Kalman 1960). Gaussian DLMs are very flexible models with well-known properties, which have been successfully applied on a number of different areas, including econometrics, engineering and climatology (Kim 1994; Pesaran et al. 1995; West 1995; III et al. 1998; West et al. 1999). Dynamic linear models with more general error structures have been recently considered by Caron et al. (2008), who also use DPM priors to provide a nonparametric specification for the distribution of both the observational and evolution errors. However, unlike the models we present this paper, Caron et al. (2008) assume that such error distributions are constant in time and allow only for evolution in the mean of the process.

There is an extensive literature on Bayesian density estimation and clustering based on discrete mixtures (either finite or infinite), particularly using Gaussian distributions (Escobar and West 1995; Richardson and Green 1997; Green and Richardson 2001). The Dirichlet process (DP) (Ferguson 1973, 1974; Antoniak 1974) has become the tool of choice in Bayesian nonparametric analysis due to its flexibility, elegance and simplicity. Some examples of recent applications of the Dirichlet processes include finance, econometrics, epidemiology, genetics and survival analysis among others (Müller et al. 2004; Kacperczyk et al. 2004; Chib and Hamilton 2002; Hirano 2002; Dunson 2005; Medvedovic and Sivaganesan 2002; Dunson et al. 2007a; Kottas et al. 2002) as well as (Bigelow and Dunson 2005; Laws and O'Hagan 2002; Dunson and Park 2008; Griffin 2007; Pennell and Dunson 2008).

Most of the classical literature on the Dirichlet process focuses on exchangeable samples. However, recent work has started to develop methods for dependent distributions, either by forming convex combinations of independent processes (Müller et al. 2004; Dunson et al. 2007b; Griffin and Steel 2007; Dunson 2006) or by introducing dependence in the elements of the stick-breaking representation of the distribution (MacEachern 1999; DeIorio et al. 2004; Gelfand et al. 2005; Griffin and Steel 2006; Rodriguez et al. 2008). Particularly relevant for this paper are the works of Griffin and Steel (2006), who induce dependency through random permutations of the atoms; Griffin and Steel (2007), who construct the mixing distribution at a new time point as a linear combination of the mixing distributions at the previous time points plus an innovation; and Tang and Ghosal (2007), who are concerned with estimating the conditional distribution of a general autoregressive process. Our approach, while sharing some of the characteristics of Griffin and Steel (2006), induces dependency in the location of the mixture components. This formulation simplifies the implementation of the model since it allows us to use any of the computational tools available for the regular Dirichlet process. It also allows us to construct models with seasonal behavior and high-order autoregressive behavior, which are hard to obtain in other settings. Finally, it provides a more intuitive interpretation of the process, as well as a direct extension of widely used models for time series analysis.

The paper is organized as follows: Section 2 reviews the basic properties of the Dirichlet process. Section 3 defines our dynamic dependent Dirichlet process and discusses its main properties, while section 4 describes efficient computational tools to fit the model. Section 5 discusses some specific examples of models that fall into the class of dynamic dependent Dirichlet processes, including distributional autoregressive models and periodic distributional models. Finally section 6 shows an application of our methodology which estimates the time-varying travel expense distribution of employees from a major development bank, and Section 7 contains a discussion and conclusion.

2 The Dirichlet process

Let $(\mathcal{X}, \mathcal{B})$ be a complete and separable metric space (typically $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{B}$ are the Borel sets on $\mathcal{X}$) with associated probability distribution K. A Dirichlet process (Ferguson 1973, 1974) with baseline measure K_0 and precision α defines a distribution on the space of probability measures $\mathcal{K} \ni K$, such that $(K(B_1), \ldots, K(B_L)) \sim \mathsf{Dir}(\alpha K_0(B_1), \ldots, \alpha K_0(B_L))$ for any partition $B_1, \ldots, B_L$ of $\mathcal{X}$. If $K \sim \mathsf{DP}(\alpha K_0)$ then for any set $B \in \mathcal{B}$

$$\mathbb{E}(K(B)) = K_0(B) \qquad \mathbb{V}(K(B)) = \frac{K_0(B)(1 - K_0(B))}{\alpha + 1}$$

which justifies the interpretation of K_0 and α as mean and precision parameters. Let $\boldsymbol{\eta}_i \sim K$ for $i = 1, \ldots, n$ and $K \sim \mathsf{DP}(\alpha K_0)$. Blackwell and MacQueen (1973) showed that, by integrating out the random distribution K, the predictive rule for a new ob-

servation $\boldsymbol{\eta}_n$ given the previous observations $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_{n-1}$ is

$$\boldsymbol{\eta}_n \sim \frac{\alpha}{\alpha + n - 1} K_0 + \sum_{l<n} \frac{1}{\alpha + n - 1} \delta_{\boldsymbol{\eta}_l}$$

Sethuraman (1994) proved that $K \sim \mathsf{DP}(\alpha K_0)$ admits a *stick breaking* representation:

$$K(\cdot) = \sum_{l=1}^{\infty} w_l^* \delta_{\boldsymbol{\eta}_l^*}(\cdot)$$

where $\{\boldsymbol{\eta}_l^*\}_{l=1}^{\infty}$ are iid draws from K_0 and $w_l^* = z_l^* \prod_{k=1}^{l-1}(1 - z_k^*)$ with $\{z_l^*\}_{l=1}^{\infty}$ iid samples from a $\mathsf{Beta}(1, \alpha)$. This readily shows that the Dirichlet process places probability one on the subspace of discrete distributions. Since this is typically not a reasonable observational model, the DP is used as the mixing measure over the parameters of some continuous distribution $\psi(\cdot|\boldsymbol{\eta})$, leading to the well known Dirichlet process Mixture models (Escobar 1994; Antoniak 1974). After integrating out the baseline measure, the predictive distribution for a future observation is therefore equal to:

$$y \sim H = \int \psi(y|\boldsymbol{\eta}) K(d\boldsymbol{\eta}) \qquad\qquad K \sim \mathsf{DP}(\alpha K_0)$$

A common choice is $\psi(\cdot|\boldsymbol{\eta}) = \mathsf{N}(\cdot|\boldsymbol{\eta} = (\mu, \sigma^2))$, yielding a model that, under mild conditions, is dense in the space of absolutely continuous distributions (Lo 1984). DPM models can be implemented by exploiting the Pòlya urn scheme described above. Indeed, since observations are exchangeable, the predictive rule also yields the full conditional distributions necessary to implement Gibbs samplers (Escobar and West 1995; MacEachern 1994; Bush and MacEachern 1996; MacEachern and Müller 1998; Neal 2000). The resulting algorithm is particularly simple and efficient when the distribution associated with the baseline measure K_0 is a conjugate prior for the mixed distribution ψ.

3 Dependent Dirichlet processes in discrete time

Given a set S, consider replacing the distribution underlying Sethuraman's stick-breaking construction with a stochastic process $\{\boldsymbol{\eta}(t) : t \in S\}$ and define

$$K_t(\cdot) = \sum_{l=1}^{\infty} w_l^*(t) \delta_{\boldsymbol{\eta}_l^*(t)}(\cdot) \tag{1}$$

where $\boldsymbol{\eta}_l^*(t)$, for $l = 1, \ldots$, are independent and identically distributed sample paths from a stochastic process $\boldsymbol{\eta}(t)$ (which plays a role akin to the baseline measure K_0) and $w_l^*(t) = z_l^*(t) \prod_{s=1}^{l-1}(1 - z_s^*(t))$ with $z_l^*(t)$ also iid sample paths from another stochastic process $z(t)$ such that $z(t) \sim \mathsf{Beta}(1, \alpha(t))$ for all $t \in D$. The resulting stochastic process is called a dependent Dirichlet process (DDP) (MacEachern 1999) and defines a

distribution on the collection of random distributions indexed by the set S, such that every K_t is marginally a Dirichlet process. In the sequel, we consider mixtures of Gaussian distributions by a discrete-time DDP. Therefore, we take $S = \mathbb{N}$ and assume our underlying stochastic process to be a general random walk. For computational reasons, we focus on "single p" DDP models where $z_l^*(t) = z_l^* \sim \mathsf{Beta}(1, \alpha)$, independently of t. More specifically, letting y_{it} be the i-th observation obtained at time $t = 1, \ldots, T$, our model reduces to:

$$\begin{aligned} y_{it}|K_t &\sim \int \mathsf{N}(y_{it}|\mathbf{F}'_{it}\boldsymbol{\theta}_t, \sigma^2) K_t(d\boldsymbol{\theta}_t, d\sigma^2) & K_t &= \sum_{l=1}^{\infty} w_l^* \delta_{(\boldsymbol{\theta}_{lt}^*, \sigma_l^{*2})} \\ w_l^* &= z_l^* \prod_{s=1}^{l-1} (1 - z_s^*) & z_l^* &\sim \mathsf{Beta}(1, \alpha) \\ \boldsymbol{\theta}_{lt}^*|\boldsymbol{\theta}_{l,t-1}^*, \sigma_l^{*2} &\sim \mathsf{N}(\mathbf{G}_t\boldsymbol{\theta}_{l,t-1}^*, \sigma_l^{*2}\mathbf{W}_t) & \boldsymbol{\theta}_{0l}^*|\sigma_l^{*2} &\sim \mathsf{N}(\mathbf{m}_0, \sigma_l^{*2}\mathbf{C}_0) \\ \sigma_l^{*2} &\sim \mathsf{IG}(s_0, s_0 S_0) \end{aligned} \tag{2}$$

where $\mathsf{N}(\mu, \sigma^2)$ denotes a Gaussian distribution with mean μ and variance σ^2, while $\mathsf{IG}(s_0, s_0 S_0)$ denotes an inverse-gamma distribution with mean $s_0 S_0/(s_0 - 1)$ and s_0 degrees of freedom (see Appendix 1).

Our model assumes that the distribution for any observation y_{it} at time t can be written as a mixture of normal components with means $\mathbf{F}'_{it}\boldsymbol{\theta}_{lt}$ and variance σ_l^2, for some known matrix $\mathbf{F}_{it}$. The parameters defining the means of these distributions are allowed to move, with the evolution matrix $\mathbf{G}_t$ and the innovation variance $\mathbf{W}_t$ controlling the direction and magnitude of the change at time t. This formulation is rather general, and by appropriately choosing the sets of matrices $\{\mathbf{F}_t\}$, $\{\mathbf{G}_t\}$ and $\{\mathbf{W}_t\}$ our model can easily accommodate patterns like trends, periodicities and dynamic regressions, or even autoregressive or moving average models for densities. The variances of the mixture components are assumed to be constant in time, but allowed to change across component; and the weights associated with each distribution are also estimated from the data, and assumed to be constant in time. Therefore, the estimates of the model can be interpreted as kernel density estimates with adaptive bandwidths and time varying kernels.

The choice of a constant weights model not only simplifies computation (MacEachern 1999), but it allows for a natural interpretation of the model as a multiprocess type II model (West and Harrison 1997). In this multiprocess interpretation, an infinite number of DLM models with the same evolutionary structure but a different set of state parameters are available at every point in time. These components represent different evolutionary paths consistent with a common underlying regime described by $\{\mathbf{G}_t\}$ and $\{\mathbf{W}_t\}$, and where each observation is allocated to one of these components with constant probability. Therefore, predictions using expected posterior means can be interpreted in this model as weighted averages of the predictions obtained under an infinite number of DLMs.

As argued by MacEachern (1999), the model in (2) can be reexpressed as a Mixture of Dirichlet processes

$$y_{it} \sim \int \mathsf{N}(y_{it}|\mathbf{F}'_{it}\boldsymbol{\theta}_t, \sigma^2)K(d\boldsymbol{\Theta}, d\sigma^2) \qquad K \sim DP(\alpha K_0) \tag{3}$$

where $\boldsymbol{\Theta} = (\boldsymbol{\theta}'_1, \ldots, \boldsymbol{\theta}'_T)'$ and K_0 is the joint distribution of $(\boldsymbol{\Theta}, \sigma^2)$ induced by the evolution equations described above, which reduces to a multivariate normal-inverse gamma where

$$\begin{aligned}
\mathbb{E}(\boldsymbol{\theta}_t|\sigma^2) &= \left[\prod_{r=1}^{t} \mathbf{G}_{t-r+1}\right] \mathbf{m}_0 \\
\mathbb{V}(\boldsymbol{\theta}_t|\sigma^2) &= \sigma^2 \left[\prod_{r=1}^{t} \mathbf{G}_{t-r+1}\right] \mathbf{C}_0 \left[\prod_{r=1}^{t} \mathbf{G}_{t-r+1}\right]' + \\
&\qquad \sigma^2 \sum_{r=1}^{t-1} \left[\prod_{s=1}^{t-r} \mathbf{G}_{t-s+1}\right] \mathbf{W}_r \left[\prod_{s=1}^{t-r} \mathbf{G}_{t-s+1}\right]' + \sigma^2 \mathbf{W}_t \\
\mathbb{C}\mathrm{ov}(\boldsymbol{\theta}_{t+k}, \boldsymbol{\theta}_t|\sigma^2) &= \left[\prod_{r=1}^{k} \mathbf{G}_{t+k-r+1}\right] \mathbb{V}(\boldsymbol{\theta}_t) \\
\mathbb{E}(\sigma^2) &= \frac{s_0}{(s_0-1)} S_0 \\
\mathbb{V}(\sigma^2) &= \frac{s_0^2}{(s_0-1)^2(s_0-2)} S_0.
\end{aligned} \tag{4}$$

This representation as a DP mixture will be exploited in section 4 to develop efficient and simple computational strategies to fit these models. However, we would like to stress that our strategy does not require the explicit computation of the moments in (4), but instead exploits forward filtering and backward sampling ideas (Carter and Kohn 1994; Fruehwirth-Schnatter 1994) to efficiently integrate out unknown parameters. Although the constant weights in (2) suggest a rather restrictive model where the same number of components is used to represent every distribution (in principle, at least the largest one needed at any time point), the model is indeed flexible. Note that a good approximation to a lower number of components can be achieved at any time point by assigning different components similar values of their parameters. Indeed, if $(\boldsymbol{\theta}_{it}, \sigma_i) \approx (\boldsymbol{\theta}_{jt}, \sigma_j)$ then $w_{it}\delta_{(\boldsymbol{\theta}_{it},\sigma_i)} + w_{jt}\delta_{(\boldsymbol{\theta}_{jt},\sigma_j)} \approx (w_{it} + w_{jt})\delta_{(\boldsymbol{\theta}_{it},\sigma_i)}$. Therefore, by having components with similar parameters but that are not allocated to any observation at certain time points, we can approximate variable weights. Therefore, the price to pay for the constant-weight assumption is, in general, a slightly larger number of atoms being used.

Besides the technical advantages already mentioned, a DDP with constant weights can provide a realistic model for the data-generation process in many circumstances. For example, in a financial applications where returns from multiple companies are being modeled, we can think of the components in the mixture as representing different industries. Therefore, the weights correspond the proportion of companies in each group

(which will typically be roughly constant) and the atoms of the mixture correspond to the industry-specific return level, which will tend to vary over time

It is also worthwhile to note that our dynamic DDP encompasses a number of other models as limiting cases. On one side by letting $\mathbf{W}_t = \mathbf{0}$ for all t we have $\boldsymbol{\theta}_t = \boldsymbol{\theta}_0$ for all t and thus $K_t = K_r$ for all t, r, which is the set up of Escobar and West (1995). On the other hand, by letting $\alpha \to 0$, we revert to the class of parametric DLMs with replicates as discussed in West and Harrison (1997). Note, however, that although increasing the value of $\mathbf{W}_t$ reduces the dependence among distributions, letting $\mathbf{W}_t \to \infty \; \forall \; t$ does not yield independent density estimates at each time point, but an improper distribution for θ_t at all times $t \geq 1$. Indeed, it is known that obtaining independent distributions from a single p DDP is not possible (MacEachern 1999).

Conditional on the mixing distribution K_t it is easy to obtain the moments of the time varying distributions,

$$\begin{aligned}
\mathbb{E}(y_{it}|K_t) &= \mathbf{F}'_{it}\left[\sum_{l=1}^{\infty} w_l^* \boldsymbol{\theta}_{lt}^*\right] \\
\mathbb{V}(y_{it}|K_t) &= \mathbf{F}'_{it}\left[\sum_{l=1}^{\infty} w_l^* \boldsymbol{\theta}_{lt}^* \boldsymbol{\theta}_{lt}^{*'} - \left\{\sum_{l=1}^{\infty} w_l^* \boldsymbol{\theta}_{lt}^*\right\}\left\{\sum_{l=1}^{\infty} w_l^* \boldsymbol{\theta}_{lt}^{*'}\right\}\right]\mathbf{F}_{it} + \sum_{l=1}^{\infty} w_l^* \sigma_l^{*2} \\
\mathbb{C}\text{ov}(y_{it}, y_{i',t+k}|K_t) &= \mathbf{F}'_{it}\left[\sum_{l=1}^{\infty} w_l^* \boldsymbol{\theta}_{lt}^* \boldsymbol{\theta}_{l,t+k}^{*'} - \left\{\sum_{l=1}^{\infty} w_l^* \boldsymbol{\theta}_{lt}^*\right\}\left\{\sum_{l=1}^{\infty} w_l^* \boldsymbol{\theta}_{l,t+k}^{*'}\right\}\right]\mathbf{F}_{i',t+k}
\end{aligned}$$

$i \neq i'$. These expressions show that the process is in general nonstationary; in particular, both the mean and the variance of the estimated distributions evolve in time. It is also possible to integrate out the unknown distribution K_t under the Dirichlet process prior, which yields (see appendix 2)

$$\mathbb{E}(y_{it}) = \mathbf{F}_t \mathbb{E}(\boldsymbol{\theta}_t) \tag{5}$$

$$\mathbb{V}(y_{it}) = \frac{1}{1+\alpha}\mathbf{F}'_{it}\mathbb{V}(\boldsymbol{\theta}_t)\mathbf{F}_{it} + \mathbb{E}(\sigma^2) \tag{6}$$

$$\mathbb{C}\text{ov}(y_{it}, y_{i',t+k}) = \frac{1}{1+\alpha}\mathbf{F}'_{it}\left[\prod_{s=1}^{k}\mathbf{G}_{t+k-s+1}\right]\mathbb{V}(\boldsymbol{\theta}_t)\mathbf{F}_{i',t+k} \quad , i \neq i' \tag{7}$$

where $\mathbb{V}(\boldsymbol{\theta}_t)$ and $\mathbb{E}(\sigma^2)$ can be obtained from (4). This shows that if the evolution process is stationary and $\mathbf{F}_{it}$ is constant for every i and t, the resulting model for the distributions is a priori centered around a stationary process. Other quantities of interest, like the correlation between the probabilities of sets under the prior, can be seen in appendix 3.

Note the similarity of (7) with the replicated Gaussian DLM model where the covariance reduces to $\mathbf{F}'_{t+k}\left[\prod_{s=1}^{k}\mathbf{G}_{t+k-s+1}\right]\mathbb{E}\left[\mathbb{V}(\boldsymbol{\theta}_t|\sigma^2)\right]\mathbf{F}_t$, implying that the covariance under the discrete time DDP is strictly smaller. This might seem counterintuitive at first sight, but it is simply a consequence of the added uncertainty in the model specification. Indeed, the use of a nonparametric mixture opens the door for observations

to come from different sample paths (atoms), in which case they are independent. This reduces the overall dependence with respect to the standard Gaussian DLM, where all observations arise from a unique sample path.

4 Inference

Inferences on the dynamic DDP can be performed using the same computational techniques employed for DP mixtures models (see, for example, MacEachern (1994); Bush and MacEachern (1996); MacEachern and Müller (1998); Neal (2000); Ishwaran and James (2001)). In what follows, we describe an extension of the algorithm in MacEachern (1994) that uses ideas underlying Forward-Filtering-Backward-Sampling (FFBS) algorithms (Carter and Kohn 1994; Fruehwirth-Schnatter 1994) to efficiently compute the predictive distributions necessary to implement the algorithm.

The algorithm uses the following reparametrization of the model: let L be the current number of components that have observations allocated to them, n^*_{lt} be the number of observations in time t assigned to group l, $n_l = \sum_t n_{lt}$, $\{\boldsymbol{\Theta}^*_1, \ldots, \boldsymbol{\Theta}^*_L\}$ be the current estimated values for those paths and $\xi_{it} = l$ iff $\boldsymbol{\Theta}_{it} = \boldsymbol{\Theta}^*_{lt}$. Given values for the structural matrices $\mathbf{F}_{it}$, $\mathbf{G}_{it}$ and $\mathbf{W}_{it}$ and after initialization of the parameters, the MCMC sampler alternates through the following steps:

1. Generate $\Theta^*_l, \sigma^{*2}_l | \{\mathbf{y} | \xi_{it} = k, \} \cdots$ using the FFBS algorithm

   (a) Forward filter using the recursions

$$\mathbf{m}_{lt} = \begin{cases} \mathbf{a}_{lt} + \mathbf{A}_{lt} e_{lt} & \text{if } n_{lt} > 0 \\ \mathbf{a}_{lt} & \text{if } n_{lt} = 0 \end{cases}$$

$$\mathbf{C}_{lt} = \begin{cases} \mathbf{R}_{lt} - \mathbf{A}_{lt} \mathbf{Q}_{lt} \mathbf{A}'_{lt} & \text{if } n_{lt} > 0 \\ \mathbf{R}_{lt} & \text{if } n_{lt} = 0 \end{cases}$$

$$s_{lt} = s_{l,t-1} + n_{lt}$$

$$s_{lt} S_{lt} = \begin{cases} s_{l,t-1} S_{l,t-1} + \mathbf{e}'_{lt} \mathbf{Q}^{-1}_{lt} \mathbf{e}_{lt} & \text{if } n_{lt} > 0 \\ s_{l,t-1} S_{l,t-1} & \text{if } n_{lt} = 0 \end{cases}$$

$$\mathbf{A}_{lt} = \mathbf{R}_{lt} \mathbf{F}^*_{lt} \mathbf{Q}^{-1}_{lt}$$

$$\mathbf{e}_{lt} = \mathbf{y}_{lt} - \mathbf{f}_{lt}$$

$$\mathbf{f}_{lt} = \mathbf{F}^{*'}_{lt} \mathbf{a}_{lt}$$

$$\mathbf{Q}_{lt} = \mathbf{F}^{*'}_{lt} \mathbf{R}_{lt} \mathbf{F}^*_{lt} + \mathbf{I}$$

$$\mathbf{a}_{lt} = \mathbf{G}_{lt} \mathbf{m}_{l,t-1}$$

$$\mathbf{R}_{lt} = \mathbf{G}_{lt} \mathbf{C}_{l,t-1} \mathbf{G}'_{lt} + \mathbf{W}_{lt}$$

where $\mathbf{y}_{lt}$ is made of all observations assigned to group l at time t, $\mathbf{F}^*_{lt}$ is a matrix whose columns are the corresponding $\mathbf{F}_{it}$ vectors and $\mathbf{I}$ is the identity matrix.

(b) Sample $\sigma^{*2}_l|\mathbf{y}_l, \cdots$ from $\mathsf{IG}(s_{lT}/2, s_{lT}S_{lT}/2)$.

(c) Sample $\boldsymbol{\theta}^*_{lT}|\sigma^{*2}_l, \mathbf{y}_l, \cdots$ from $\mathsf{N}(\mathbf{m}_{lT}, \mathbf{C}_{lT})$.
Then recursively sample $\boldsymbol{\theta}^*_{lt}|\boldsymbol{\theta}^*_{l,t+1}, \sigma^{*2}_l, \mathbf{y}_l$ from $\mathsf{N}(\mathbf{d}_{lt}, \mathbf{D}_{lt})$ where

$$\begin{aligned}\mathbf{d}_{lt} &= \mathbf{m}_{lt} + \mathbf{B}_{lt}\left(\boldsymbol{\theta}_{l,t+1} - \mathbf{a}_{l,t+1}\right)\\ \mathbf{D}_{lt} &= \mathbf{C}_{lt} - \mathbf{B}_{lt}\mathbf{R}_{l,t+1}\mathbf{B}'_{lt}\\ \mathbf{B}_{lt} &= \mathbf{C}_{lt}\mathbf{G}_{t+1}R^{-1}_{l,t+1}\end{aligned}$$

2. Sample $\xi_{it}|\mathbf{y}, \boldsymbol{\xi}^-, \cdots$ from a multinomial distribution with probabilities:

$$\begin{aligned}q_l &= n^-_l\, p(y_{it}|\mathbf{y}^-, \boldsymbol{\xi}^-)\\ &= n^-_l\, \mathsf{T}_{s^-_{lT}}\left(y_{it}|\mathbf{F}'_{it}\mathbf{h}^-_{lt}, S^-_{lT}(1 + \mathbf{F}'_{it}\mathbf{H}^-_{lt}\mathbf{F}_{it})\right)\\ q_0 &= \alpha p(y_{it}|S_0)\\ &= \alpha\mathsf{T}_{s_0}\left(y_{it}|\mathbf{F}'_{it}\mathbf{h}_{0t}, S_0(1 + \mathbf{F}'_{it}\mathbf{H}_{0t}\mathbf{F}_{it})\right)\end{aligned}$$

where the superindex indicates removal of observation (i,t) from the sample, q_l for $l = 1, \ldots, L^-$ is the probability of allocating observation (i,t) to cluster l, q_0 is the probability of allocating the observation to a new cluster, $\mathbf{h}_{lT} = \mathbf{m}_{lT}$, $\mathbf{H}_{lT} = \mathbf{C}_{lT}$ and

$$\begin{aligned}\mathbf{h}_{lt} &= \mathbf{m}_{lt} + \mathbf{B}_{lt}\left(\mathbf{h}_{l,t+1} - \mathbf{a}_{l,t+1}\right)\\ \mathbf{H}_{lt} &= \mathbf{C}_{lt} - \mathbf{B}_{lt}(\mathbf{H}_{l,t+1} - \mathbf{R}_{l,t+1})\mathbf{B}'_{lt}\end{aligned}$$

Both $\mathbf{h}_{0t}$ and $\mathbf{H}_{0t}$ can be calculated using the same recursions with $n_{0t} = 0$ for all t.

The filtering and smoothing relations described above are generalizations of those found in Carter and Kohn (1994), Fruehwirth-Schnatter (1994) and West and Harrison (1997), intended to deal with missing data or multiple observations. Note that Step 2 can be computationally expensive since it implies running Forward-Filtering/Backward-Smoothing (FFBS) steps for each observation in the sample. A careful implementation requires at least one and at most two such steps for each observation: one to recalculate the parameters for the group to which the observation currently belongs and possibly another one to calculate those of the group were it is to be assigned.

Typically, the matrices $\mathbf{G}_t$ and $\mathbf{W}_t$ governing the evolution of the system might be unknown. However, since $\mathbf{W}_t$ and $\mathbf{G}_t$ define the moments of the baseline measure, inference on these quantities can be performed as discussed in Escobar and West (1998). Therefore, inferences on structural parameters like periodic components or autoregressive coefficients can be accommodated very easily. Details on some specific examples are discussed in section 5.

4.1 Smoothing and predicting density estimates

The original goal of our analysis is to obtain density estimates that borrow information across different periods and predict the shape of the density at future periods. Given D_T, which stands for all the information up to time T, the optimal estimator for the density at time $t < T$ under squared error loss corresponds to the posterior predictive distribution,

$$\hat{h}_t(\cdot|D_T) = \mathbb{E}\left[\int \mathsf{N}(\cdot|\mathbf{F}_t'\boldsymbol{\theta}_t, \sigma^2)K_t(d\boldsymbol{\theta}_t, d\sigma^2)\Big|D_T\right] = \int \mathsf{N}(\cdot|\mathbf{F}_t'\boldsymbol{\theta}_t, \sigma^2)\mathbb{E}\left[K_t(d\boldsymbol{\theta}_t, d\sigma^2)\big|D_T\right]. \tag{8}$$

We call this a filtered density estimate. In the specific case of the nonparametric DLM models discussed above, equation (8) reduces to

$$\hat{h}_t(y|D_T) = \int \left[\sum_{l=1}^{L} \frac{1}{\alpha+L}\mathsf{N}(y|\mathbf{F}_t'\boldsymbol{\theta}_{lt}^*, \sigma_l^{*2}) + \frac{\alpha}{\alpha+L}\mathsf{T}_{s_0}\left(y|\mathbf{F}_t'\mathbf{h}_{0t}, S_0[1+\mathbf{F}_t'\mathbf{H}_{0t}\mathbf{F}_t]\right)\right] p(\boldsymbol{\Theta}_1^*, \ldots, \boldsymbol{\Theta}_L^*, \sigma_1^{*2}, \ldots, \sigma_L^{*2}|D_T) \, d\boldsymbol{\Theta}_1^*, \ldots d\boldsymbol{\Theta}_L^* d\sigma_1^{*2} \ldots d\sigma_L^{*2} \tag{9}$$

Given a sample from the posterior distribution of $\{\boldsymbol{\Theta}_1^*, \ldots, \boldsymbol{\Theta}_L^*\}$ and $\{\sigma_1^{*2}, \ldots, \sigma_L^{*2}\}$, the integral in (9) can be easily evaluated for any value of y using Monte-Carlo integration. Also, k-step ahead density predictions, $\hat{f}_{t+k}(\cdot|D_t)$, can be obtained in a similar way.

5 Examples

This section discusses some specific implementations of the modeling framework discussed so far. We focus mainly on distribution autoregressive models and periodic density processes, but distributional trends and mixtures of dynamic regression models, as well as more complicated models constructed as superpositions of all these, are straightforward. We also use the DAR(1) model to extend the discussion about computational strategies for structural parameters in Section 4.

5.1 First Order Distribution Autoregressive Models

Autoregressive (AR) models are one of the most popular tools in finance and econometrics, both because of their flexibility and interpretability. For example, as noted by a number of authors (Box and Jenkins 1974; West 1997; Aguilar et al. 1999), high order stationary AR processes can be seen as approximations to Moving Average (MA) processes. Even more, Wold's decomposition ensures that high order AR processes are good approximations to any stationary process.

AR models have been used to model not only the mean structure of the time series,

but also to understand other aspects of the distributions of interest. For example, ARCH models (Engle 1982) use an AR process to model the volatility of the process. The goal of this section is to push the idea of AR processes beyond that of a model for *the moments* of distributions and get an equivalent formulation for the whole distribution. For simplicity, we start our discussion with the first-order, distribution autoregressive process (DAR(1)), which takes the form

$$y_{it} \sim \mathsf{N}(\mu_{it}, \sigma_{it}^2) \qquad (\mu_{it}, \sigma_{it}^2) \sim K_t$$
$$K_t = \sum_{l=1}^{\infty} w_l^* \delta_{(\mu_{lt}^*, \sigma_l^{*2})} \qquad \mu_{lt}^* \sim \mathsf{N}(\phi \mu_{l,t-1}^*, \sigma_l^{*2} U)$$

The name DAR(1) comes from the fact that the stochastic process defining the location of the Gaussian distributions used to represent the unknown density is an autoregressive process with autocorrelation ϕ. This is a special case of the general model described in section 3 where $\mathbf{F}_t = 1$, $\mathbf{G}_t = \phi$ and $\mathbf{W}_t = U$. Therefore, the correlation a priori induced on the observations is

$$\mathbb{C}\text{or}(y_{i,t}, y_{i',t+k}) = \frac{\phi^k}{1+\alpha}$$

This formulation extends the latent AR process models (West and Harrison 1997) to infinite mixtures. As in the latent AR process, and unlike the typical Gaussian AR(1) process, $\phi = 0$ implies uncorrelated but dependent observations. Indeed, the case $\phi = 0$ generates identifiability issues since it is not possible to separate the noise of the underlying process from the observational noise.

The model is completed by placing priors on ϕ, U, μ_0 and α. For computational simplicity a conditionally conjugate distribution for the variance of the autoregressive process can be used, $U \sim \mathsf{IG}(a_U, b_U)$. Also, in order to ensure that the model is centered around a stationary process, we set

$$\phi \sim \mathsf{N}\left(0, \tau^2\right) \mathbf{1}_{(-1,1)} \qquad \mu_{l0}^* \sim \mathsf{N}\left(0, \frac{\sigma_l^{*2} U}{1-\phi^2}\right) \quad \forall \; l$$

where $\mathsf{N}(a,b)\mathbf{1}_\Omega$ denotes the normal distribution with mean a, variance b and restricted to the set Ω. Finally, the DP precision factor α is given a $\mathsf{G}(a_\alpha, b_\alpha)$, which is conditionally conjugate.

Implementation of this model is a straightforward extension of that in section 4. Conditional on ϕ, U and α, the model is a discrete-time DDP. On the other hand, conditional on the allocation indicators, the sample paths $\{(\boldsymbol{\mu}_l^*, \sigma_l^{*2})\}_{l=1}^L$ are iid samples from the baseline measure. Therefore, samples from ϕ and U can be easily obtained using the following full-conditional distributions.

- The variance of the autoregressive process can be obtained by sampling U from

its full conditional distribution,

$$U|\cdots \sim \mathsf{IG}\left(a_U + \frac{T}{2}, b_U + \sum_l \left[\frac{(1-\phi^2)\mu_{l0}^*}{\sigma_l^{*2}} + \sum_{t=1}^{T} \frac{(\mu_{lt}^* - \mu_{l,t-1}^*)}{\sigma_l^{*2}}\right]\right)$$

- The full conditional distribution for the autocorrelation coefficient takes the form

$$p(\phi|\cdots) \propto (1-\phi^2)^{-L/2} \exp\left\{-\frac{1-\phi^2}{2U}\sum_{l=1}^{L}\frac{\mu_{l0}^{*2}}{\sigma_l^{*2}}\right\} \exp\left\{-\frac{1}{2}\left[\frac{b}{U}+\frac{1}{\tau^2}\right]\left[\phi - d\left(\frac{b}{U}+\frac{1}{\tau^2}\right)^{-1}\right]\right\}$$

where

$$b = \sum_{t=1}^{T}\sum_{l=1}^{L}\frac{\mu_{l,t-1}^{*2}}{\sigma_l^{*2}} \qquad d = \sum_{t=1}^{T}\sum_{l=1}^{L}\frac{\mu_{l,t}^*\mu_{l,t-1}^*}{\sigma_l^{*2}}$$

Note that this expression does not correspond to any known distribution. However, we recognize the third term (which happens to contain most of the information provided by the observations) as a normal kernel. Therefore an efficient independent-proposal Metropolis step can be devised to sample from this full-conditional distribution. Given the current value of the autoregression parameter $\phi_{(c)}$ in the previous iteration, propose $\phi_{(p)} \sim \mathsf{N}\left(d\left(\frac{b}{U}+\frac{1}{\tau^2}\right)^{-1}, \left(\frac{b}{U}+\frac{1}{\tau^2}\right)^{-1}\right)\mathbf{1}_{(-1,1)}$. Then, accept this proposal with probability

$$\min\left\{1, \left(\frac{1-\phi_{(p)}^2}{1-\phi_{(c)}^2}\right)^{-L/2} \exp\left\{-\frac{\phi_{(c)}^2 - \phi_{(p)}^2}{2U}\sum_{l=1}^{L}\frac{\mu_{l0}^{*2}}{\sigma_l^{*2}}\right\}\right\}$$

Otherwise retain the previous value $\phi_{(c)}$.

5.2 General Distribution Autoregressive Models

Extending the previous ideas to a DAR(p) is straightforward. The model takes the form

$$y_{it} \sim \mathsf{N}(\mu_{it}, \sigma_{it}^2) \qquad (\mu_{it}, \sigma_{it}^2) \sim K_t$$
$$K_t = \sum_{l=1}^{\infty} w_l^* \delta_{(\mu_{lt}^*, \sigma_l^{*2})} \qquad \mu_{lt}^* \sim \mathsf{N}\left(\sum_{r=1}^{p}\phi_r \mu_{l,t-r}^*, \sigma_l^{*2}U\right)$$

In terms of inference, the DAR(p) requires a slight adaptation of the FFBS algorithm described in section 4 due to the fact that $\mathbf{W}_t$ is a singular matrix. This modification is described in West and Harrison (1997), Chapter 15.3.2 for the Gaussian AR(p) model.

5.3 Periodic density processes

Annual, monthly and daily patterns are usually found in economic and financial time series, and developing models that can account for these patterns is critical in many applications. We can easily include seasonal behavior in our models by borrowing from the DLM literature. For example, consider a form-free p seasonal distributional model such that

$$\mathbf{F}_t = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \mathbf{F} \qquad \mathbf{G}_t = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 0 \end{pmatrix} = \mathbf{G}$$

where $\mathbf{F}$ is a $p \times 1$ vector and $\mathbf{G}$ is a $p \times p$ matrix. Note that $\mathbf{G}$ is a permutation matrix, such that $\mathbf{G}^p = \mathbf{I}$, and the i-th entry $\boldsymbol{\theta}_{lt}$ describes the behavior of the i-th periodic term of the l-th mixture component of the model. For example, if modeling annual effects present on monthly data we would take $p = 12$ and θ_{1lt} would represent the estimated location of the l-th mixture component in January, estimated at period t. Unlike DAR models, this structure enforces strong similarities only between distributions separated by p times periods.

6 Illustrations

6.1 Simulation study

This section presents a simulation study where we compare the performance of our dynamic density estimation model against regular kernel density estimators. The true model we use to simulate the data corresponds to a sequence of thirteen distributions, $\{h_t(\cdot)\}_{t=1}^{13}$, each being a mixture of two constant normal distributions with fixed locations and scales but time varying weights. Specifically,

$$h_t(y) = \epsilon_t \mathsf{N}(y| - 1.5, 1) + (1 - \epsilon_t)\mathsf{N}(y|1.5, 1)$$

where $\epsilon_t = 0.15 + 0.05t$ for $t = 1, \ldots, 13$. Note that this is a particularly complex situation for our model as the data generation mechanism is not in the class of models defined in the paper. Indeed, this is one case where we could reasonably expect that independent estimators (like those obtained using regular kernel density estimation) can have a better performance. For each simulation, we generate 20 observations from each one of these distributions, for a total sample size of 260 observations. The small sample sizes introduce an additional complication to the density estimation process, and allow us to demonstrate 1) the advantages of borrowing information across time and 2) that kernel density estimators can be highly unreliable for small sample sizes. This last point is particularly relevant to understand the results from the real-life application we present in the following section.

We ran a total of 15 simulations, each one using an independent data set generated as described above. For each of these data sets we fitted a DDP model that uses a simple random walk to describe the evolution of the atoms, i.e.,

$$\boldsymbol{\theta}_t = \theta_t \qquad \mathbf{F}_t = \mathbf{G}_t = 1 \qquad \mathbf{W}_t = U$$

This is a reasonable model in this situation as it centers the prior for the distribution around the posterior for the previous period, and it is the same model we use in the next section for our real-life application. Results are based on 20,000 iterations of the Gibbs sampler described in section 4, obtained after a burn-in period of 5,000 samples.

We obtained filtered density estimators $\hat{h}_t(y|D_{13})$ at each time point t as described in Section 4.1, and computed the L^1 distance between the corresponding estimate and the true distribution generating the data,

$$d_t^{DDP} = \int_{-\infty}^{\infty} |h_t(y) - \hat{h}_t(y|D_{13})| dy.$$

The above integral was approximated numerically using cubic quadrature over a grid of 200 equally-spaced points in the interval $[-8, 8]$. Simultaneously, we computed kernel density estimates independently for each time point, along with the corresponding L^1 distance d_t^{KDE}. Bandwidths were obtained through crossvalidation (see Silverman (1986) for more details). Figure 1 shows boxplots of the differences $d_t^{DDP} - d_t^{KDE}$ across all 15 simulations.

Figure 1 reveals that, in spite of the fact that the data generation mechanism is not a member of the class of models being fitted, the density estimates obtained form our model are typically closer to the true densities than regular kernel density estimates. In a typical simulation, kernel density estimates improve over the filtered estimates in no more than 3 out of 13 periods, and in around half of the simulations the filtered estimates are uniformly better.

In order to provide some additional insight into the behavior of the model we present in Figure 2 four filtered density estimates obtained in one of our simulations, along with the true distribution and kernel density estimates. The filtered distributions at $t = 5$ and $t = 9$ are clearly closer to the true distributions than the kernel density estimates. At $t = 1$ and $t = 13$ the advantage is not so clear cut, however, we can see that the filtered distributions tend to be less bumpy and better identify the location of the modes. In any case, the differences with respect to the true distribution seem to be driven by the small sample size and the results from the dynamic DDP do not appear to be worse than regular kernel density estimates. As we discussed in Section 3, the price to pay for using a constant-weight model in this setting is a relatively larger number of components in the mixture than otherwise expected; in this specific example, the dynamic DDP uses uses between 5 and 9 components to represent the sequence of distributions.

Finally, we present in Figure 3 the one-step-ahead density prediction for period 13 based on the information up to time 12, $h_{13}(y|D_{12})$, for the same data set as Figure

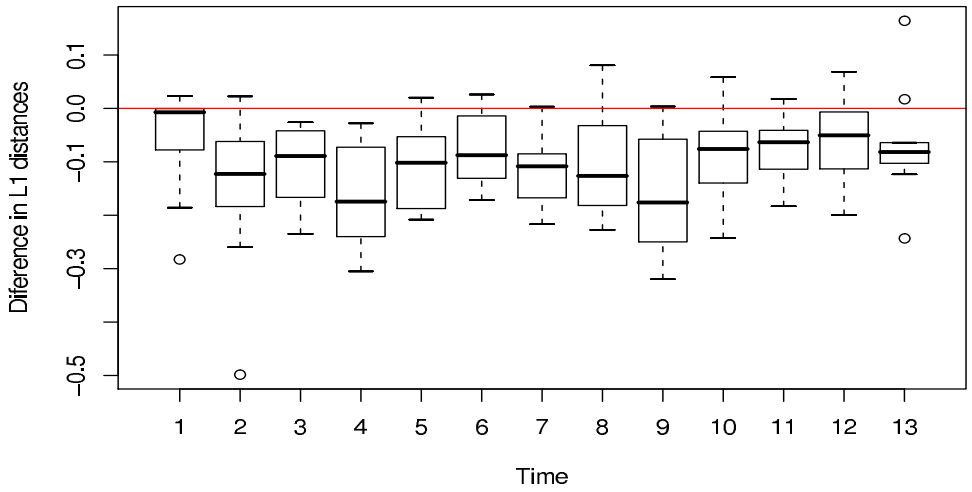

Figure 1: Differences in L^1 distances with respect to the rue distribution between the DDP model and a regular kernel density estimate for our 15 simulations. Each boxplot corresponds to one of the thirteen time points used in the study. The dynamic density estimation models typically yield more accurate density estimates, even though the data generation mechanism is not a member of the class.

2. This prediction can also be interpreted as the mean of the prior distribution on densities for $t = 13$. Although the point estimate (given by the blue (dark grey) line) differs significantly from the true distribution, the set of posterior samples (grey lines) seem to overlap with the true distribution (red (light grey) line).

6.2 Time-Varying travel expense distributions

Travel costs of employees for business trips can be an important source of cost for an organization. Assessing and mitigating them can have a major impact in reducing costs, and knowing during what periods of the year does the organization spend overall more than in other months. Knowing the travel distribution costs might help to determine what are the months where one should use internet phone (IP) instead of traveling if one has to reduce costs. The costs include from the airplane ticket fare itself to the hotel, transportation as well as meals all expressed in US dollars. The analysis performed here can of course be carried out to other problems such as estimating the claim distribution of insurance companies which is a loss to them. In what follows, we concentrate on the travel costs of employees from a major development bank between January 2005 and May 2007, applied to a total of 29 months.

Figure 4 shows the value of claims (in US dollars) in each period. Note that the distributions are right-skewed, with most of the claims being small (under $1,000) but some reaching over $20,000. Also the number of claims per month varies significantly,

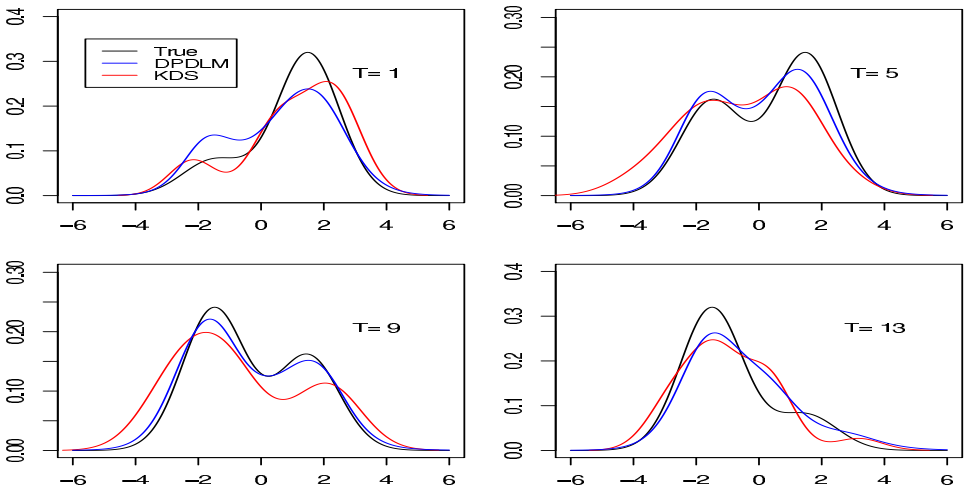

Figure 2: Examples of density estimates obtained in one representative simulation. We also present plots of the true distribution and kernel density estimates to enable comparisons.

from 3 (in January 2006) to 59 (in December 2006). Figure 5 shows kernel density estimates for the log-value of claims at six different months spanning the whole period. As before, crossvalidation was used to select the bandwidth. Clearly, distributions are multimodal and heavy-tailed even after transformations, although variances seem to be roughly constant. Also, the plots do not reveal any evident periodicity in the shape of the distributions.

In the sequel, we consider models on the logarithm of the value of the claims. Since no periodicity is evident from the density estimates, we employ again a simple random walk process for the evolution of the atoms. We assume $\mathbf{m}_0 \sim \mathsf{N}(\mu, \kappa^2)$ where $\mu = 6.0$ and $\kappa^2 = 4.0$. This choice reflects approximately the location and dispersion of the logarithm of the data. However, density estimates were similar under our sensitivity analysis, which included values of μ between 0.0 and 12.0 and values of κ^2 between 1 and 25. Prior parameters for σ^2 were chosen as $s_0 = 1.0$ and $S_0 = 1.0$, while U and α were assigned priors $\mathsf{IG}(2.0, 1.0)$ and $\mathsf{G}(1.0, 1.0)$ respectively. Again, results were robust to moderate changes in these prior parameters.

A variant of the MCMC sampler described in Section 4 was used to fit this model. All results are based on 20,000 iterations obtained after a burn-in period of 5,000 samples. No convergence problems were evident from inspection of trace plots.

The main features present in Figures 6 and 7 are those of multimodality, skewness and fat tails of the distribution of logarithm of claim amounts. The most interesting feature is when and how the mode of the travel costs during the month of December has shifted from a value of 7 to a value of 5. The opposite has happened for the month of May, where one sees that the probability mass has shifted from a low value of 4 to

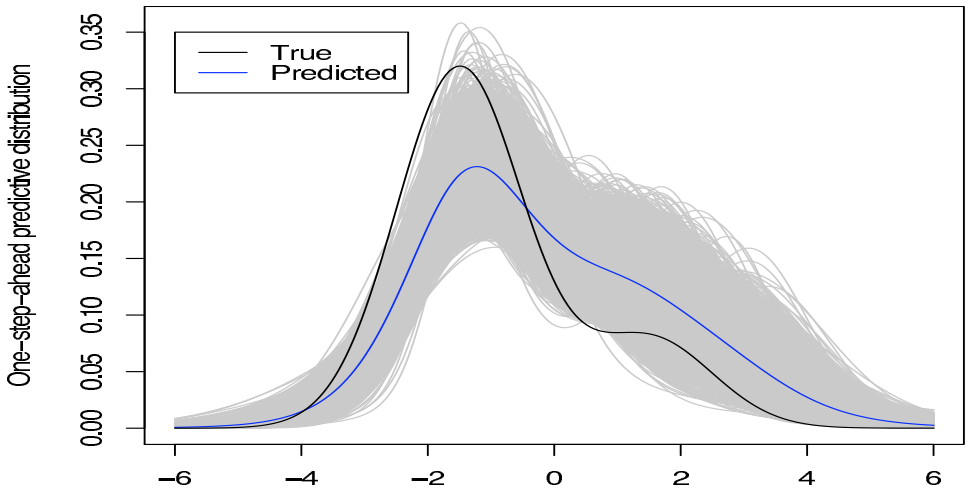

Figure 3: One step ahead predictive distribution $h_{13}(y|D_{12})$ corresponding to the same data set as in Figure 2 compared against the true density at time $t = 13$. Grey lines correspond to 1000 randomly chosen samples from the posterior distribution.

around a higher value of 7. This behavior for both the months of December and May throughout the years 2005 until 2007 could be due to the effective implementation of a new telecommunication technology (IP) during the year 2006. The shifting of the modes is clearly not only more evident in figure 6 than in figure 5, but also the peakedness of such modes.

The simulation study in Section 6.1 comes in handy to understand the marked differences between the dynamic density estimates and the simple kernel density estimates. Indeed, in the simulation we noted that, by borrowing information across time, the filtered estimates from our model are able to better capture the location of the different modes even if just the weights of the components are changing. This leads to the more noticeable peaks in the filtered density estimates. This is specially important because, for small sample sizes like the ones available in this study, estimates of the bandwidths for regular kernel density estimation (which affect the "peakedness" of the distribution) is very unreliable.

Figure 7 also depicts one-step-ahead predictive distributions for the last 5 months in our data set. They correspond to the best guess for the distribution on period t given the information at period $t-1$, as well as the mean of the prior on the distribution at time t. As expected, the peaks of the distributions on the right column are less marked than the peaks on the left column. Also, one-step-ahead predictive distributions tend to have larger variables than filtered distributions.

In general, understanding the influence of the dynamic prior on the density estimates is important . In particular, we need to make sure that the model is not overfitting the data. When a simulation analysis is not available, we can use "hold-out" data for this

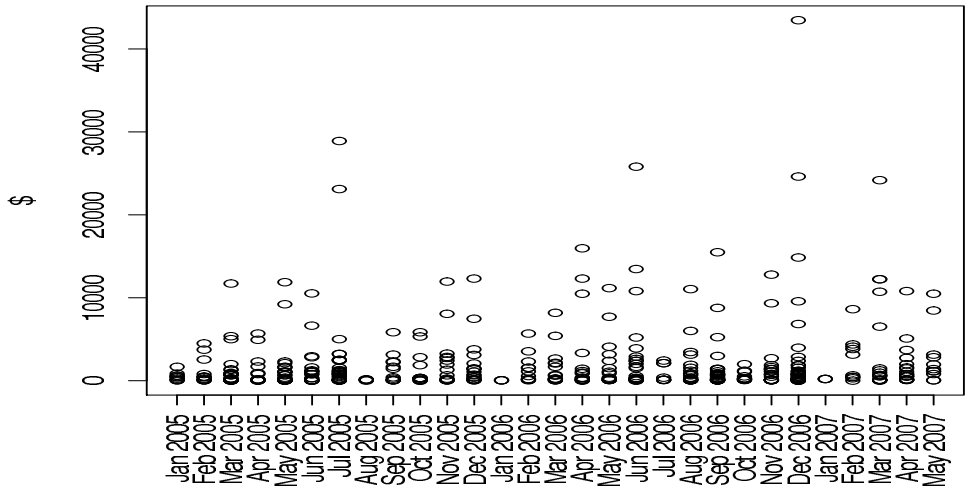

Figure 4: Claim amounts ($) as a function of time. Distributions are very skewed, and there is some evidence of multimodality.

purpose; by dividing the data in two groups, an analysis can be carried out on one of the groups and the other one can be used to asses fit. However, some care must exercised for small sample sizes like the ones in this example, as little information is available on the tails of the distribution at each time point. An alternative to this procedure consists on comparing one-step-ahead predictive distributions (which are based on information up to period $t-1$) with the actual observations obtained at time t. This is done in the left column of Figure 7, showing that the data is indeed consistent with the model assumptions.

7 Discussion

In this paper we discuss a class of models for dynamic density estimation in discrete time that allows us to borrow information across adjacent observations and obtain robust inferences. The main advantages of our method over similar approaches are simplicity, as the MCMC sampler required for implementation is an extension of well known methods, and flexibility, as including seasonal, high order autoregressive or regression terms to explain the evolution of the distribution is straightforward. We provide details for the most general location-and-scale mixture models, with the development for location mixtures constituting a straightforward simplification.

The illustration in this paper focuses on the distribution of travel reimbursement claims, but it can be immediately adapted to the estimation and prediction of insurance claim probability distributions together with their associated risk measures such as Value-at-Risk (VaR). Other applications include stochastic volatility models, risk management, actuarial science, credit analysis and FX options, as well as epidemiology,

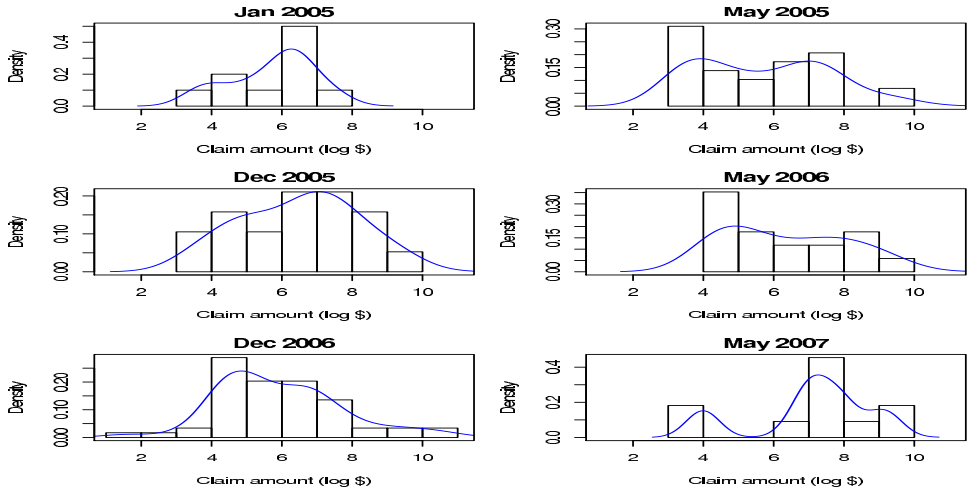

Figure 5: Kernel density estimates for distribution of the logarithm of claim amounts (log $) at six different periods. Distributions are multimodal, but variances appear roughly constant. There is no evidence of periodicities in the shape of the distribution.

climatology and engineering, some of which are the focus of current research. However, despite their generality, it is important to emphasize that the models described in this paper induce dependence in the distributions themselves and assume that observations are exchangeable within every time point given that distribution. Therefore, our model is not suitable for the analysis of longitudinal studies where the same experimental unit is followed at different times points.

The sampler we have described in this paper marginalizes over the collection of unknown distributions and samples the paths of the different components in the mixture. However, there is ample literature on MCMC techniques for Dirichlet processes that can be exploited to obtain alternative exact or approximate samplers. Some attractive options that can be readily implemented are truncation approximations or variational methods, with the latter being specially appealing for very large sample samples. In all these alternatives, the use of FFBS schemes is still key to obtain efficient algorithms.

Another important feature of the model is that, although only the means of the mixtures components are allowed to vary in time, the model is rich enough to allow for the variance of the resulting distributions to vary in time. This might seem surprising at first sight, but it is a simple consequence of the mixture structure.

There are two main drawbacks with our model formulation. As discussed by MacEachern (1999), DDPs with constant weights do not posses the intuitive asymptotic behavior expected as $\mathbf{W}_t \to \infty$. Indeed, it is impossible to generate a collection of independent distributions within this framework. However, this is hardly a limitation (at least in the context of financial application) since in most cases the existence of dependence is not in question and it is rarely the case that dependence tests are required.

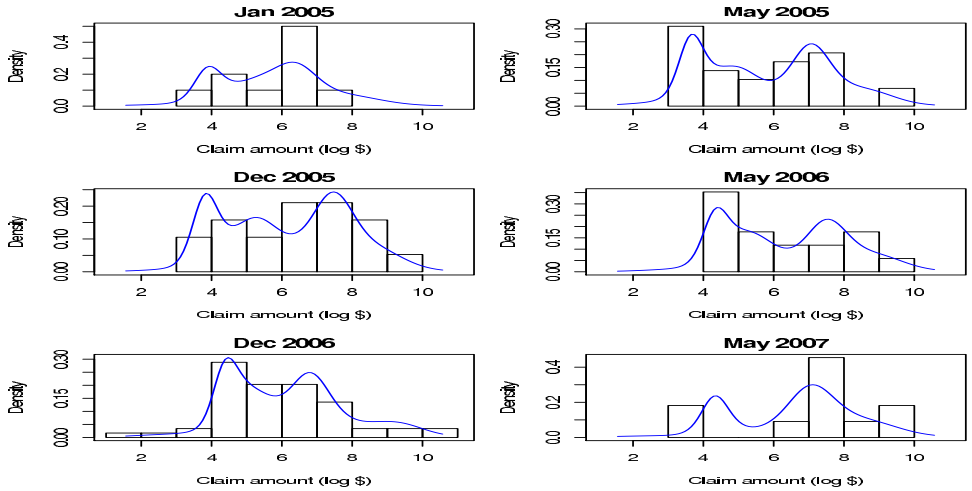

Figure 6: Retrospective dynamic density estimates for the distribution of the logarithm of claim amounts (log \$) $p(y_t|D_T)$ at six different periods. Differences with Figure 5 are due not only to adaptive smoothing features of the method, but also to the information borrowed across time.

Finally, identifiability can be an issue with single-observation time series. Indeed, it is uncommon in finance and econometric applications to have replicates at any given time. It is hard for constant weight constructions to differentiate between variability at the observational level, variability in the evolution of the parameters, and multimodality unless replicates are available. The intuition in the context of discrete-time DDPs is straightforward: are observed changes in the process due to a unimodal process with fairly large observational/evolution variance or to an almost constant multimodal distribution? Work is in progress on this specific topic.

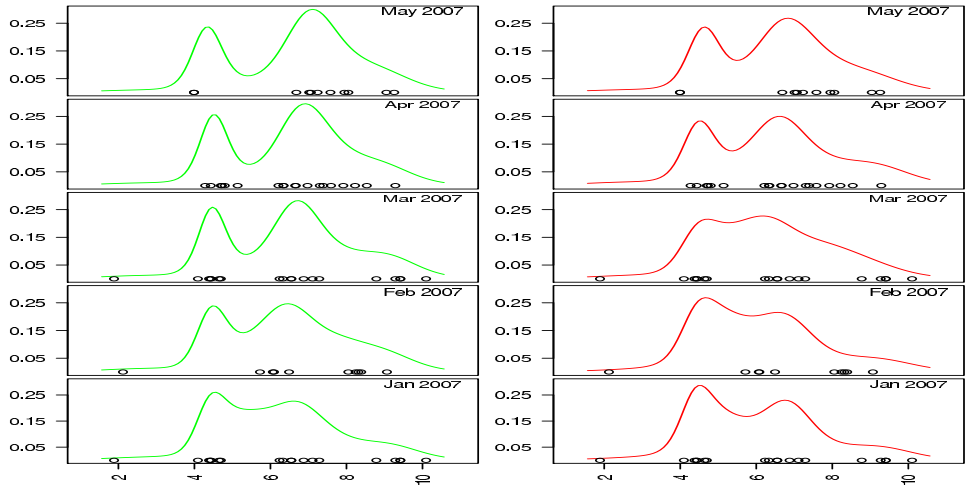

Figure 7: Dynamic density estimates $h_t(y|D_T)$ (left column) and one-step ahead predictive distributions, $h_t(y|D_{t-1})$ (right column) for claims in 2007. Dots correspond to actual observations.

1 Appendix 1: Notation

This appendix establishes the notation and parametrizations we used in the paper.

1.1 Gamma distribution

We denote $X \sim \mathsf{G}(a, b)$ if

$$p(x|a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp\{-bx\}$$

1.2 Inverse-Gamma distribution

We denote $X \sim \mathsf{IG}(a, b)$ if

$$p(x|a, b) = \frac{b^a}{\Gamma(a)} \left(\frac{1}{x}\right)^{a+1} \exp\left\{-\frac{b}{x}\right\}$$

1.3 T distribution

We denote $X \sim \mathsf{T}_\nu(\mu, \sigma^2)$ if

$$p(x) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\pi\nu}} \left(1 + \frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-(\nu+1)/2}$$

2 Appendix 2: Covariance structure in the discrete-time DDP

The result for the expectation is straightforward since

$$\mathbb{E}\left[\mathbb{E}(y_{it}|K_t\right] = \mathbf{F}'_{it}\mathbb{E}\left[\mathbb{E}(\boldsymbol{\theta}_t|K_t\right] = \mathbf{F}'_{it}\sum_{l=1}^{\infty}\mathbb{E}(w_l)\left[\mathbb{E}(\boldsymbol{\theta}^*_{lt}\right] = \mathbf{F}'_{it}\mathbb{E}(\boldsymbol{\theta}_t)\sum_{l=1}^{\infty}\mathbb{E}(w_l) = \mathbf{F}'_{it}\mathbb{E}(\boldsymbol{\theta}_t).$$

Next we show the result for the covariance. Computation for the variance follows an almost identical procedure. Note that our model implies that $y_{it}|\boldsymbol{\theta}_{it}, \sigma_i^2 = \mathbf{F}_t\boldsymbol{\theta}_{it} + \epsilon_{it}$ with errors $\epsilon_{it} \sim \mathsf{N}(0, \sigma_i^2)$ independent for every pair (i, t). Then

$$\begin{aligned}\mathbb{C}\mathrm{ov}(y_{i,t}, y_{i',t+k}) = \mathbb{C}\mathrm{ov}(\mathbf{F}_t\boldsymbol{\theta}_{it}, \mathbf{F}_{t+k}\boldsymbol{\theta}_{i',t+k}) + \mathbb{C}\mathrm{ov}(\epsilon_{it}, \mathbf{F}_{t+k}\boldsymbol{\theta}_{i',t+k}) \\ + \mathbb{C}\mathrm{ov}(\mathbf{F}_t\boldsymbol{\theta}_{it}, \epsilon_{i',t+k}) + \mathbb{C}\mathrm{ov}(\epsilon_{it}, \epsilon_{i',t+k})\end{aligned}$$

where the last three terms are zero as long as either $i \neq i'$ or $k \neq 0$. Therefore

$$\mathbb{C}\mathrm{ov}(y_{i,t}, y_{i',t+k}) = \mathbf{F}_t\mathbb{C}\mathrm{ov}(\boldsymbol{\theta}_{it}, \boldsymbol{\theta}_{i',t+k})\mathbf{F}'_{t+k}.$$

Now,

$$\begin{aligned}\mathbb{E}(\boldsymbol{\theta}_{it}\boldsymbol{\theta}_{i',t+k}) &= \mathbb{E}\left(\sum_{l=1}^{\infty} w_l^*\delta_{(\boldsymbol{\theta}^*_{lt})}\sum_{l=1}^{\infty} w_l^*\delta_{(\boldsymbol{\theta}^*_{l,t+k})}\right) \\ &= \mathbb{E}\left(\sum_{l=1}^{\infty}\sum_{r=1}^{\infty} w_l^* w_r^*\delta_{(\boldsymbol{\theta}^*_{lt})}\delta_{(\boldsymbol{\theta}^*_{r,t+k})}\right) \\ &= \mathbb{E}\left(\sum_{l=1}^{\infty} w_l^{*2}\delta_{(\boldsymbol{\theta}^*_{lt})}\delta_{(\boldsymbol{\theta}^*_{l,t+k})}\right) + \mathbb{E}\left(\sum_{l=1}^{\infty}\sum_{r=1,r\neq l}^{\infty} w_l^* w_r^*\delta_{(\boldsymbol{\theta}^*_{lt})}\delta_{(\boldsymbol{\theta}^*_{r,t+k})}\right) \\ &= \frac{1}{1+\alpha}\mathbb{E}(\boldsymbol{\theta}_t\boldsymbol{\theta}_{t+k}) + \frac{\alpha}{1+\alpha}\mathbb{E}(\boldsymbol{\theta}_t)\mathbb{E}(\boldsymbol{\theta}_{t+k}).\end{aligned}$$

Therefore

$$\begin{aligned}\mathbb{C}\mathrm{ov}(\boldsymbol{\theta}_{it}, \boldsymbol{\theta}_{i',t+k}) &= \mathbb{E}(\boldsymbol{\theta}_{it}\boldsymbol{\theta}_{i',t+k}) - \mathbb{E}(\boldsymbol{\theta}_{it})\mathbb{E}(\boldsymbol{\theta}_{i',t+k}) \\ &= \frac{1}{1+\alpha}\mathbb{E}(\boldsymbol{\theta}_t\boldsymbol{\theta}_{t+k}) - \frac{1}{1+\alpha}\mathbb{E}(\boldsymbol{\theta}_t)\mathbb{E}(\boldsymbol{\theta}_{t+k}) \\ &= \frac{1}{1+\alpha}\mathbb{C}\mathrm{ov}(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+k}).\end{aligned}$$

Since the pair $(\boldsymbol{\theta}^*_{1t}, \boldsymbol{\theta}^*_{1,t+k})$ is sampled from the baseline measure K_0.

3 Appendix 3: Covariance structure across probability distributions

Dealing with the density-valued transition probability is clearly not a straightforward matter. Instead, insight into the dependence at the random density level could be given through second order moments such as $\mathbb{C}\text{ov}(H_t(A), H_{t+k}(A))$. For simplicity, we assume that $\mathbf{F}_{it} = \mathbf{F}_t$ for all i. A priori, for any set A,

$$\begin{aligned}
\mathbb{E}(H_t(A)H_{t+k}(A)) &= \mathbb{E}\left[\sum_{l=1}^{\infty}\sum_{r=1}^{\infty} w_l^* w_r^* \Phi(A|\mathbf{F}_t'\boldsymbol{\theta}_{lt}^*, \sigma_l^{*2})\Phi(A|\mathbf{F}_{t+r}'\boldsymbol{\theta}_{r,t+k}^*, \sigma_l^{*2})\right] \\
&= \sum_{l=1}^{\infty}\mathbb{E}(w_l^{*2})\mathbb{E}\left(\Phi(A|\mathbf{F}_t'\boldsymbol{\theta}_{lt}^*, \sigma_l^{*2})\Phi(A|\mathbf{F}_{t+r}'\boldsymbol{\theta}_{l,t+k}^*, \sigma_l^{*2})\right) \\
&\quad + \sum_{l=1}^{\infty}\sum_{r\neq l}^{\infty}\mathbb{E}(w_l^* w_r^*)\mathbb{E}\left(\Phi(A|\mathbf{F}_t'\boldsymbol{\theta}_{lt}^*, \sigma_l^{*2})\right)\mathbb{E}\left(\Phi(A|\mathbf{F}_{t+r}'\boldsymbol{\theta}_{r,t+k}^*, \sigma_r^{*2})\right) \\
&= \mathbb{E}\left(\Phi(A|\mathbf{F}_t'\boldsymbol{\theta}_t, \sigma^2)\Phi(A|\mathbf{F}_{t+r}'\boldsymbol{\theta}_{t+k}, \sigma^2)\right)\sum_{l=1}^{\infty}\mathbb{E}(w_l^{*2}) \\
&\quad + \mathbb{E}\left(\Phi(A|\mathbf{F}_t'\boldsymbol{\theta}_t, \sigma^2)\right)\mathbb{E}\left(\Phi(A|\mathbf{F}_{t+k}'\boldsymbol{\theta}_{t+k}, \sigma^2)\right)\sum_{l=1}^{\infty}\sum_{r\neq l}^{\infty}\mathbb{E}(w_l^* w_r^*) \\
&= \mathbb{E}\left(\Phi(A|\mathbf{F}_t'\boldsymbol{\theta}_t, \sigma^2)\Phi(A|\mathbf{F}_{t+r}'\boldsymbol{\theta}_{t+k}, \sigma^2)\right)\frac{1}{1+\alpha} \\
&\quad + \mathbb{E}\left(\Phi(A|\mathbf{F}_t'\boldsymbol{\theta}_t, \sigma^2)\right)\mathbb{E}\left(\Phi(A|\mathbf{F}_{t+k}'\boldsymbol{\theta}_{t+k}, \sigma^2)\right)\frac{\alpha}{1+\alpha}
\end{aligned}$$

where $\Phi(A|\mu, \tau^2) = \int_A \frac{1}{\sqrt{2\pi\tau}}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\tau}\right)^2\right\}dx$. Therefore,

$$\begin{aligned}
\mathbb{C}\text{ov}(H_t(A), H_{t+k}(A)) &= \mathbb{E}\left(\Phi(A|\mathbf{F}_t'\boldsymbol{\theta}_t, \sigma^2)\Phi(A|\mathbf{F}_{t+r}'\boldsymbol{\theta}_{t+k}, \sigma^2)\right)\frac{1}{1+\alpha} \\
&\quad + \mathbb{E}\left(\Phi(A|\mathbf{F}_t'\boldsymbol{\theta}_t, \sigma^2)\right)\mathbb{E}\left(\Phi(A|\mathbf{F}_{t+k}'\boldsymbol{\theta}_{t+k}, \sigma^2)\right)\frac{\alpha}{1+\alpha} \\
&\quad - \mathbb{E}\left(\Phi(A|\mathbf{F}_t'\boldsymbol{\theta}_t, \sigma^2)\right)\mathbb{E}\left(\Phi(A|\mathbf{F}_{t+k}'\boldsymbol{\theta}_{t+k}, \sigma^2)\right) \\
&= \frac{1}{1+\alpha}\left[\mathbb{E}\left(\Phi(A|\mathbf{F}_t'\boldsymbol{\theta}_t, \sigma^2)\Phi(A|\mathbf{F}_{t+r}'\boldsymbol{\theta}_{t+k}, \sigma^2)\right)\right. \\
&\quad \left. -\mathbb{E}\left(\Phi(A|\mathbf{F}_t'\boldsymbol{\theta}_t, \sigma^2)\right)\mathbb{E}\left(\Phi(A|\mathbf{F}_{t+k}'\boldsymbol{\theta}_{t+k}, \sigma^2)\right)\right].
\end{aligned}$$

References

Aguilar, O., Huerta, G., Prado, R., and West, M. (1999). "Bayesian inference on latent structure in time series." In *Bayesian Statistics 6*, 3–26. Oxford University Press. 348

Antoniak, C. (1974). "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems." *Annals of Statistics*, 2: 1152–1174. 340, 342

Bigelow, J. and Dunson, D. (2005). "Semiparametric classification in hierarchical functional data analysis." *Institute of Statistics and Decision Sciences, Duke University*, Discussion paper 05-18. 340

Blackwell, D. and MacQueen, J. (1973). "Ferguson Distribution via Pólya Urn Schemes." *The Annals of Statistics*, 1: 353–355. 341

Box, G. and Jenkins, G. (1974). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day, 2nd edition. 348

Bush, C. and MacEachern, S. (1996). "A Semiparametric Bayesian Model for Randomised Block Designs." *Biometrika*, 83: 275–285. 342, 346

Caron, F., Davy, M., Doucet, A., and Duflos, E. (2008). "Bayesian Inference for Linear Dynamic Models with Dirichlet Process Mixtures." *IEEE Transaction on Signal Processing*, to appear. 340

Carter, C. K. and Kohn, R. (1994). "On Gibbs sampling for state space models." *Biometrika*, 81: 541–553. 344, 346, 347

Chaveesuk, R., Srivaree-Ratana, C., and Smith, A. E. (1999). "Alternative neural network approaches to corporate bond rating." *Journal of Engineering Valuation and Cost Analysis*, 2: 117–131. 340

Chib, S. and Hamilton, B. (2002). "Semiparametric Bayes Analysis of Longitudinal Data Treatment Models." *Journal of Econometrics*, 110: 67–89. 340

DeIorio, M., Müller, P., Rosner, G., and MacEachern, S. (2004). "An ANOVA Model for Dependent Random Measures." *Journal of the American Statistical Association*, 205–215. 341

Dunson, D. (2005). "Bayesian semiparametric isotonic regression for count data." *Journal of the American Statistical Association*, 100: 618–627. 340

— (2006). "Bayesian dynamic modeling of latent trait distributions." *Biostatistics*, 7: 551–568. 341

Dunson, D., Herring, A., and Mulherin-Engel, S. (2007a). "Bayesian selection and clustering of polymorphisms in functionally-related genes." *Journal of the American Statistical Association*. 340

Dunson, D. and Park, J.-H. (2008). "Kernel stick breaking processes." *Biometrika*, to appear. 340

Dunson, D., Pillai, N., and Park, J.-H. (2007b). "Bayesian density regression." *Journal of the Royal Statistical Society - Series B*, 69: 163–183. 341

Engle, R. (1982). "Evaluation and comparison of EEG traces: Latent Structure in Non-Stationary Time Series." *Econometrica*, 50: 987–1008. 349

Escobar, M. (1994). "Estimating Normal Means with a Dirichlet Process Prior." *Journal of the American Statistical Association*, 89: 268–277. 342

Escobar, M. and West, M. (1995). "Bayesian Density Estimation and Inference Using Mixtures." *Journal of the American Statistical Association*, 90: 577–588. 340, 342, 345

— (1998). "Computing Nonparametric Hierarchical Models." In Dey, D., Müller, P., and Sinha, D. (eds.), *Practical nonparametric and semiparametric Bayesian statistics*, 1–22. Springer - Verlag (Berlin, New York). 347

Fellingham, G. and Kottas, A. (2007). "Parametric and Nonparametric Bayesian Methods to Model Health Insurance Claims Costs." *University of California at Santa Cruz, Department of Applied Math and Statistics Technical Reports*. 340

Ferguson, T. (1973). "A Bayesian Analysis of Some Nonparametric Problems." *Annals of Statistics*, 1: 209–230. 340, 341

— (1974). "Prior Distributions on Spaces of Probability Measures." *Annals of Statistics*, 2: 615–629. 340, 341

Fruehwirth-Schnatter, S. (1994). "Data augmentation and dynamic linear models." *Journal of Time Series Analysis*, 15: 183–202. 344, 346, 347

Gelfand, A., Kottas, A., and MacEachern, S. (2005). "Bayesian Nonparametric Spatial Modeling With Dirichlet Process Mixing." *Journal of the American Statistical Association*, 100: 1021–1035. 341

Green, P. and Richardson, S. (2001). "Modelling Heterogeneity With and Without the Dirichlet Process." *Scandinavian Journal of Statistics*, 28: 355–375. 340

Griffin, J. (2007). "The Ornstein-Uhlenbeck Dirichlet Process and other time-varying processes for Bayesian nonparametric inference." *University of Kent, Technical Reports*. 340

Griffin, J. and Steel, M. (2006). "Order-Based Dependent Dirichlet Processes." *Journal of the American Statistical Association*, 101: 179–194. 341

— (2007). "Nonparametric Inference in Time Series Problems." mimeo, Dept. of Statistics, University of Warwick. 341

Hirano, K. (2002). "Semiparametric Bayesian Inference in Autoregressive Panel Data Models." *Econometrica*, 7: 781–799. 340

Huang, Z., Chen, H., Hsu, C.-J., and Chen, A. (2004). "Credit rating analysis with support vector machines and neural networks: a market comparative study." *Decision Support Systems*, 37: 543–558. 340

III, C. L., Carpenter, S., and Stow, C. (1998). "Forecasting PCB Concentrations in Lake Michigan Salmonids: A Dynamic Linear Model Approach." *Ecological Applications*, 8: 659–668. 340

Ishwaran, H. and James, L. (2001). "Gibbs Sampling Methods for Stick-Breaking Priors." *Journal of the American Statistical Association*, 96: 161–173. 346

Kacperczyk, M., Damien, P., and Walker, S. (2004). "A new class of Bayesian Semiparametric Models with Applications to Option Pricing." *Sauder School of Business, University of Michigan. Available at Social Science Research Network (SSRN)*. 340

Kalman, R. (1960). "A new approach to linear filtering and prediction problems." *Transactions of the ASME - Journal of Basic Engineering*, 82: 35–45. 340

Kim, C.-J. (1994). "Dynamic linear models with Markov-switching." *Journal of Econometrics*, 60: 1–22. 340

Kottas, A., Branco, M., and Gelfand, A. (2002). "A Nonparametric Bayesian Modeling Approach for Cytogenetic Dosimetry." *Biometrics*, 58: 593–600. 340

Laws, D. and O'Hagan, A. (2002). "A Hierarchical Bayes Model for Multilocation Auditing." 51: 431–450. 340

Lo, A. (1984). "On a class of Bayesian Nonparametric Estimates: I. Density Estimates." *Annals of Statistics*, 12: 351–357. 340, 342

MacEachern, S. (1994). "Estimating Normal Means with a Conjugate Style Dirichlet Process Prior." *Commnunications in Statistics, Part B - Simulation and Computation*, 23(460): 727–741. 342, 346

— (1999). "Dependent Nonparametric Processes." In *ASA Proceedings of the Section on Bayesian Statistical Science*, 50–55. 341, 342, 343, 345, 357

MacEachern, S. and Müller, P. (1998). "Estimating Mixture of Dirichlet Process Models." *Journal of Computational and Graphical Statistics*, 7: 223–238. 342, 346

Medvedovic, M. and Sivaganesan, S. (2002). "Bayesian Infinite Mixture Model-Based Clustering of Gene Expression Profiles." *Bioinformatics*, 18: 1194–1206. 340

Müller, P., Quintana, F., and Rosner, G. (2004). "Hierarchical Meta-Analysis over Related Non-parametric Bayesian Models." *Journal of the Royal Statistical Society - Series B*, 66: 735–749. 340, 341

Neal, R. (2000). "Markov Chain Sampling Methods for Dirichlet Process Mixture Models." *Journal of Computational and Graphical Statistics*, 9: 249–265. 342, 346

Pennell, M. and Dunson, D. (2008). "Nonparametric Bayes Testing of Changes in a Response Distribution with an Ordinal Predictor." *Biometrics*, to appear. 340

Pesaran, M. H., Smith, R. P., and Im, K.-S. (1995). "Dynamic Linear Models for Heterogeneous Panels." *Faculty of Economics, University of Cambridge*, (Cambridge Working Papers in Economics). 340

Reiss, R. and Thomas, M. (1999). "A new class of Bayesian estimator in Paretian excess-of-loss reinsurance." *ASTIN Bulletin*, 29: 339–349. 340

Richardson, S. and Green, P. (1997). "On Bayesian Analysis of Mixtures with an Unknown Number of Components." *Journal of the Royal Statistical Society - Series B*, 59(4): 731–792. 340

Rodriguez, A., Dunson, D. B., and Gelfand, A. E. (2008). "The nested Dirichlet process." *Journal of the American Statistical Association, to appear with Discussion*. 341

Sethuraman, J. (1994). "A Constructive Definition of Dirichlet Priors." *Statistica Sinica*, 4: 639–650. 342

Silverman, B. (1986). *Density Estimation*. Chapman and Hall, London. 352

Smyth, G. and Jorgensen, B. (2002). "Fitting Tweedie's compound Poisson model to insurance claims data: dispersion modelling." *ASTIN Bulletin*, 32(1): 143–157. 340

Tang, Y. and Ghosal, S. (2007). "A Consistent Nonparametric Bayesian Procedure for Estimating Autoregressive Conditional Densities." *Computational Statistics & Data Analysis*, 51: 4424–4437. 341

West, M. (1995). "Bayesian Time Series: Models and Computations for the Analysis of Time Series in the Physical Sciences." In *XV Workshop on Maximum Entropy and Bayesian Methods*. Oxford University Press. 340

— (1997). "Time Series Decomposition." *Biometrika*, 84: 489–494. 348

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*. Reading, MA: Springer - Verlag, New York, second edition. 340, 343, 345, 347, 349, 350

West, M., Prado, R., and Krystal, A. (1999). "Evaluation and comparison of EEG traces: Latent Structure in Non-Stationary Time Series." *Journal of the American Statistical Association*, 94: 1083–1095. 340

Acknowledgments

The authors would like to thank the anonymous referees as well as Brad Carlin, Alan Gelfand, David Dunson and Mike West for their helpful comments. This work was supported in part by the Intramural Research Program of the NIH, National Institute of Environmental Health Science.