A Bayesian Nonparametric Markovian Model for Nonstationary Time Series

Maria DeYoreo and Athanasios Kottas *

Abstract

Stationary time series models built from parametric distributions are, in general, limited in scope due to the assumptions imposed on the residual distribution and autoregression relationship. We present a modeling approach for univariate time series data, which makes no assumptions of stationarity, and can accommodate complex dynamics and capture non-standard distributions. The model for the transition density arises from the conditional distribution implied by a Bayesian nonparametric mixture of bivariate normals. This results in a flexible autoregressive form for the conditional transition density, defining a time-homogeneous, nonstationary Markovian model for real-valued data indexed in discrete time. To obtain a computationally tractable algorithm for posterior inference, we utilize a square-root-free Cholesky decomposition of the mixture kernel covariance matrix. Results from simulated data suggest the model is able to recover challenging transition densities and nonlinear dynamic relationships. We also illustrate the model on time intervals between eruptions of the Old Faithful geyser. Extensions to accommodate higher order structure and to develop a state-space model are also discussed.

KEY WORDS: Autoregressive models; Bayesian nonparametrics; Dirichlet process mixtures; Markov chain Monte Carlo; nonstationarity; time series.

*M. DeYoreo (maria.deyoreo@stat.duke.edu) is Postdoctoral Researcher, Department of Statistical Science, Duke University, and A. Kottas (thanos@ams.ucsc.edu) is Professor of Statistics, Department of Applied Mathematics and Statistics, University of California, Santa Cruz.

1 Introduction

Consider a time series of continuous random variables $(Z_1, \ldots, Z_n)$ observed at equally spaced time points $t = 1, \ldots, n$. It is common to assume dependence on lagged terms, or that Z_t depends on $(Z_{t-1}, \ldots, Z_{t-p})$, for some $p \geq 1$. The relationship between Z_t and $(Z_{t-1}, \ldots, Z_{t-p})$ is generally assumed to be linear, with error terms arising from a given parametric distribution. The simplest scenario involves $p = 1$ and normally distributed errors, referred to as a first-order Gaussian autoregression.

Time series are generally assumed to be time-homogeneous, that is, the transition density that defines the conditional distribution of Z_t given $(Z_{t-1}, \ldots, Z_{t-p})$ does not change with time. A stronger assumption is that of stationarity, which requires that the finite dimensional distributions of the time series are invariant under time shifts. Weak stationarity requires only the mean to be constant across time and the covariance function to be invariant under time shifts.

Stationary time series models are not appropriate for many applications. Stochastic systems may go through structural changes, and as a consequence, the data they produce may require models which change across time. While stationarity is a convenient property, stationary models do not allow for this type of evolution, as they assume constant means and variances across time. For instance, economic time series are commonly believed to be nonstationary (e.g., Früwirth-Schnatter, 2006).

Additionally, customary parametric time series models (both stationary and nonstationary) are generally restrictive in terms of the transition and marginal densities they imply. Parametric stationary densities are unable to accommodate time series that exhibit asymmetric or non-standard marginal distributions. Tong (1990) gives an example of a real time series that possesses a bimodal marginal distribution. Conditional distributions may also be multimodal, for instance when the stock-market is volatile, price changes may be more likely to be large in magnitude than near zero, hence it is reasonable to expect a bimodal distribution (Wong and Li, 2000).

Various parametric models have been developed to capture nonlinear autoregressive

(AR) behavior and/or relax the stationarity assumption. Time-varying autoregressions (TVAR) naturally extend AR models, by allowing the parameters to evolve in time, and thus can be used to describe nonstationary time series. TVAR models have a dynamic linear model (DLM) representation and belong to the larger class of Markovian state-space models. Such models require specification of an observation density and a state evolution density, which need not rely on normality or linearity, though these are common assumptions.

The DLM framework can be made more flexible by combining multiple DLMs, referred to as multiprocess models (West and Harrison, 1999). Mixture models of various forms have been used to move away from parametric assumptions, and capture changes over time in a series which may not be described well by a single model. The threshold autoregressive (TAR) model (Tong, 1987; Geweke and Terui, 1993) describes an AR process whose parameters switch according to the value of a previous observation, and is a special case of the Markov switching autoregressive model. We refer to Tong (1990) for a review of nonlinear time series, and Früwirth-Schnatter (2006) for a thorough review of mixture models for time series. Mixture autoregressive models (Juang and Rabiner, 1985; Wong and Li, 2000) are also special cases of Markov switching AR models, in which the parameters of the autoregression change according to a hidden Markov process.

The models discussed above generally achieve nonstationarity or nonlinearity by allowing parameters to switch or evolve in time. These models are naturally suited to problems in which a single parametric model holds in a given interval of time. For instance, the TAR structure assumes only one linear submodel applies at any particular time, with abrupt changes at the thresholds. In contrast, mixture models can be obtained by introducing hierarchical priors on model parameters, to yield a set of parametric models which are favored with different probabilities across time. These models possess the ability to capture features which could not be accommodated under the assumption of a single parametric distribution at a particular point in time. To this end, a mixture modeling approach involving Bayesian nonparametric techniques was first proposed by Müller et al. (1997). More recently, Di Lucca et al. (2013) have utilized dependent Dirichlet process priors to build countable mixtures of AR models as well as variations of this model. Antoniano-Villalobos

and Walker (2016) developed stationary time series models which contain general transition and invariant densities. Existing mixture models for time series are discussed further in Section 2.4, relative to our proposed model.

Here, we present a general framework for modeling univariate time series data, which assumes time-homogeneity but makes no assumptions of stationarity, and can accommodate complex, nonlinear dynamics as well as non-standard distributions. The proposed model for the transition density takes the form of a location-scale mixture of normal densities, with means and mixture weights which depend on the previous state(s). This structure arises from the conditional distribution implied by a Bayesian nonparametric mixture of bivariate normals. Key to the posterior simulation method is a square-root-free Cholesky decomposition of the mixture kernel covariance matrix. As demonstrated with synthetic and real data, the model enables general inference for time-homogeneous, nonstationary Markovian processes indexed in discrete time.

The rest of the paper is organized as follows. The methodology is presented in Section 2, including the model formulation for the transition density, and methods for prior specification and posterior simulation (technical details for the latter are included in the appendices). To place our contribution within the relevant literature, we also discuss certain classes of mixture models for discrete-time Markovian processes. In Section 3, the modeling approach is illustrated with simulated data examples, and it is also applied to a standard data set on waiting times between successive eruptions of the Old Faithful geyser. For the real data example, we also consider comparison with a parametric TAR model and with a more structured version of the proposed mixture model which ensures existence of a stationary distribution for the Markov chain; this latter model is essentially the one developed by Antoniano-Villalobos and Walker (2016). While the model development and data illustrations are focused on univariate time series with first-order dependence, in Section 4, we discuss possible extensions to accommodate higher order structure and to develop a state-space model. Finally, Section 5 concludes with a summary.

2 Methodology

2.1 Model Formulation

Here, we present the model for nonstationary time series. We focus on the case with first-order Markovian dependence, discussing the extension to modeling higher order time series in Section 4. Hence, the observed time series, $(z_1, \ldots, z_n)$, is assumed to be a realization from a time-homogeneous, real-valued, first-order Markov chain, and thus the likelihood, conditional on z_1, is given by $\prod_{t=2}^{n} f(z_t \mid z_{t-1})$, where $f(z_t \mid z_{t-1})$ is the transition density.

To flexibly model the transition density, we use the conditional density $f(y \mid x)$ induced by a nonparametric mixture of bivariate normal distributions for $f(x, y)$. More specifically, $f(x, y) \equiv f(x, y \mid G) = \int \mathrm{N}(x, y \mid \boldsymbol{\mu}, \Sigma)\, \mathrm{d}G(\boldsymbol{\mu}, \Sigma)$, with a Dirichlet process (DP) prior (Ferguson, 1973) for the random mixing distribution G. In the ensuing model expressions, we work with a truncated version of G motivated by the DP constructive definition (Sethuraman, 1994), which is also the approach we follow for posterior simulation (Ishwaran and James, 2001). Under a truncated DP at level L, the joint density can be expressed as $f(x, y \mid G) = \sum_{l=1}^{L} p_l \mathrm{N}(x, y \mid \boldsymbol{\mu}_l, \Sigma_l)$. Here, the $(\boldsymbol{\mu}_l, \Sigma_l)$ are independent and identically distributed (i.i.d.) from the DP base distribution G_0, and the weights $(p_1, \ldots, p_L)$ are determined through stick-breaking from latent beta$(1, \alpha)$ random variables. In particular, $p_1 = v_1$, $p_l = v_l \prod_{r=1}^{l-1}(1 - v_r)$, for $l = 2, \ldots, L-1$, and $p_L = \prod_{r=1}^{L-1}(1 - v_r)$, where $v_1, \ldots, v_{L-1} \overset{i.i.d.}{\sim} \text{beta}(1, \alpha)$. The choice of the truncation level L is discussed in Section 3.

Partitioning $\boldsymbol{\mu}_l$ and Σ_l with superscripts x and y, the conditional distribution for $f(y \mid x, G)$ implied by $f(x, y \mid G)$ is used as the model for the transition density:

$$f(z_t \mid z_{t-1}, G) = \sum_{l=1}^{L} q_l(z_{t-1}) \mathrm{N}\left(z_t \mid \mu_l^y + \Sigma_l^{yx}(\Sigma_l^{xx})^{-1}(z_{t-1} - \mu_l^x), \Sigma_l^{yy} - (\Sigma_l^{yx})^2(\Sigma_l^{xx})^{-1}\right) \tag{1}$$

with

$$q_l(z_{t-1}) = p_l \mathrm{N}(z_{t-1} \mid \mu_l^x, \Sigma_l^{xx}) / \left\{ \sum\nolimits_{m=1}^{L} p_m \mathrm{N}(z_{t-1} \mid \mu_m^x, \Sigma_m^{xx}) \right\}. \tag{2}$$

The transition density is therefore a location-scale mixture of normal transition densities, with means which depend on the previous state in a linear fashion, and weights which favor

mixture component l if z_{t-1} is near μ_l^x.

The model structure in (1) and (2) defines a flexible time-homogeneous, nonstationary Markov chain model. It allows for very general transition density shapes that can change flexibly across the state space, owing to the local adjustment provided by the mixture weights. The model also enables rich nonlinear dynamic relationships, which can be explored through, for instance, the conditional expectation

$$\mathrm{E}(Z_t \mid Z_{t-1} = z_{t-1}, G) = \sum_{l=1}^{L} q_l(z_{t-1})\{\mu_l^y + \Sigma_l^{yx}(\Sigma_l^{xx})^{-1}(z_{t-1} - \mu_l^x)\}.$$

This is a mixture of linear functions, but with state-dependent weights which can thus uncover nonlinear dynamics, in addition to non-Gaussian transition densities.

As discussed above, the transition density in (1) arises from the well-studied DP mixture of normals model. Conditional on an initial value z_1, the likelihood $\prod_{t=2}^{n} f(z_t \mid z_{t-1}, G)$ is a product of conditional densities, each being a mixture of normals. The mixture weights, given by (2), contain $\{\mu_l^x\}$ and $\{\Sigma_l^{xx}\}$ in the denominator, and each mixture component variance in (1) contains a complex function of the elements of Σ_l. Hence, with respect to posterior simulation, there does not exist a choice of G_0 which allows the full conditional distributions for μ_l^x, Σ_l^{xx}, Σ_l^{yy}, or Σ_l^{yx} to be recognizable as standard distributions.

These difficulties are alleviated to some extent by employing a square-root-free Cholesky decomposition of the covariance matrix Σ (Daniels and Pourahmadi, 2002; Webb and Forster, 2008; DeYoreo and Kottas, 2015), which expresses Σ in terms of a unit lower triangular matrix β and a diagonal matrix Δ with positive elements, such that $\Sigma = \beta^{-1}\Delta(\beta^{-1})^T$. The utility of this parametrization lies in the following property. If $(Y_1, \ldots, Y_m) \sim \mathrm{N}(\boldsymbol{\mu}, \beta^{-1}\Delta(\beta^{-1})^T)$, with $(\delta_1, \ldots, \delta_m)$ on the diagonal of Δ, then the joint distribution of Y can be expressed in a recursive form: $Y_1 \sim \mathrm{N}(\mu_1, \delta_1)$, and $(Y_k \mid Y_1, \ldots, Y_{k-1}) \sim \mathrm{N}(\mu_k - \sum_{j=1}^{k-1} \beta_{k,j}(y_j - \mu_j), \delta_k)$, for $k = 2, \ldots, m$. With this parameterization of the mixture kernel covariance matrix, the mixture transition density (1) admits the form

$$f(z_t \mid z_{t-1}, G) = \sum_{l=1}^{L} q_l(z_{t-1})\mathrm{N}(z_t \mid \mu_l^y - \beta_l(z_{t-1} - \mu_l^x), \delta_l^y) \tag{3}$$

with

$$q_l(z_{t-1}) = p_l \mathrm{N}(z_{t-1} \mid \mu_l^x, \delta_l^x) / \left\{ \sum\nolimits_{m=1}^{L} p_m \mathrm{N}(z_{t-1} \mid \mu_m^x, \delta_m^x) \right\} \tag{4}$$

where, in the case of the 2×2 covariance matrix Σ, β represents the only free element of the lower triangular matrix, and Δ has diagonal elements (δ^x, δ^y).

Let $\boldsymbol{\eta}_l = (\mu_l^x, \mu_l^y, \beta_l, \delta_l^x, \delta_l^y)$, for $l = 1, \ldots, L$, denote the mixing parameters. The mixture transition density can be broken by introducing latent configuration variables $\{U_2, \ldots, U_n\}$ taking values in $\{1, \ldots, L\}$, with $\Pr(U_t = l) = q_l(z_{t-1})$, such that the augmented hierarchical model for the data becomes:

$$\begin{aligned} z_t \mid z_{t-1}, U_t, \{\boldsymbol{\eta}_l\} &\overset{ind.}{\sim} \mathrm{N}(\mu_{U_t}^y - \beta_{U_t}(z_{t-1} - \mu_{U_t}^x), \delta_{U_t}^y), \quad t = 2, \ldots, n \\ U_t \mid z_{t-1}, \boldsymbol{p}, \{\boldsymbol{\eta}_l\} &\overset{ind.}{\sim} \sum_{l=1}^{L} \frac{p_l \mathrm{N}(z_{t-1} \mid \mu_l^x, \delta_l^x)}{\sum_{m=1}^{L} p_m \mathrm{N}(z_{t-1} \mid \mu_m^x, \delta_m^x)} I(U_t = l), \quad t = 2, \ldots, n \\ \boldsymbol{\eta}_l \mid \boldsymbol{\psi} &\overset{i.i.d.}{\sim} G_0(\boldsymbol{\eta}_l \mid \boldsymbol{\psi}), \quad l = 1, \ldots, L \end{aligned} \tag{5}$$

and the prior density for $\boldsymbol{p} = (p_1, \ldots p_L)$ is given by a special case of the generalized Dirichlet distribution: $f(\boldsymbol{p} \mid \alpha) = \alpha^{L-1} p_L^{\alpha-1} (1 - p_1)^{-1} (1 - (p_1 + p_2))^{-1} \times \cdots \times (1 - \sum_{l=1}^{L-2} p_l)^{-1}$ (Connor and Mosimann, 1969). The base distribution G_0 comprises independent components: $\mathrm{N}(m^x, v^x)$ and $\mathrm{N}(m^y, v^y)$ for μ_l^x and μ_l^y; $\mathrm{IG}(\nu^x, s^x)$ and $\mathrm{IG}(\nu^y, s^y)$ for δ_l^x and δ_l^y; and $\mathrm{N}(\theta, c)$ for β_l. This choice is conjugate for $\{\delta_l^y\}$, $\{\beta_l\}$, and $\{\mu_l^y\}$. The full Bayesian model is completed with conditionally conjugate priors on $\boldsymbol{\psi} = (m^x, v^x, m^y, v^y, s^x, s^y, \theta, c)$, the hyperparameters of G_0:

$$\begin{aligned} m^x \sim \mathrm{N}(a_m^x, b_m^x),\ m^y \sim \mathrm{N}(a_m^y, b_m^y),\ v^x \sim \mathrm{IG}(a_v^x, b_v^x),\ v^y \sim \mathrm{IG}(a_v^y, b_v^y), \\ s^x \sim \mathrm{Ga}(a_s^x, b_s^x),\ s^y \sim \mathrm{Ga}(a_s^y, b_s^y),\ \theta \sim \mathrm{N}(a_\theta, b_\theta),\ c \sim \mathrm{IG}(a_c, b_c) \end{aligned} \tag{6}$$

and a gamma prior for the DP precision parameter, $\alpha \sim \mathrm{Ga}(a_\alpha, b_\alpha)$.

2.2 Posterior Inference

Samples from the posterior distribution of model (5) are obtained using a combination of Gibbs sampling and Metropolis-Hastings steps. Details of the Markov chain Monte Carlo (MCMC) algorithm are provided in Appendix A, focusing particular attention on vector $(p_1, \ldots, p_L)$, which requires the most care in developing an effective sampling strategy.

Using the posterior samples, full inference is readily available for the transition density, $f(z_{t+1} \mid z_t, G)$, for any value of z_t. In particular, point and interval estimates can be obtained for the forecast distribution, $f(z_{n+1} \mid z_n, G) = \sum_{l=1}^{L} q_l(z_n)\mathrm{N}(z_{n+1} \mid \mu_l^y - \beta_l(z_n - \mu_l^x), \delta_l^y)$. The posterior mean estimate corresponds to the posterior predictive density for the next observation, since it can be shown that $p(z_{n+1} \mid \text{data}) = \mathrm{E}\{f(z_{n+1} \mid z_n, G) \mid \text{data}\}$. Point estimates for forecasts further than one step ahead may be obtained fairly easily, and entire distributions are also available, albeit at somewhat greater computational expense.

It may also be of interest to compare the predictive performance of the model with alternative models through one-step-ahead predictive distributions, $p(z_t \mid \boldsymbol{z}_{(t-1)})$. Here, $\boldsymbol{z}_{(m)}$ denotes the observed series up to time m, for $m = 2, ..., n$, such that $\boldsymbol{z}_{(n)}$ corresponds to the full data vector. As detailed in Appendix B, it is possible to compute the value of the posterior predictive density $p(z_t \mid \boldsymbol{z}_{(t-1)})$ at any observed z_t, using the posterior samples from fitting the model once to $\boldsymbol{z}_{(n)}$. We use these one-step-ahead posterior predictive ordinates to supplement graphical model comparison for the data example of Section 3.2.

2.3 Prior Specification

We discuss prior specification for the hyperparameters $\boldsymbol{\psi}$ of G_0, aiming to select appropriately diffuse priors which use only a small amount of prior information. Recall that $\mathrm{E}(Z_t \mid Z_{t-1} = z_{t-1}, G) = \sum_{l=1}^{L} q_l(z_{t-1})(\mu_l^y + \beta_l\mu_l^x - \beta_l z_{t-1})$. As a default approach, we assume that on average Z_{t-1} does not inform Z_t in the prior, so that $\mathrm{E}(\beta_l) = a_\theta = 0$. To fix b_θ, a_c, and b_c, the parameters contributing to $\mathrm{Var}(\beta_l)$, we note that the β_l parameters can be thought of as component-specific autoregressive coefficients, and that stationarity for each Gaussian mixture component requires $|\beta_l| < 1$. We thus select b_θ, a_c, and b_c such that $\mathrm{Var}(\beta_l) = b_\theta + \{b_c/(a_c - 1)\} = 1$ to favor a prior for the β_l that places most mass on values in

the stationary region, but also allows for nonstationarity in the mixture components. Here, we set a_c to a small value that ensures finite mean for the IG(a_c, b_c) prior distribution, and we follow a similar approach for the shape parameters of the other inverse-gamma priors.

Let d and r be a proxy for the center and range of the data, respectively. Based again on the conditional expectation, μ_l^y can be centered around d with variance that is consistent with the scale of the data; for instance, $a_m^y = d$ and $\text{Var}(\mu_l^y) = b_m^y + \{b_v^y/(a_v^y - 1)\} = (r/4)^2$. Similarly, δ_l^y can be centered at a value representing the data variance, that is, $\text{E}(\delta_l^y) = a_s^y\{b_s^y(\nu^y - 1)\}^{-1} = (r/4)^2$. With a_v^y, ν^y, and a_s^y fixed at relatively small values, b_m^y, b_v^y, and b_s^y can be specified from these expressions.

Finally, the μ_l^x and δ_l^x parameters correspond to the means and variances of the Gaussian densities that define the mixture weights in (4). The parameters δ_l^x control how quickly the weights decay as z_{t-1} gets farther from μ_l^x. Component l receives large weight for z_{t-1} around μ_l^x, with the relative weight decreasing according to a Gaussian distribution. Hence, using again the rough values for the center and range of the data, we set $\text{E}(\mu_l^x) = a_m^x = d$, $\text{Var}(\mu_l^x) = b_m^x + \{b_v^x/(a_v^x - 1)\} = (r/4)^2$, and $\text{E}(\delta_l^x) = a_s^x\{b_s^x(\nu^x - 1)\}^{-1} = (r/4)^2$.

2.4 Related Mixture Models for Time Series

Carvalho and Tanner (2005, 2006) model nonlinear time series through finite mixtures of generalized linear models, or experts, resulting in time series models with transition densities similar to (1). However, they approach the problem from a maximum likelihood perspective, and require the use of model selection criteria to determine the optimal size of the mixture. Wood et al. (2011) consider parametric mixture modeling for time series in which the weights are time-dependent and the lag is unknown.

While Bayesian nonparametric techniques have become extremely popular in density estimation, regression, and other applications, they have been used to a lesser extent in the context of time series. Müller et al. (1997) first made use of DP priors to build a model for nonstationary time series. They propose a finite mixture of AR models with local weights, where the parameters of the autoregressions and the parameters of the mixture weights arise from a random distribution which is assigned a DP prior. Tang and Ghosal (2007b)

establish posterior consistency for transition densities which can be expressed as DP mixtures of Gaussian AR kernels. Tang and Ghosal (2007a) consider a particular version of this class of models, involving a hyperbolic tangent transformation of lagged terms. Di Lucca et al. (2013) apply a dependent DP (DDP) mixture (MacEachern, 2000) for the transition densities, focusing mainly on the common weights version of the DDP. The DP atoms arise from a normal distribution with means linear on the previous observation. Their primary model is a location mixture of AR models, with mixing on the AR parameters. Mena and Walker (2005) construct structured transition densities to obtain strongly stationary AR models. Lau and So (2008) also considered DP mixtures of AR processes. Caron et al. (2008) and Fox et al. (2011) assume DP mixture errors within a DLM framework.

The proposed mixture model can be modified such that the Markov chain has a stationary distribution. In particular, consider the restricted version of the bivariate normal kernel for the joint DP mixture from which the transition density is defined, such that $\mu^x = \mu^y \equiv \mu$ and $\Sigma^{xx} = \Sigma^{yy} \equiv \sigma^2$. Then, it can be shown that the density $f(\cdot \mid G) = \sum_{l=1}^{L} p_l \mathrm{N}(\cdot \mid \mu_l, \sigma_l^2)$ satisfies $\int_A f(u \mid G)\mathrm{d}u = \int \{\int_A f(y \mid x, G)\mathrm{d}y\} f(x \mid G)\mathrm{d}x$, for all measurable $A \subset \mathbb{R}$, that is, $f(\cdot \mid G)$ is a stationary (invariant) density. This constraint yields transition density

$$f(z_t \mid z_{t-1}, G) = \sum_{l=1}^{L} q_l(z_{t-1}) \mathrm{N}(z_t \mid \mu_l - \beta_l(z_{t-1} - \mu_l), \sigma_l^2(1 - \beta_l^2)) \tag{7}$$

with $q_l(z_{t-1}) \propto p_l \mathrm{N}(z_{t-1} \mid \mu_l, \sigma_l^2)$, and $\beta_l \in (-1, 1)$. This is essentially the model studied by Antoniano-Villalobos and Walker (2016), although the version they implemented did not involve mixing over the scale parameter σ. The modeling framework of Antoniano-Villalobos and Walker (2016) begins much like ours, in that a transition mechanism is obtained as the conditional density from a bivariate mixture distribution. The authors do not apply a truncation approximation to the mixing distribution, and instead develop a posterior simulation method based on introduction of multiple sets of latent variables and a trans-dimensional MCMC algorithm. The model developed by Antoniano-Villalobos and Walker (2016) was previously proposed by Martinez-Ovando and Walker (2011), however it was then thought to be intractable due to the infinite sum appearing in the denominator of

the transition density mixture weights. Although we utilize a truncation approximation to the DP, the sum in the denominator of the weights in (2) still presents challenges in terms of posterior simulation. We develop a tractable MCMC algorithm by reparameterizing the covariance matrices in $f(x, y \mid G)$ and working with the stick-breaking weights to develop a slice sampler which indirectly provides samples for $\boldsymbol{p}$ (see Appendix A).

We refer to the special case of the nonparametric mixture model discussed above as the "stationary" mixture model. However, it is important to note that the particular restriction ensures existence of a stationary distribution, but not its uniqueness, which would be required to develop conditions for additional properties of the stationary Markov chain, such as ergodicity. Some results in this direction are studied in Carvalho and Tanner (2005) under the mixture of experts formulation for the transition density.

3 Data Illustrations

We now illustrate the proposed model on two simulated data sets (Section 3.1) and apply it to the waiting times between eruptions of the Old Faithful geyser (Section 3.2). For the real data example, we also consider comparison with a parametric TAR model and with the stationary mixture model discussed in Section 2.4 as a special case of the proposed model.

In all cases, MCMC inference for the nonparametric mixture model was implemented in R, saving every 20-th iteration after burn-in, and with a posterior sample of $5,000$ used for inference. For the Old Faithful data example, for which $n = 272$, it took about 2.5 hours to collect $50,000$ posterior samples (without particular emphasis on optimizing the MCMC code). We follow the approach to prior specification described in Section 2.3.

The DP truncation level L is specified using the expectation of the partial sum of the original DP weights, $\mathrm{E}(\sum_{l=1}^{L} p_l \mid \alpha) = 1 - \{\alpha/(\alpha+1)\}^L$. This expression can be averaged over the prior for α to estimate the marginal prior expectation $\mathrm{E}(\sum_{l=1}^{L} p_l)$, which is used to specify L given any desired tolerance level for the approximation. For instance, under a gamma$(0.5, 0.5)$ prior on α, $\mathrm{E}(\sum_{l=1}^{L} p_l)$ is 0.9997 with $L = 30$ and 0.99999 with $L = 50$. We used a value of L in this range for all data examples, and monitored the number of

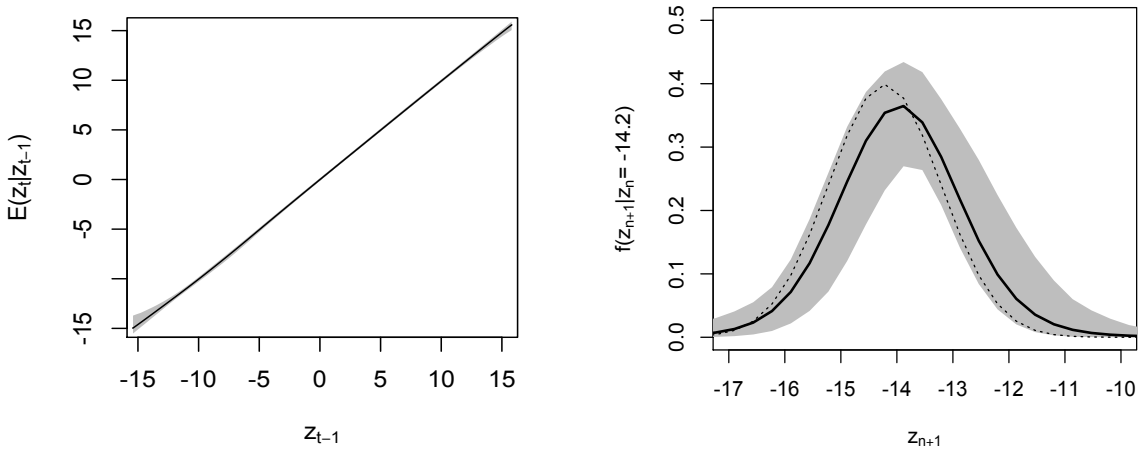

Figure 1: Brownian motion simulation. The left panel plots the posterior mean estimate (solid line) and 95% credible intervals (gray shaded region) for $\mathrm{E}(Z_t \mid Z_{t-1} = z_{t-1})$ plotted over a grid in z_{t-1}. The true expectation is indistinguishable from the model's estimate. The right panel shows the posterior mean (solid line) and 95% credible intervals (gray shaded region) for the forecast density, $f(z_{n+1} \mid z_n = -14.2)$, compared to the truth (dotted line).

effective components to ensure it never reached the upper bound.

3.1 Simulated Data

We first consider a data set generated from standard Brownian motion to test the model in a nonstationary setting, albeit with linear Gaussian transition densities (Section 3.1.1). Next, we demonstrate the capacity of the model to uncover non-linear, non-Gaussian dynamics, using synthetic data from skew-normal transition densities with varying skewness and dispersion (Section 3.1.2).

3.1.1 Brownian Motion

Standard Brownian motion is a nonstationary process defined by the transition density $f(z_t \mid z_{t-1}) = \mathrm{N}(z_{t-1}, 1)$. A standard Brownian motion path is generated assuming $n = 500$. Trivially, $\mathrm{E}(Z_t \mid Z_{t-1} = z_{t-1}) = z_{t-1}$ in this model. The inference from the model indicates it is detecting this trend with little uncertainty (Figure 1, left panel). The value of the last observation is -14.2, one of the smallest values in the entire series. The forecast density

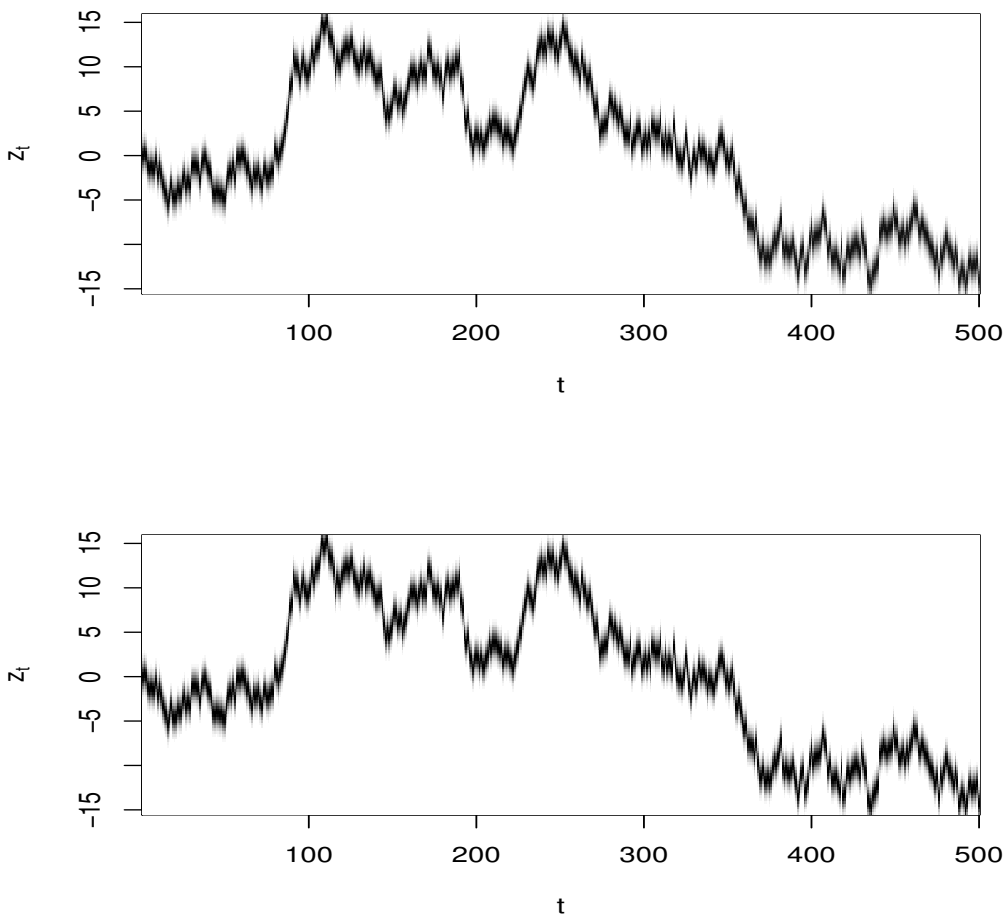

Figure 2: Brownian motion simulation. For each $t = 2, ..., n$, the top panel shows the posterior mean estimates for $f(z_t \mid z_{t-1})$, and the bottom panel plots the corresponding true densities. Darker colors indicate larger density values. Refer to Section 3.1.1 for details.

for the next observation is displayed in Figure 1 (right panel). While the 95% posterior credible intervals contain the true density, the mode of the point estimate favors slightly larger values, likely due to the fact that -14.2 is an extreme value in this series.

The top panel of Figure 2 plots the posterior mean estimates for $f(z_t \mid z_{t-1})$ for each $t = 2, ..., n$ and for the corresponding observed z_{t-1}. In particular, for each index t on the horizontal axis, $\mathrm{E}\{f(z_t \mid z_{t-1}, G) \mid \text{data}\}$ is plotted on the vertical axis such that darker colors represent larger density values. The associated true densities $f(z_t \mid z_{t-1})$, again given the observed z_{t-1}, are plotted in the bottom panel of Figure 2.

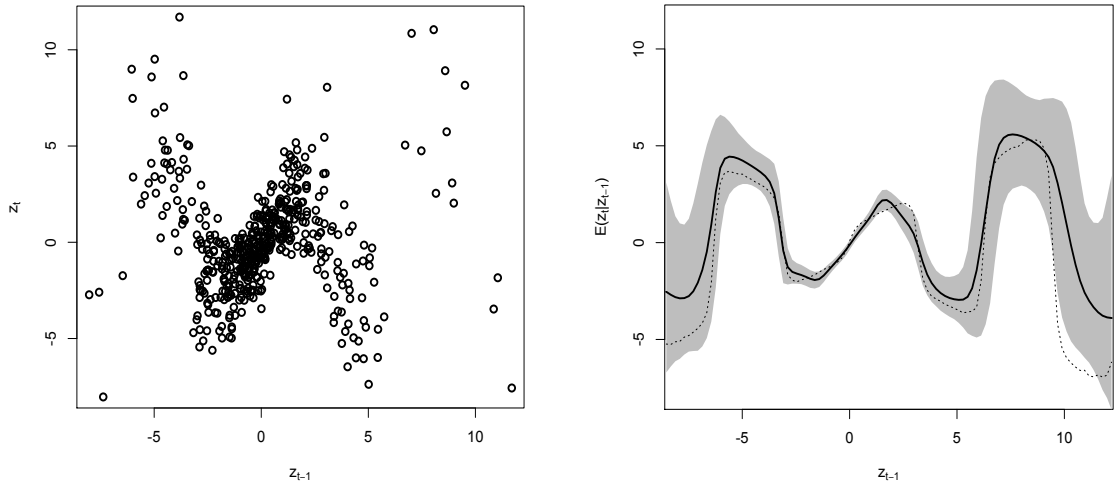

Figure 3: Skew-normal simulation. The left panel plots the simulated data as pairs of points (z_{t-1}, z_t). The right panel shows the posterior mean (solid line) and 95% credible intervals (gray shaded region) for $\mathrm{E}(Z_t \mid Z_{t-1} = z_{t-1})$ plotted over a grid in z_{t-1}; the true expectation is shown as a dotted line.

In summary, all visual displays indicate that the model is capturing the dynamics quite well, even though its transition densities are substantially less structured than the transition densities of the underlying Brownian motion.

3.1.2 Skew-normal Transition Densities

To generate a time series that exhibits challenging transition densities which evolve over time in a non-standard fashion, we assume each observation is generated from a skew-normal distribution (Azzalini, 1985), with scale and skewness parameters which are functions of the previous observation. In particular, we generate $z_t \mid z_{t-1} \sim \mathrm{SN}(z_t \mid 0, 1+0.7|z_{t-1}|, 0.1+4\sin(z_{t-1}))$, for $t = 2, \ldots, n$. Here, $\mathrm{SN}(y \mid \xi, \omega, \alpha)$ denotes the skew-normal distribution with density $(\omega\pi)^{-1} \exp\{-(y-\xi)^2/(2\omega^2)\}\Phi(\alpha(x-\xi)/\omega)$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function. The sinusoidal or periodic trend in skewness parameter α yields conditional distributions with various directions and degrees of skewness, and the decreasing followed by increasing linear trend in scale parameter ω leads to distributions which are more peaked when z_{t-1} is near 0.

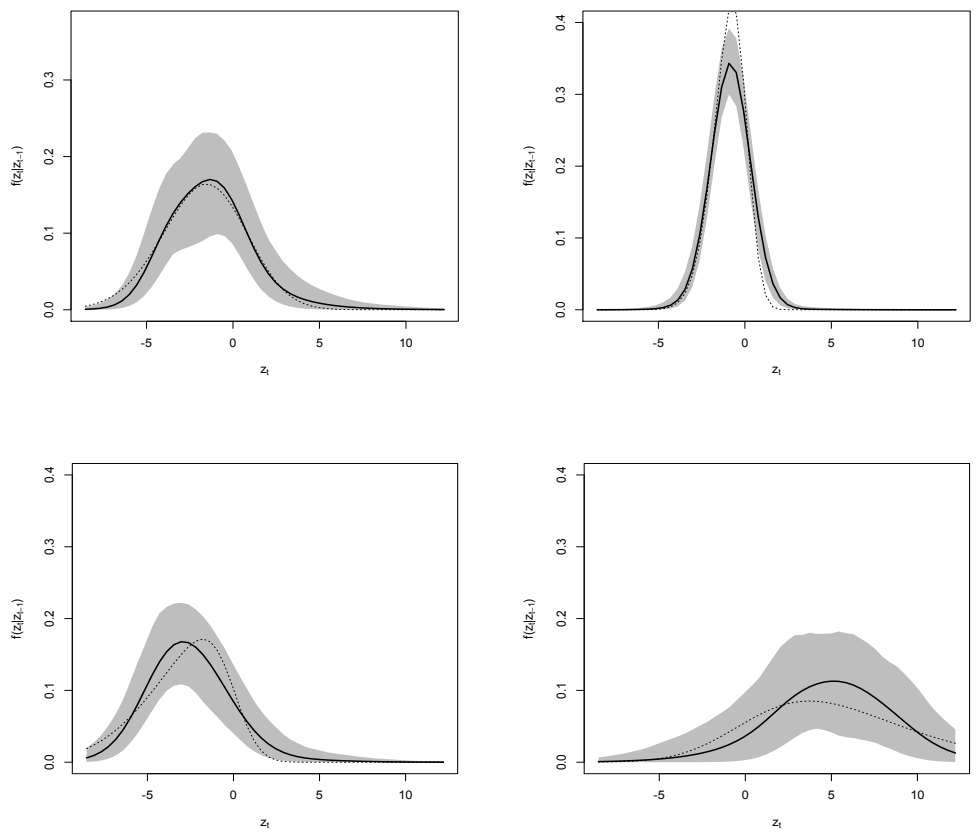

Figure 4: Skew-normal simulation. Posterior mean (solid line) and 95% credible intervals (gray shaded region) for transition densities $f(z_t \mid z_{t-1})$, for $z_{t-1} = -2.85$ (top left), $z_{t-1} = -0.5$ (top right), $z_{t-1} = 4.2$ (bottom left), and $z_{t-1} = 8.85$ (bottom right). The corresponding true densities are plotted as dotted lines.

A time series $(z_2, \ldots, z_{500})$ was simulated from this model assuming an initial value $z_1 = 0$. Figure 3 (left panel) shows the simulated data $\{(z_{t-1}, z_t),\, t = 2, \ldots, 500\}$. Notice the oscillating trend in location, and the larger variation in z_t for z_{t-1} far from 0. Figure 3 (right panel) plots posterior mean and interval estimates for $\mathrm{E}(Z_t \mid Z_{t-1} = z_{t-1})$ along with the data-generating expectation trend. The point estimate captures successfully the overall non-linear trend, and the 95% credible intervals contain the truth everywhere except for a small region around $z_{t-1} = 10$, where there is very little data.

In this case, the true densities $f(z_t \mid z_{t-1})$ do not depict a strong trend analogous to the one in Figure 2, but the model was again successful in capturing the evolution of the skew-normal transition densities through the corresponding posterior mean estimates (results

now shown). To demonstrate the capacity of the model to uncover varying density shapes and the corresponding uncertainty quantification, in Figure 4 we display point estimates and 95% uncertainty bands for $f(z_t \mid z_{t-1})$ at four particular values of z_{t-1}. Notice the wide uncertainty bands for the density at $z_{t-1} = 8.85$ (bottom right panel) and the narrow uncertainty bands when $z_{t-1} = -0.5$ (top right panel), which reflects the lack of data above $z_{t-1} = 5$ and the large amount of data in the region near $z_{t-1} = 0$.

3.2 Waiting Times Between Eruptions of the Old Faithful Geyser

For our real data illustration, we consider the time intervals between successive eruptions of the Old Faithful geyser. The specific data set is available through R (dataset `faithful`) and it consists of 272 measurements $\{z_t,\, t = 1, \ldots, 272\}$, where z_t represents the waiting time in minutes before eruption t. The data are included in Figure 6 in the form of a plot of z_t versus z_{t-1}, for $t = 2, \ldots, 272$.

There are some interesting features present in the data. When z_{t-1} is below 60, there is a large cluster of points around $z_t = 80$, and a small number of points extending down to about $z_t = 50$, indicating a distribution with a mode near 80 but with a heavy left tail or a small additional mode near 50. Moving to larger values of z_{t-1}, there are two clusters of points, one centered around 55 and one around 80. These features are captured by the mixture model in (3) and (4), as shown in Figure 5 with the estimated transition densities $f(z_t \mid z_{t-1} = 50)$ and $f(z_t \mid z_{t-1} = 80)$. Moreover, inference for $\mathrm{E}(Z_t \mid Z_{t-1} = z_{t-1})$ is shown in Figure 6 (right panel), and the posterior mean estimate and 95% credible intervals for the forecast density, $f(z_{n+1} \mid z_n = 74)$, are given in Figure 7 (right panel). The estimated forecast density has a primary mode near 80 and a heavy left tail with a suggestion of additional modes around 55 and 65; this is a plausible shape for the forecast density given the cross-section of data around the last observation $z_{272} = 74$.

Next, we discuss results from comparison with a parametric model and with the special case of the proposed model that incorporates the stationarity restriction. The stationary mixture model, given in (7), was implemented by appropriately modifying the MCMC algorithm described in Appendix A, and with priors that were comparable to the ones used

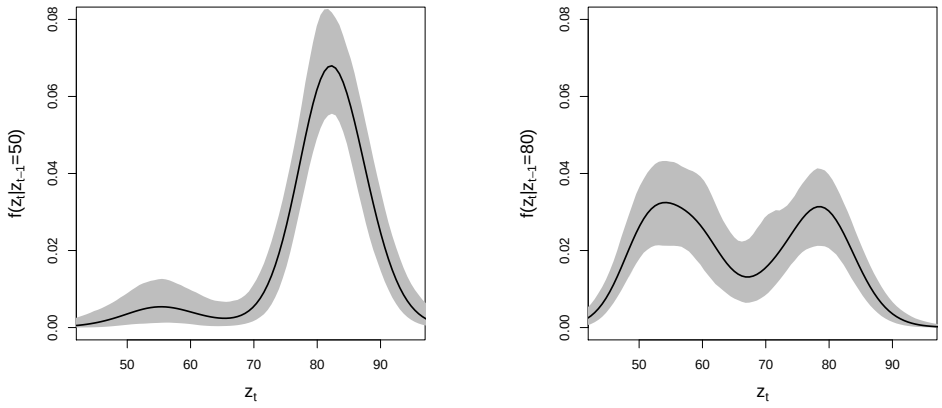

Figure 5: Old Faithful data. Posterior mean (solid line) and 95% credible intervals (gray shaded region) for transition densities $f(z_t \mid z_{t-1})$, at $z_{t-1} = 50$ (left panel) and $z_{t-1} = 80$ (right panel), under the general mixture model.

for the general mixture model. Regarding the choice of a parametric model for comparison, inspection of the data suggests the TAR model structure as a plausible, simpler alternative to capture the nonlinear dynamics in $\mathrm{E}(Z_t \mid Z_{t-1} = z_{t-1})$. We thus consider a Gaussian-based TAR model with threshold that depends on the previous value in the time series, and with two regimes. More specifically, $z_t \mid z_{t-1} \sim \mathrm{N}(\phi_0^{(1)} + \phi_1^{(1)} z_{t-1}, \tau^{(1)})$ if $z_{t-1} \leq r$, and $z_t \mid z_{t-1} \sim \mathrm{N}(\phi_0^{(2)} + \phi_1^{(2)} z_{t-1}, \tau^{(2)})$ if $z_{t-1} > r$. The model was implemented with the R package BAYSTAR, "Bayesian analysis of threshold autoregressive models". We used data-based, informative priors, in particular: the prior on the intercept parameters was normal centered at the midpoint of the time series, with variance equal to the approximate variance of the time series; the AR coefficient parameters were given normal priors with mean 0 and variance 2; and the τ parameters were assigned inverse-gamma priors centered at the residual mean square error of fitting an AR(1) model to the data, and with small shape parameters. The posterior means of the AR coefficients were 0.17 and -0.68, and the posterior mean for the threshold r was 64.4.

Inference results for $\mathrm{E}(Z_t \mid Z_{t-1} = z_{t-1})$ and for the forecast density are reported in Figures 6 and 7, respectively. Focusing first on the conditional expectation estimates, the TAR model uncovers a nonlinear shape that is overall comparable to the one estimated by

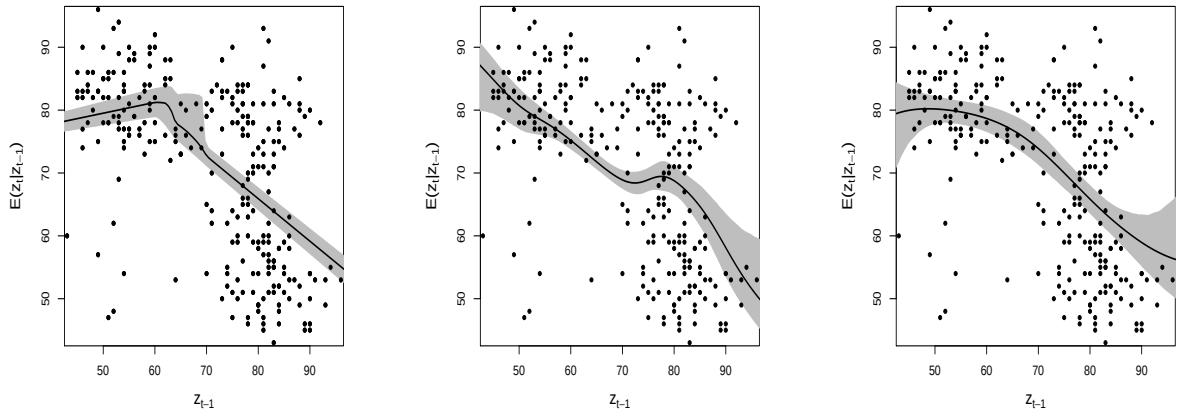

Figure 6: Old Faithful data. Posterior mean (solid line) and 95% credible intervals (gray shaded region) for $\mathrm{E}(Z_t \mid Z_{t-1} = z_{t-1})$ plotted over a grid in z_{t-1}, under the TAR model (left panel), the stationary mixture model (middle panel), and the general mixture model (right panel). Included in each panel are the data shown as pairs of points (z_{t-1}, z_t), for $t = 2, \ldots, 272$.

the general mixture model, although the latter produces a smoother point estimate and uncertainty bands that increase at the data boundaries. The stationary mixture model also yields more plausible uncertainty quantification than the parametric model, but estimates a nonlinearity for $\mathrm{E}(Z_t \mid Z_{t-1} = z_{t-1})$ at a range of z_{t-1} values that is distinctly different from the other two models. Regarding the forecast density estimates, the TAR model is unable to capture the non-standard shape uncovered by the general mixture model. The stationary mixture model estimates a bimodal forecast density, but with a significant difference in the magnitude of the peaks relative to the unrestricted mixture model; in particular, the more pronounced mode around 65 does not seem to be compatible with the cross-section of data around $z_{272} = 74$. The superior predictive performance of the general mixture model is further supported by one-step-ahead predictions. Using the approach described in Appendix B, we computed the one-step-ahead posterior predictive ordinates for the last 90 observations (about 1/3 of the observed time series). The sum of the log-ordinates was -364.9 for the stationary mixture model, and -327.4 for the unrestricted mixture model. The corresponding value for the TAR model was -344.0, that is, based on this criterion, the parametric model performs better than the stationary mixture model.

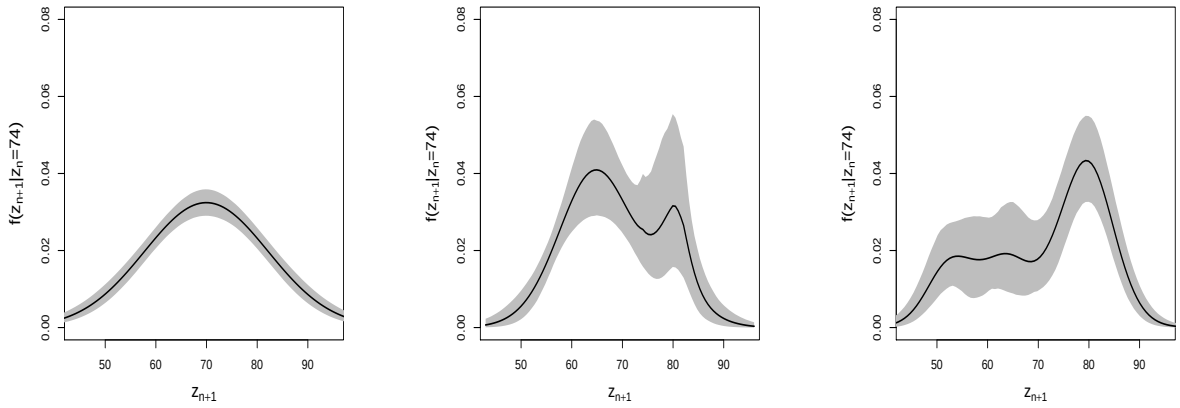

Figure 7: Old Faithful data. Posterior mean (solid line) and 95% credible intervals (gray shaded region) for the forecast density, $f(z_{n+1} \mid z_n = 74)$, under the TAR model (left panel), the stationary mixture model (middle panel), and the general mixture model (right panel).

For problems where one has information regarding stationarity of the data generating process, the stationary mixture model may provide a natural starting point for the analysis. In fact, as demonstrated in Antoniano-Villalobos and Walker (2016), this model is able to estimate effectively transition densities from nonstationary processes, which however are driven by standard distributions. This was confirmed by a reanalysis of the Brownian motion simulated data example – which is also one of the examples of Antoniano-Villalobos and Walker (2016) – for which we obtained from the stationary mixture model results very similar to the ones reported in Section 3.1.1. However, our experience suggests that the stationary mixture model may not be sufficiently flexible for settings that involve transition densities with non-standard shapes. This is not surprising upon inspecting the model structure in (7), and contrasting it with (3) and (4). In particular, note that the stationarity restriction forces a single set of mixing parameters μ_l used to inform both the means of the Gaussian AR mixture components and the locations of the associated mixture weights. Moreover, the σ_l^2 control the dispersion of both the Gaussian mixture components and of the Gaussian densities that define the mixture weights.

4 Extensions

The data illustrations suggest the ability of the first-order model to uncover a variety of conditional density shapes, and approximate well the truth contained in simulated data. However, some applications may require additional features in the model formulation.

The first-order model can be extended to accommodate higher order structure (say, based on r lagged terms), where again the transition density, $f(z_t \mid z_{t-1}, \ldots, z_{t-r})$, is implied by a joint DP mixture model. Hence, the transition density has a similar form to (1), but now the means of the Gaussian mixture components and the mixture weights depend on the previous r states. Let superscript y correspond to Z_t and x to $(Z_{t-r}, \ldots, Z_{t-1})$ in the vector $\boldsymbol{\mu}$ of length $r+1$ and the $(r+1) \times (r+1)$ matrix Σ. Under the reparameterization of Σ used in the first-order case, the Gaussian mixture kernels have the form $\mathrm{N}(z_t \mid \mu_l^y - \sum_{j=1}^{r} \beta_{l,(r+1,j)}(z_{t+j-r-1} - \mu_{l,j}^x), \delta_l^y)$, for $l = 1, \ldots, L$. Gibbs sampling steps are thus preserved for μ_l^y and δ_l^y, as well as for the last row of the matrix β. However, more care is needed in devising an MCMC algorithm to sample $\boldsymbol{\delta}_l^x$, $\boldsymbol{\mu}_l^x$ (each a vector of length r) and the first r rows of β, particularly when r is of order larger than 2 or 3.

Turning to an application oriented extension, in population biology, the size of a wild population is often monitored over time. Yearly estimated biomass may be recorded for a specific species, and the trend in population size indicates how the species is faring, and is indicative of greater environmental conditions. A state-space modeling framework is suitable for such applications, since the observed biomass is not an exact measurement of population size. Rather, biomass is viewed as a noisy version of the underlying population size, and a key goal is to forecast population size in the future. The proposed model can be incorporated into a state-space framework, with the addition of an observation equation. The observations are now viewed as arising from latent unobserved states, which evolve in time according to the Markovian model. Denote the observed data by $(y_1, \ldots, y_n)$, and the underlying latent states by $(z_1, \ldots, z_n)$. Assume $y_t \mid z_t, \boldsymbol{\theta} \sim f(y_t \mid z_t, \boldsymbol{\theta})$, for some parametric distribution $f(y_t \mid z_t, \boldsymbol{\theta})$, with the latent states evolving according to the nonparametric mixture model for $f(z_t \mid z_{t-1})$. In the population dynamics example,

environmental covariates may also be available. These can be treated as random, and modeled jointly with y_t at the observation level, or incorporated at the state level.

The introduction of latent states is also useful in modeling ordinal time series data, as it is often assumed that $Y_t = j$ if and only if $Z_t \in (\gamma_{j-1}, \gamma_j)$, for $j = 1, \ldots, C$. However, rather than working with a restrictive parametric distribution for the latent continuous responses, they can be modeled with the proposed nonparametric Markovian model.

5 Summary

We have proposed a modeling approach for nonstationary time series which allows for non-standard transition densities and nonlinear autoregressions. The transition density of the Markovian model admits a representation as a location-scale mixture of normal densities, with means and mixture weights that depend on observations from previous time points. This model structure arises from the conditional distribution induced from a Dirichlet process mixture of multivariate normals. We have discussed methods for posterior inference and prior specification, and illustrated the model with synthetic and real data,

including comparison with a special case of the mixture model that ensures existence of a stationary distribution for the Markov chain. Although the methodology has been developed and applied for directly observable time series with first-order dependence, we have discussed possible extensions to model higher order Markov chains, and to expand the model structure to a state-space setting.

Acknowledgments

The work of the first author was supported by the National Science Foundation under award SES 1131897. The work of the second author was supported in part by the National Science Foundation under awards DMS 1310438 and DMS 1407838. The authors wish to thank two reviewers for constructive feedback and for comments that improved the presentation of the material in the paper.

Appendix A. The Markov chain Monte Carlo algorithm

Here, we provide the details of the MCMC method for posterior simulation from the non-parametric mixture model developed in Section 2.1.

The posterior full conditional distributions for α and the components of vector ψ are standard as they are assigned conditionally conjugate priors. Each U_t, $t = 2, \ldots, n$ is sampled from a discrete distribution on $\{1, \ldots, L\}$, with probabilities $(\tilde{p}_{1,t}, \ldots, \tilde{p}_{L,t})$, where $\tilde{p}_{l,t} \propto p_l \mathrm{N}(z_t \mid \mu_l^y - \beta_l(z_{t-1} - \mu_l^x), \delta_l^y)\mathrm{N}(z_{t-1} \mid \mu_l^x, \delta_l^x)$, for $l = 1, \ldots, L$.

Next, consider the mixing parameters. Letting $\{U_j^* : j = 1, \ldots, n^*\}$ be the n^* distinct values of $(U_2, \ldots, U_n)$, and $M_l = |\{U_t : U_t = l\}|$, we obtain the full conditional

$$p(\boldsymbol{\eta}_l \mid \ldots, \text{data}) \propto G_0(\boldsymbol{\eta}_l \mid \boldsymbol{\psi}) \left\{ \prod_{j=1}^{n^*} \prod_{\{t:U_t=U_j^*\}} \mathrm{N}(z_t \mid \mu_l^y - \beta_l(z_{t-1} - \mu_l^x), \delta_l^y) \right\} \left\{ \prod_{r=1}^{L} \prod_{\{t:U_t=r\}} q_r(z_{t-1}) \right\}.$$

Therefore, if $l \in \{U_j^*\}$, μ_l^y is sampled from a normal distribution with variance $(v^y)^* = [(\nu^y)^{-1} + M_l(\delta_l^y)^{-1}]^{-1}$, and mean $(v^y)^*[(\nu^y)^{-1}m^y + (\delta_l^y)^{-1}\sum_{\{t:U_t=U_j^*\}}(z_t + \beta_l(z_{t-1} - \mu_l^x))]$. If component l is empty, that is, $l \notin \{U_j^*\}$, then $\mu_l^y \sim \mathrm{N}(m^y, v^y)$. The updates for δ_l^y and β_l also require only Gibbs sampling. If $l \in \{U_j^*\}$, then $\delta_l^y \sim \mathrm{IG}(\nu^y + 0.5M_l, s^y + 0.5\sum_{\{t:U_t=l\}}(z_t - \mu_l^y + \beta_l(z_{t-1} - \mu_l^x))^2)$ and β_l is sampled from a normal with variance $c^* = [c^{-1} + (\delta_l^y)^{-1}\sum_{\{t:U_t=l\}}(z_{t-1} - \mu_l^x)^2]^{-1}$ and mean $c^*[c^{-1}\theta + (\delta_l^y)^{-1}\sum_{\{t:U_t=l\}}(z_{t-1} - \mu_l^x)(\mu_l^y - z_t)]$. If $l \notin \{U_j^*\}$, then we sample from G_0: $\delta_l^y \sim \mathrm{IG}(\nu^y, s^y)$ and $\beta_l \sim \mathrm{N}(\theta, c)$.

No matter the choice of G_0, the full conditionals for μ_l^x and δ_l^x are not proportional to any standard distribution, as these parameters are contained in the sum of L terms in the denominator of $q_l(z_{t-1})$. The posterior full conditional $p(\mu_l^x \mid \ldots, \text{data})$, when $l \in \{U_j^*\}$, is given by

$$\mathrm{N}\left(\mu_l^x \mid m^x, v^x\right) \prod_{\{t:U_t=l\}} \mathrm{N}\left(z_t \mid \mu_l^y - \beta_l(z_{t-1} - \mu_l^x), \delta_l^y\right) \mathrm{N}(z_{t-1} \mid \mu_l^x, \delta_l^x) \left(\prod_{t=2}^{n} \sum_{m=1}^{L} p_m \mathrm{N}(z_{t-1} \mid \mu_m^x, \delta_m^x) \right)^{-1}.$$

This can be written as $p(\mu_l^x | \ldots, \text{data}) \propto \mathrm{N}(\mu_l^x \mid (m^x)^*, (v^x)^*)(\prod_{t=2}^{n}\sum_{m=1}^{L} p_m \mathrm{N}(z_{t-1} \mid \mu_m^x, \delta_m^x))^{-1}$, with $(v^x)^* = ((v^x)^{-1} + M_l(\delta_l^x)^{-1} + M_l\beta_l^2(\delta_l^y)^{-1})$ and $(m^x)^* = (v^x)^*((v^x)^{-1}m^x +$

$(\delta_l^x)^{-1}\sum_{\{t:U_t=l\}} z_{t-1} + (\delta_l^y)^{-1}\beta_l^2\sum_{\{t:U_t=l\}}(z_{t-1} + (z_t - \mu_l^y)/\beta_l))$. We use a random-walk Metropolis step to update μ_l^x. For $l \notin \{U_j^*\}$, $p(\mu_l^x \mid \ldots, \text{data})$ is proportional to $\text{N}(\mu_l^x \mid m^x, v^x)[\prod_{t=2}^n \sum_{m=1}^L p_m \text{N}(z_{t-1} \mid \mu_m^x, \delta_m^x)]^{-1}$, and in this case we use a Metropolis-Hastings algorithm, proposing a candidate value μ_l^x from the base distribution $\text{N}(m^x, v^x)$.

The full conditional and sampling strategy for δ_l^x are similar to those for μ_l^x. We have

$$p(\delta_l^x \mid \ldots, \text{data}) \propto \text{IG}(\delta_l^x \mid \nu^x, s^x) \prod_{\{t:U_t=l\}} \text{N}(z_{t-1} \mid \mu_l^x, \delta_l^x) \left(\prod_{t=2}^n \sum_{m=1}^L p_m \text{N}(z_{t-1} \mid \mu_m^x, \delta_m^x)\right)^{-1},$$

which for an active component, is written as proportional to

$$\text{IG}\left(\delta_l^x \mid \nu^x + 0.5M_l, s^x + 0.5\sum_{\{t:U_t=l\}} (z_{t-1} - \mu_l^x)^2\right)\left(\prod_{t=2}^n \sum_{m=1}^L p_m \text{N}(z_{t-1} \mid \mu_m^x, \delta_m^x)\right)^{-1}.$$

For non-active components, the full conditional is $\text{IG}(\delta_l^x \mid \nu^x, s^x)(\prod_{t=2}^n \sum_{m=1}^L p_m \text{N}(z_{t-1} \mid \mu_m^x, \delta_m^x))^{-1}$. We use a similar strategy for sampling δ_l^x as we did with μ_l^x, using a random-walk Metropolis algorithm for the active components of δ_l^x, working on the log-scale and sampling $\log(\delta_l^x)$, and proposing the non-active components from $G_0(\delta_l^x) = \text{IG}(\nu^x, s^x)$.

We next discuss the updating scheme for the vector $\boldsymbol{p} = (p_1, \ldots, p_L)$, which poses the main challenge for posterior simulation. The full conditional for $\boldsymbol{p}$ has the form

$$f(\boldsymbol{p} \mid \alpha)\prod_{l=1}^L p_l^{M_l}\left(\prod_{t=2}^n \sum_{m=1}^L p_m \text{N}(z_{t-1} \mid \mu_m^x, \delta_m^x)\right)^{-1}.$$

In standard DP mixture models, the implied generalized Dirichlet prior for $f(\boldsymbol{p} \mid \alpha)$ combines with $\prod_{l=1}^L p_l^{M_l}$ to form another generalized Dirichlet distribution. However, in this case there is an additional term. Metropolis–Hastings algorithms with various proposal distributions were explored to sample the vector p, resulting in very low acceptance rates. We instead devise an alternative sampling scheme, in which we work directly with the latent beta-distributed random variables which determine the probability vector $\boldsymbol{p}$ arising from the DP truncation approximation. Recall that $p_1 = v_1$, $p_l = v_l\prod_{r=1}^{l-1}(1 - v_r)$, for $l = 2, \ldots, L-1$, and $p_L = \prod_{r=1}^{L-1}(1 - v_r)$, where $v_1, \ldots, v_{L-1} \overset{i.i.d.}{\sim} \text{beta}(1, \alpha)$. Equiva-

lently, let $\zeta_1, \ldots, \zeta_{L-1} \overset{i.i.d.}{\sim} \text{beta}(\alpha, 1)$, and define $p_1 = 1 - \zeta_1$, $p_l = (1 - \zeta_l) \prod_{r=1}^{l-1} \zeta_r$, and $p_L = \prod_{r=1}^{L-1} \zeta_r$. Rather than updating directly $\boldsymbol{p}$, we work with the ζ_l, a sample for which implies a particular probability vector $\boldsymbol{p}$.

The full conditional for ζ_l, $l = 1, \ldots, L-1$, has the form

$$p(\zeta_l \mid \ldots, \text{data}) \propto \text{beta}\left(\zeta_l \mid \alpha + \sum_{r=l+1}^{L} M_r, M_l + 1\right) \left(\prod_{t=2}^{n} d(z_{t-1})\right)^{-1} \tag{A.1}$$

where

$$d(z_{t-1}) = \text{N}(z_{t-1} \mid \mu_1^x, \delta_1^x)(1-\zeta_1) + \sum_{l=2}^{L-1} \text{N}(z_{t-1} \mid \mu_l^x, \delta_l^x)(1-\zeta_l) \prod_{s=1}^{l-1} \zeta_s + \text{N}(z_{t-1} \mid \mu_L^x, \delta_L^x) \prod_{s=1}^{L-1} \zeta_s.$$

Also, let $c_{t,l} = \text{N}(z_{t-1} \mid \mu_l^x, \delta_l^x)$, which is constant with respect to each ζ_l. The form of the full conditional in (A.1) suggests the use of a slice sampler to update each ζ_l one at a time. The slice sampler is implemented by drawing auxiliary random variables $u_t \sim \text{uniform}(0, (d(z_{t-1}))^{-1})$, $t = 2, ..., n$, and then sampling $\zeta_l \sim \text{beta}(\alpha + \sum_{r=l+1}^{L} M_r, M_l + 1)$, but restricted to the set $\{\zeta_l : u_t < (d(z_{t-1}))^{-1}, t = 2, ..., n\}$. The term $d(z_{t-1})$ can be expressed as $d(z_{t-1}) = \zeta_l w_{1t} + w_{0t}$, for any $l = 1, ..., L-1$, where

$$w_{1t} = -c_{t,l} \prod_{s=1}^{l-1} \zeta_s + \left(\sum_{m=l+1}^{L-1} c_{t,m}(1 - \zeta_m) \prod_{s=1, s\neq l}^{m-1} \zeta_s \right) + c_{t,L} \prod_{s=1, s\neq l}^{L-1} \zeta_s$$

and, if $l = 1$, $w_{0t} = c_{t,1}$, otherwise $w_{0t} = c_{t,1}(1-\zeta_1) + \sum_{s=2}^{l-1} c_{t,s}(1-\zeta_s) \prod_{r=1}^{s-1} \zeta_r + c_{t,l} \prod_{s=1}^{l-1} \zeta_s$. Then, the set $\{\zeta_l : d(z_{t-1}) < u_t^{-1}\}$ is $\{\zeta_l : \zeta_l w_{1t} < u_t^{-1} - w_{0t}\}$. This takes the form of $\{\zeta_l : \zeta_l < (u_t w_{1t})^{-1} - w_{0t}(w_{1t})^{-1}\}$ when w_{1t} is positive, and has the form $\{\zeta_l : \zeta_l > (u_t w_{1t})^{-1} - w_{0t}(w_{1t})^{-1}\}$ otherwise. Therefore, the truncated–beta random draw for ζ_l must lie in the interval $(\max_{\{t:w_{1t}<0\}}[(u_t w_{1t})^{-1} - w_{0t}(w_{1t})^{-1}], \min_{\{t:w_{1t}>0\}}[(u_t w_{1t})^{-1} - w_{0t}(w_{1t})^{-1}])$. The inverse CDF random variate generation method can be used to sample from these truncated beta random variables. This strategy results in direct draws for the ζ_l.

Appendix B. Computing posterior predictive ordinates

We describe here an approach to computing one-step-ahead posterior predictive ordinates, $p(z_t \mid \boldsymbol{z}_{(t-1)})$, where $\boldsymbol{z}_{(m)} = (z_2, ..., z_m)$, for $m = 2, ..., n$, is the observed series up to time m. The objective is to compute $p(z_t \mid \boldsymbol{z}_{(t-1)})$ for any desired number of observations z_t, using the samples from the posterior distribution given the full data vector $\boldsymbol{z}_{(n)}$.

Denote by $\boldsymbol{\Theta} = (\{\boldsymbol{\eta}_l : l = 1, ..., L\}, \boldsymbol{p}, \alpha, \boldsymbol{\psi})$ all model parameters, excluding the latent configuration variables. We abbreviate $f(z_t \mid z_{t-1}, G)$ in (3) to $f(z_t \mid z_{t-1})$, but note that, given the $\boldsymbol{\eta}_l$ and $\boldsymbol{p}$, the mixture model for the transition density can be computed at any values z_t and z_{t-1}. Let $B_{(m)}$ be the normalizing constant of the posterior distribution for $\boldsymbol{\Theta}$ given $\boldsymbol{z}_{(m)}$, and $p(\boldsymbol{\Theta}) = \{\prod_{l=1}^{L} G_0(\boldsymbol{\eta}_l \mid \boldsymbol{\psi})\} f(\boldsymbol{p} \mid \alpha) p(\alpha) p(\boldsymbol{\psi})$ be the prior for $\boldsymbol{\Theta}$. Then,

$$p(\boldsymbol{\Theta} \mid \boldsymbol{z}_{(n-1)}) = \frac{p(\boldsymbol{\Theta}) \prod_{t=2}^{n-1} f(z_t \mid z_{t-1})}{B_{(n-1)}} = \frac{p(\boldsymbol{\Theta}) \prod_{t=2}^{n} f(z_t \mid z_{t-1})}{B_{(n-1)}\, f(z_n \mid z_{n-1})} = \frac{B_{(n)}\, p(\boldsymbol{\Theta} \mid \boldsymbol{z}_{(n)})}{B_{(n-1)}\, f(z_n \mid z_{n-1})}$$

and therefore $p(z_n \mid \boldsymbol{z}_{(n-1)}) = \int f(z_n \mid z_{n-1}) p(\boldsymbol{\Theta} \mid \boldsymbol{z}_{(n-1)})\, \mathrm{d}\boldsymbol{\Theta} = B_{(n)}/B_{(n-1)}$. In addition, $\int \{f(z_n \mid z_{n-1})\}^{-1} p(\boldsymbol{\Theta} \mid \boldsymbol{z}_{(n)})\, \mathrm{d}\boldsymbol{\Theta} = B_{(n-1)}/B_{(n)}$, and thus

$$p(z_n \mid \boldsymbol{z}_{(n-1)}) = \left(\int \{f(z_n \mid z_{n-1})\}^{-1} p(\boldsymbol{\Theta} \mid \boldsymbol{z}_{(n)})\, \mathrm{d}\boldsymbol{\Theta} \right)^{-1}. \tag{B.1}$$

Similarly, $p(\boldsymbol{\Theta} \mid \boldsymbol{z}_{(n-2)}) = \{B_{(n)} p(\boldsymbol{\Theta} \mid \boldsymbol{z}_{(n)})\} / \{B_{(n-2)} f(z_n \mid z_{n-1}) f(z_{n-1} \mid z_{n-2})\}$. Hence, $p(z_{n-1} \mid \boldsymbol{z}_{(n-2)}) = \int f(z_{n-1} \mid z_{n-2}) p(\boldsymbol{\Theta} \mid \boldsymbol{z}_{(n-2)})\, \mathrm{d}\boldsymbol{\Theta} = \frac{B_{(n)}}{B_{(n-2)}} \int \{f(z_n \mid z_{n-1})\}^{-1} p(\boldsymbol{\Theta} \mid \boldsymbol{z}_{(n)})\, \mathrm{d}\boldsymbol{\Theta}$. Then, observing that $\int \{f(z_n \mid z_{n-1}) f(z_{n-1} \mid z_{n-2})\}^{-1} p(\boldsymbol{\Theta} \mid \boldsymbol{z}_{(n)})\, \mathrm{d}\boldsymbol{\Theta} = B_{(n-2)}/B_{(n)}$, we obtain an expression for $p(z_{n-1} \mid \boldsymbol{z}_{(n-2)})$ that involves the product of the two integrals above. Extending the derivation for $p(z_{n-1} \mid \boldsymbol{z}_{(n-2)})$, we obtain

$$p(z_t \mid \boldsymbol{z}_{(t-1)}) = \left(\int \left\{ \prod_{s=t}^{n} f(z_s \mid z_{s-1}) \right\}^{-1} p(\boldsymbol{\Theta} \mid \boldsymbol{z}_{(n)})\, \mathrm{d}\boldsymbol{\Theta} \right)^{-1} \left(\int \left\{ \prod_{s=t+1}^{n} f(z_s \mid z_{s-1}) \right\}^{-1} p(\boldsymbol{\Theta} \mid \boldsymbol{z}_{(n)})\, \mathrm{d}\boldsymbol{\Theta} \right)$$

for any $t = 3, ..., n-1$, with the expression for $t = n$ given in (B.1). These expressions allow us to estimate any posterior predictive ordinate $p(z_t \mid \boldsymbol{z}_{(t-1)})$, using Monte Carlo integration based on the samples from $p(\boldsymbol{\Theta} \mid \boldsymbol{z}_{(n)})$.

References

Antoniano-Villalobos, I. and Walker, S. G. (2016), "A nonparametric model for stationary time series," *Journal of Time Series Analysis*, 37, 126–142.

Azzalini, A. (1985), "A class of distributions which includes the normal ones," *Scandinavian Journal of Statistics*, 12, 171–178.

Caron, F., Davy, M., Doucet, A., Duflos, E., and Vanheeghe, P. (2008), "Bayesian Inference for linear dynamic models with Dirichlet process mixtures," *IEEE Transactions on Signal Processing*, 56, 71–84.

Carvalho, A. and Tanner, M. (2005), "Modeling nonlinear time series with local mixtures of generalized linear models," *Canadian Journal of Statistics*, 33, 97–113.

— (2006), "Modeling nonlinearities with mixtures of experts time series models," *International Journal of Mathematics and Mathematical Sciences*, 2006.

Connor, R. and Mosimann, J. (1969), "Concepts of independence for proportions with a generalization of the Dirichlet distribution," *Journal of the American Statistical Association*, 64, 194–206.

Daniels, M. and Pourahmadi, M. (2002), "Bayesian analysis of covariance matrices and dynamic models for longitudinal data," *Biometrika*, 89, 553–566.

DeYoreo, M. and Kottas, A. (2015), "A fully nonparametric modeling approach to binary regression," *Bayesian Analysis*, 10, 821–847.

Di Lucca, M., Guglielmi, A., Müller, P., and Quintana, F. (2013), "A simple class of Bayesian autoregression models," *Bayesian Analysis*, 8, 63–88.

Ferguson, T. (1973), "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, 1, 209–230.

Fox, E., Sudderth, E., Jordan, M., and Willsky, A. (2011), "Bayesian nonparametric inference for switching dynamic linear models," *IEEE Transactions on Signal Processing*, 59, 1569–1585.

Früwirth-Schnatter, S. (2006), *Finite Mixture and Markov Switching Models*, Springer.

Geweke, J. and Terui, N. (1993), "Bayesian threshold autoregressive models for nonlinear time series," *Journal of Time Series Analysis*, 14, 441–454.

Ishwaran, H. and James, L. (2001), "Gibbs sampling methods for stick-breaking priors," *Journal of the American Statistical Association*, 96, 161–173.

Juang and Rabiner (1985), "Mixture autoregressive hidden Markov models for speech signals," *IEEE Transactions and Acoustic, Speech, and Signal Processing*, 1404–1413.

Lau, J. and So, M. (2008), "Bayesian mixture of autoregressive models," *Computational Statistics and Data Analysis*, 53, 38–60.

MacEachern, S. (2000), "Dependent Dirichlet processes," Tech. rep., The Ohio State University, Department of Statistics.

Martinez-Ovando, J. and Walker, S. G. (2011), "Time-series modeling, stationarity, and Bayesian nonparametric methods," Tech. rep., Banco de Mexico.

Mena, R. and Walker, S. (2005), "Stationary autoregressive models via a Bayesian nonparametric approach," *Journal of Time Series Analysis*, 26, 789–805.

Müller, P., West, M., and MacEachern, S. (1997), "Bayesian models for nonlinear autoregressions," *Journal of Time Series Analysis*, 18, 593–614.

Sethuraman, J. (1994), "A constructive definition of Dirichlet priors," *Statistica Sinica*, 4, 639–650.

Tang, Y. and Ghosal, S. (2007a), "A consistent nonparametric Bayesian procedure for estimating autoregressive conditional densities," *Computational Statistics and Data Analysis*, 51, 4424–4437.

— (2007b), "Posterior consistency of Dirichlet mixtures for estimating a transition density," *Journal of Statistical Planning and Inference*, 137, 1711–1726.

Tong, H. (1987), "On a Threshold Model," in *Pattern Recognition and Signal Processing*, ed. Chen, C., Amsterdam: Sijhoff and Nordhoff.

— (1990), *Non-Linear Time Series: A Dynamical System Approach*, Oxford University Press.

Webb, E. and Forster, J. (2008), "Bayesian model determination for multivariate ordinal and binary data," *Computational Statistics and Data Analysis*, 52, 2632–2649.

West, M. and Harrison, J. (1999), *Bayesian Forecasting and Dynamic Models*, New York: Springer.

Wong, C. S. and Li, W. K. (2000), "On a mixture autoregressive model," *Journal of the Royal Statistical Society, Series B*, 62, 95–115.

Wood, S., Rosen, O., and Kohn, R. (2011), "Bayesian mixtures of autoregressive models," *Journal of Computational and Graphical Statistics*, 20, 174–195.