Title

A Bayesian nonparametric Markovian model for non-stationary time series

Permalink
https://escholarship.org/uc/item/6zq900tj

Journal
Statistics and Computing, 27(6)

ISSN
0960-3174

Authors
DeYoreo, Maria
Kottas, Athanasios

Publication Date
2017-11-01

DOI
10.1007/s11222-016-9702-x

Peer reviewed

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 nonstandard distributions. The model arises from a Bayesian nonparametric mixture of normals specification for the joint distribution of successive observations in time. This implies 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 more 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 and predictive densities. 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, Department of Applied Mathematics and Statistics, University of California, Santa Cruz. 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.

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.

Stationary time series models built from parametric distributions 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 are often restrictive in terms of the transition and marginal densities they imply. Time series may exhibit marginal distributions which are asymmetric, and predictive distributions which are nonlinear in the effects of past observations on the mean, and heteroskedasticity. Economic time series are generally believed to be nonstationary, often exhibiting distinct periods of high and low volatility, motivating the development of stochastic volatility and autoregressive (AR) conditional heteroskedasticity models, among others (Früwirth-Schnatter, 2006).

Various parametric models have been developed to capture nonlinear 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 non-linear 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 (2015) developed stationary time series models which contain flexible transition and invariant densities. Existing mixture models for time series are discussed further in Section 2.4, relative to our proposed model.

The restrictions commonly imposed on the residual distribution and autoregression relationship limit the scope of parametric AR models. Here, we present a general framework for modeling univariate time series data, which makes no assumptions of stationarity, and can accommodate complex dynamics and capture non-standard distributions. The model arises from a Bayesian nonparametric mixture of normals specification for the joint distribution of successive observations in time. This implies a flexible AR model structure for the conditional transition density. In particular, 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). 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 posterior simulation and prior specification. 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. While the model presented and all data illustrations are focused on univariate time series data with first-order dependence, we discuss possible extensions to accommodate higher order structure, and to develop a state-space model in Section 4. 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 $\prod_{t=2}^{n} f(z_t \mid z_{t-1})$. The model for the transition density, $f(z_t \mid z_{t-1})$, is induced by a nonparametric mixture of bivariate normal distributions for $f(z_{t-1}, z_t)$, which can accommodate a wide range of density shapes and complex dependencies between Z_t and Z_{t-1}.

More specifically, let $f(z_{t-1}, z_t) \equiv f(z_{t-1}, z_t; G) = \int \mathrm{N}(z_{t-1}, z_t; \mu, \Sigma)\, \mathrm{d}G(\mu, \Sigma)$, with a Dirichlet process (DP) prior (Ferguson, 1973) placed on the random mixing distribution G.

Thus, any two successive observations in time are distributed as a DP mixture of bivariate normals. 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

$$f(z_{t-1}, z_t; G) \approx \sum_{l=1}^{L} p_l \mathrm{N}(z_{t-1}, z_t; \mu_l, \Sigma_l). \tag{1}$$

The weights $(p_1, \dots, p_L)$ are determined through stick-breaking from latent beta$(1, \alpha)$ random variables (where $p_L = 1 - \sum_{l=1}^{L-1} p_l$), and the (μ_l, Σ_l) are independent and identically distributed (i.i.d.) from some base distribution G_0. Partitioning μ_l and Σ_l with superscripts x and y corresponding to z_{t-1} and z_t, respectively, the conditional transition density implied by (1) can be written as

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

with

$$q_l(z_{t-1}) = p_l \mathrm{N}(z_{t-1}; \mu_l^x, \Sigma_l^{xx}) / \left\{ \sum\nolimits_{m=1}^{L} p_m \mathrm{N}(z_{t-1}; \mu_m^x, \Sigma_m^{xx}) \right\}. \tag{3}$$

This 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. This defines a general time-homogeneous Markovian model which can handle nonstationary time series.

As discussed above, the transition density in (2) arises from the flexible and well-studied DP mixture of normals model for two successive observations in time. 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 associated mixture weights, given by (3), contain $\{\mu_l^x\}$ and $\{\Sigma_l^{xx}\}$ in the denominator, and each mixture component variance in (2) 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), 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 $(Y_1, \ldots, Y_m) \sim \mathrm{N}(\mu, \beta^{-1}\Delta(\beta^{-1})^T)$, 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 (2) admits the form

$$f(z_t \mid z_{t-1}; G) = \sum_{l=1}^{L} q_l(z_{t-1}) \mathrm{N}(z_t; \mu_l^y - \beta_l(z_{t-1} - \mu_l^x), \delta_l^y) \tag{4}$$

with

$$q_l(z_{t-1}) = p_l \mathrm{N}(z_{t-1}; \mu_l^x, \delta_l^x) / \left\{ \sum_{m=1}^{L} p_m \mathrm{N}(z_{t-1}; \mu_m^x, \delta_m^x) \right\} \tag{5}$$

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 $\eta_l = (\mu_l^x, \mu_l^y, \beta_l, \delta_l^x, \delta_l^y)$, 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:

$$\begin{aligned} z_t \mid z_{t-1}, U_t, \{\eta_l\} &\overset{ind.}{\sim} \mathrm{N}(\mu_{U_t}^y - \beta_{U_t}(z_{t-1} - \mu_{U_t}^x), \delta_{U_t}^y), \quad t = 2, \ldots, n \\ U_t \mid z_{t-1}, p, \mu^x, \delta^x &\overset{ind.}{\sim} \sum_{l=1}^{L} \frac{p_l \mathrm{N}(z_{t-1}; \mu_l^x, \delta_l^x)}{\sum_{m=1}^{L} p_m \mathrm{N}(z_{t-1}; \mu_m^x, \delta_m^x)} I(U_t = l), \quad t = 2, \ldots, n \\ \eta_l \mid \psi &\overset{i.i.d.}{\sim} G_0(\eta_l \mid \psi), \quad l = 1, \ldots, L \end{aligned} \tag{6}$$

and the prior density for $p = (p_1, \ldots p_L)$ is given by a special case of the generalized Dirichlet distribution: $f(p \mid \alpha) = \alpha^{L-1} p_L^{\alpha-1}(1 - p_1)^{-1}(1 - (p_1 + p_2))^{-1} \times \cdots \times (1 - \sum_{l=1}^{L-2} p_l)^{-1}$

(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 $\psi = (m^x, v^x, m^y, v^y, s^x, s^y, \theta, c)$, the hyperparameters of G_0:

$$\begin{aligned} m^x \sim \mathrm{N}(a_m^x, b_m^x),\ m^y \sim \mathrm{N}(a_m^y, b_m^y),\ v^x \sim \mathrm{IG}(a_v^x, b_v^x),\ v^y \sim \mathrm{IG}(a_v^y, b_v^y), \\ s^x \sim \mathrm{Ga}(a_s^x, b_s^x),\ s^y \sim \mathrm{Ga}(a_s^y, b_s^y),\ \theta \sim \mathrm{N}(a_\theta, b_\theta),\ c \sim \mathrm{IG}(a_c, b_c) \end{aligned} \tag{7}$$

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

2.2 Posterior Inference

Samples from the full posterior distribution of the model are obtained using a combination of Gibbs sampling and Metropolis-Hastings steps. Here, we describe posterior simulation details for all model parameters, focusing particular attention on the vector $(p_1, \dots, p_L)$, which requires the most care in developing an effective updating strategy.

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

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

$$p(\eta_l \mid \dots, \text{data}) \propto G_0(\eta_l \mid \psi) \left\{ \prod_{j=1}^{n^*} \prod_{\{t: U_t = U_j^*\}} \mathrm{N}(z_t; \mu_l^y - \beta_l(z_{t-1} - \mu_l^x), \delta_l^y) \right\} \left\{ \prod_{l=1}^{L} \prod_{\{t: U_t = l\}} q_l(z_{t-1}) \right\}.$$

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 $\delta_l^y \sim \mathrm{IG}(\nu^y + 0.5 M_l, s^y +$

$0.5\sum_{\{t:U_t=l\}}(z_t - \mu_l^y + \beta_l(z_{t-1} - \mu_l^x))^2)$ 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

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

This can be written as $p(\mu_l^x | \ldots, \text{data}) \propto \mathrm{N}(\mu_l^x; (m^x)^*, (v^x)^*)(\prod_{t=2}^{n}\sum_{m=1}^{L} p_m \mathrm{N}(z_{t-1}; \mu_m^x, \delta_m^x))^{-1}$, 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 $\mathrm{N}(\mu_l^x; m^x, v^x)[\prod_{t=2}^{n}\sum_{m=1}^{L} p_m \mathrm{N}(z_{t-1}; \mu_m^x, \delta_m^x)]^{-1}$, and in this case we use a Metropolis-Hastings algorithm, proposing a candidate value μ_l^x from the base distribution $\mathrm{N}(m^x, v^x)$.

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

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

which for an active component, is written as proportional to

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

For non-active components, the full conditional is $\mathrm{IG}(\delta_l^x; \nu^x, s^x)(\prod_{t=2}^{n}\sum_{m=1}^{L} p_m \mathrm{N}(z_{t-1}; \mu_m^x, \delta_m^x))^{-1}$. 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 $G_0(\delta_l^x) = \mathrm{IG}(\nu^x, s^x)$.

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

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

In standard DP mixture models, the implied generalized Dirichlet prior for $f(p \mid \alpha)$ 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 p arising from the DP truncation approximation.

Recall that the joint prior for p corresponds to a generalized Dirichlet distribution, which can be constructed from latent beta random variables through stick-breaking. Let $v_1, \ldots, v_{L-1} \overset{i.i.d.}{\sim} \text{beta}(1, \alpha)$, and define $p_1 = v_1$, $p_l = v_l \prod_{r=1}^{l-1}(1 - v_r)$, for $l = 2, \ldots, L-1$, and $p_L = \prod_{r=1}^{L-1}(1-v_r)$. Equivalently, 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 p, we work with the ζ_l, a sample for which implies a particular probability vector p.

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

$$p(\zeta_l \mid \ldots, \text{data}) \propto \text{beta}\left(\zeta_l; \alpha + \sum_{r=l+1}^{L} M_r, M_l + 1\right) \left(\prod_{t=2}^{n} d(z_{t-1}) \right)^{-1} \tag{8}$$

where

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

Also, let $c_{t,l} = \mathrm{N}(z_{t-1}; \mu_l^x, \delta_l^x)$, which is constant with respect to each ζ_l. The form of the full conditional in (8) 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, which implies a corresponding probability vector p.

At any time point t, an entire distribution can be obtained for $f(z_{t+1} \mid Z_t = z_t; G)$, for any z_t, providing full inference for the transition density. This conditional distribution can be evaluated at the last time point, conditional on Z_n, to give a forecasting distribution, $f(z_{n+1} \mid Z_n = z_n; G) = \sum_{l=1}^{L} q_l(z_n) \mathrm{N}(z_{n+1}; \mu_l^y - \beta_l(z_n - \mu_l^x), \delta_l^y)$. Full inference is readily available for any z_{n+1}, yielding an entire forecasting distribution. The point estimate of this distribution is the posterior predictive density for the next observation, since it can be shown that $p(z_{n+1} \mid Z_n = z_n; \text{data}) = \mathrm{E}(f(z_{n+1} | Z_n = z_n; G) \mid \text{data})$. 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.

2.3 Prior Specification

We now discuss prior specification for the hyperparameters ψ of G_0, aiming to specify appropriately diffuse priors which use only a small amount of prior information. Recall that the model for the transition density $f(z_t \mid z_{t-1})$ was motivated by a DP mixture of bivariate normals, that is $f(z_{t-1}, z_t; G) = \int \mathrm{N}(z_{t-1}, z_t; \mu, \Sigma) \mathrm{d}G(\mu, \Sigma)$, with G having a DP prior. In the limit, as $\alpha \to 0^+$, this model consists of a single mixture component, $\mathrm{N}(z_{t-1}, z_t; \mu, \Sigma)$. An approximate center d and range r of the data are used to center and scale the mixture kernel appropriately.

Based on the form of G_0, we obtain $\mathrm{E}(z_{t-1}) = \mathrm{E}(\mu^x) = a^x_m$ and $\mathrm{E}(z_t) = \mathrm{E}(\mu^y) = a^y_m$, and therefore set $a^x_m = a^y_m = d$. We find $\mathrm{Cov}(z_{t-1}, z_t) = \mathrm{E}(\Sigma) + \mathrm{Cov}(\mu)$, and use this expression to scale the prior for μ. The marginal prior variances for the components of μ are $\mathrm{var}(\mu^x) = b^x_m + (a^x_v - 1)^{-1} b^x_v$ and $\mathrm{var}(\mu^y) = b^y_m + (a^y_v - 1)^{-1} b^y_v$. Fixing small values for a^x_v and a^y_v to ensure large variance for v^x and v^y, and assuming $\mathrm{var}(\mu^x) \approx (r/4)^2$ and $\mathrm{var}(\mu^y) \approx (r/4)^2$, one can then obtain reasonable values for b^x_m, b^x_v, b^y_m, and b^y_v. This completes specification of the parameters associated with μ^x and μ^y.

We now discuss two approaches to prior specification for the hyperparameters associated with Σ. One approach involves obtaining the prior expectation of the diagonal elements of Σ, and setting each of these equal $(r/4)^2$, while also ensuring that the implied prior on the correlation $-\beta\delta^x((\beta^2\delta^x + \delta^y)\delta^x)^{-1/2}$ is approximately uniform on $(-1, 1)$. We find that $\mathrm{E}(\Sigma^{xx}) = \mathrm{E}(\delta^x) = (b^x_s(\nu^x - 1))^{-1} a^x_s$, and fixing a^x_s and ν^x, determine b^x_s so that $\mathrm{E}(\Sigma^{xx}) = (r/4)^2$. The prior on β should be centered around 0, supporting independence in the normal kernel, that is, $a_\theta = 0$. Then, again taking prior expectations, $\mathrm{E}(\Sigma^{yy}) = (b_\theta + b_c(a_c - 1)^{-1})\mathrm{E}(\delta^x) + (b^y_s(\nu^y - 1))^{-1} a^y_s$. Fixing ν^y, a_c, and a^y_s, this sum can be set equal to $(r/4)^2$, where the particular values of b_θ, b_c, and b^y_s are determined so that the induced prior on the correlation is approximately uniform on $(-1, 1)$, as verified through prior simulation.

An alternative, more automatic, strategy arises from considering the distributions implied on β, δ^x, and δ^y if Σ is inverse-Wishart distributed, a setting under which we are accustomed to specifying priors for covariance matrices. A common noninformative $\mathrm{IW}(a, B)$ specification for Σ involves fixing a small value for the degrees of freedom parameter a, and assuming B to be a diagonal matrix with diagonal (B_1, B_2). We can, for instance, fix $a = 4$, the smallest possible integer value such that Σ has finite expectation, and assume $\mathrm{E}(\Sigma) = \mathrm{diag}\{(r/4)^2, (r/4)^2\}$, that is, $(B_1, B_2) = ((r/4)^2, (r/4)^2)$. If $\Sigma \sim \mathrm{IW}(a, B)$, then as a consequence, $\delta^x \sim \mathrm{IG}(0.5(a-1), 0.5B_1)$, $\delta^y \sim \mathrm{IG}(0.5a, 0.5B_2)$, and $\beta \mid \delta^y \sim \mathrm{N}(0, \delta^y B_1^{-1})$ (DeYoreo and Kottas, 2015). Therefore, we set $\nu^x = 0.5(a-1)$ and $\nu^y = 0.5a$, and let $\mathrm{E}(s^x) = \mathrm{E}(s^y) = 0.5(r/4)^2$. Exponential priors for s^y and s^x yield $b^x_s = b^y_s = 2(r/4)^{-2}$.

After marginalizing out θ, the G_0 component for β becomes $\mathrm{N}(a_\theta, b_\theta + c)$, so we let $a_\theta = 0$, and $b_\theta + \mathrm{E}(c) = B_1^{-1}\mathrm{E}(\delta^y) = 0.5(\nu^y - 1)^{-1}$. Assuming $b_\theta = \mathrm{E}(c)$, and fixing a_c, b_θ and b_c can be determined accordingly.

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 (2). 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 the DP 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 are assumed to be i.i.d from some 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 normal kernels, with means given by functions of previous observations. Tang and Ghosal (2007a) consider a particular version of this class of models, involving a hyperbolic tangent transformation of lagged terms, which can approximate any linear autoregressive model arbitrarily closely. 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 then a countable location mixture of AR models, with mixing taking place on the AR parameters. Mena and Walker (2005) construct transition densities nonparametrically, but restrict the transition densities further 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.

Stationarity arises as a special case of (2), occurring when the two marginal densities are identical. To preserve stationarity, one can set $\mu_l^x = \mu_l^y$ and $\Sigma_l^{xx} = \Sigma_l^{yy}$. This is the version of the model studied by Antoniano-Villalobos and Walker (2015), who focus on building flexible stationary models, as stationarity is desirable in some settings, but achieving both stationarity and flexibility in the transition and invariant densities is a challenge. Their modeling framework begins much like ours, in that a transition mechanism is obtained as the conditional density from a bivariate distribution. The authors do not apply a truncation approximation to G, thus inference under this model requires the introduction of multiple sets of latent variables and a trans-dimensional MCMC algorithm for posterior simulation. The model developed by Antoniano-Villalobos and Walker (2015) 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. Note that under our parameterization, constraints to stationary yield a transition density $\sum_{l=1}^{L} q_l(z_{t-1})\mathrm{N}(z_t; \mu_l - \beta_l(z_{t-1} - \mu_l), \sigma_l^2(1-\beta_l^2))$ with $q_l(z_{t-1}) \propto p_l \mathrm{N}(z_{t-1}; \mu_l, \sigma_l^2)$, and $\beta_l \in (-1, 1)$.

Although we utilize a truncation approximation to the DP from the outset, the sum in the denominator of the weights in (3) still presents challenges in terms of posterior simulation. We developed a tractable MCMC algorithm by reparameterizing the covariance matrices in (1) and, rather than working directly with the probability vector p which is difficult to update efficiently, working with the stick-breaking weights to develop a slice sampler which indirectly provides samples for p.

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). In all cases, MCMC inference was implemented in R, saving every 20-th iteration after burn-in, and a Monte Carlo sample size of $5,000$ was used for inference. We follow the approach to prior specification described in Section 2.3, using the second method for fixing hyperparameters associated with β, δ^x, and δ^y, which follows from considering the distributions implied under an inverse-Wishart specification for Σ.

The DP truncation level L was chosen using standard DP properties: under a gamma$(0.5, 0.5)$ prior on α, the expectation of the partial sum of the original DP weights, $\mathrm{E}(\sum_{l=1}^{L} p_l)$, 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 effective components to ensure it never reached the upper bound.

3.1 Simulated Data

3.1.1 Skew-normal Transition Densities

To generate a time series that exhibits challenging transition densities which evolve over time in a plausible 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 $z_t \mid z_{t-1} \sim \mathrm{SN}(z_t; 0, 1 + 0.7|z_{t-1}|, 0.1 + 4\sin(z_{t-1}))$, for $t = 2, \ldots, n$. Here, $\mathrm{SN}(y; \xi, \omega, \alpha)$ denotes a skew-normal distribution with density $(\omega\pi)^{-1}\exp(-(y-\xi)^2/(2\omega^2))\Phi(\alpha(x-\xi)/\omega)$, where $\Phi(\cdot)$ denotes 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.

A time series $(z_2, \ldots, z_{500})$ was simulated from this model assuming an initial value $z_1 = 0$. Figure 1 (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. We estimate $\mathrm{E}(Z_t \mid Z_{t-1} = z_{t-1})$ by evaluating $\sum_{l=1}^{L} q_l(z_{t-1})\{\mu_l^y - \beta_l(z_{t-1} - \mu_l^x)\}$ over a grid in z_{t-1}, providing point estimates and uncertainty quantification for the expectation of the next observation in a series given that the previous observation was z_{t-1}. Figure 1 (right panel) displays these results along with the data-generating expectation trend. The estimates generally match fairly closely and the 95% credible intervals contain the truth

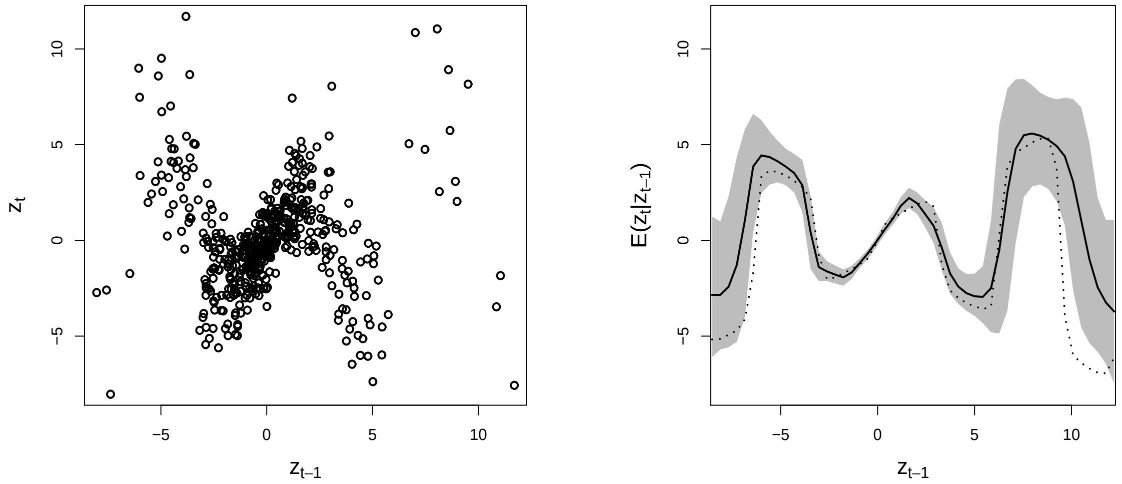

Figure 1: Skew-normal simulation. The left panel plots the simulated data as pairs of points (z_{t-1}, z_t). The right panel shows the posterior mean (solid line) and 95% credible intervals (gray shaded region) for $\mathrm{E}(Z_t \mid Z_{t-1} = z_{t-1})$ plotted over a grid in z_{t-1}; the true expectation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

everywhere except for a small region around $z_{t-1} = 10$, where there is very little data.

We also compute posterior predictive densities $p(z_t \mid z_{t-1}; G)$, for $t = 2, \ldots, n$. These densities are displayed in Figure 2 (top panel). We plot each index t on the horizontal axis, and the corresponding predictive density for z_t on the vertical axis, using darker colors to represent larger values. The true predictive densities are given also in Figure 2 (bottom panel) and the data is shown in each plot. While these inferences are based on only a posterior point estimate for $f(z_t \mid z_{t-1})$, we also have available full inference which we display in the form of point estimates and 95% uncertainty bands for $f(z_t \mid Z_{t-1} = z_{t-1})$ at four particular values of z_{t-1} in Figure 3. 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 5 or 6 and the large amount of data in the region near 0.

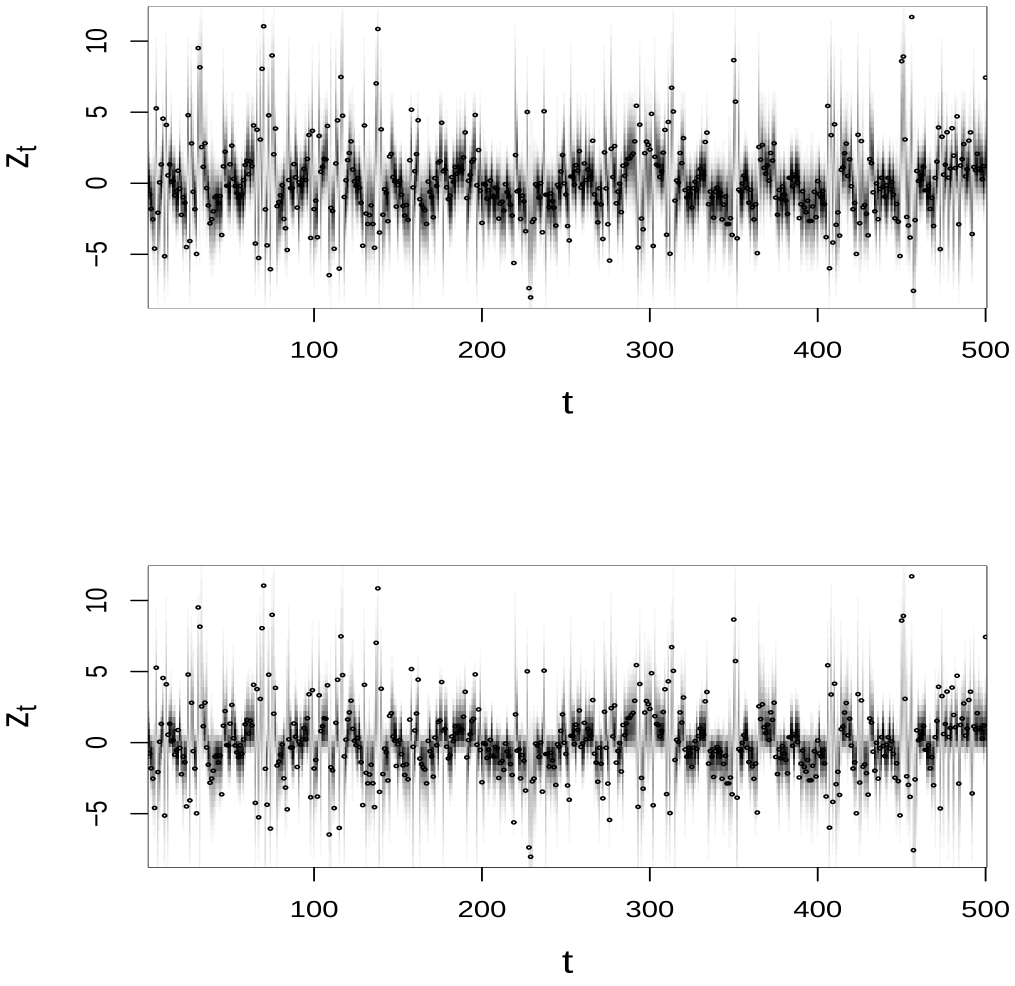

Figure 2: Skew-normal simulation. Predictive densities $p(z_t \mid z_{t-1})$ for each $t = 2, \ldots, n$. The top panel displays the estimates from the model and the bottom panel shows the true densities. Darker colors indicate larger values. The data is also included in each plot.

3.1.2 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 4, left panel). The value of the last observation is approximately -14.2, one of the smallest values in the entire series. The

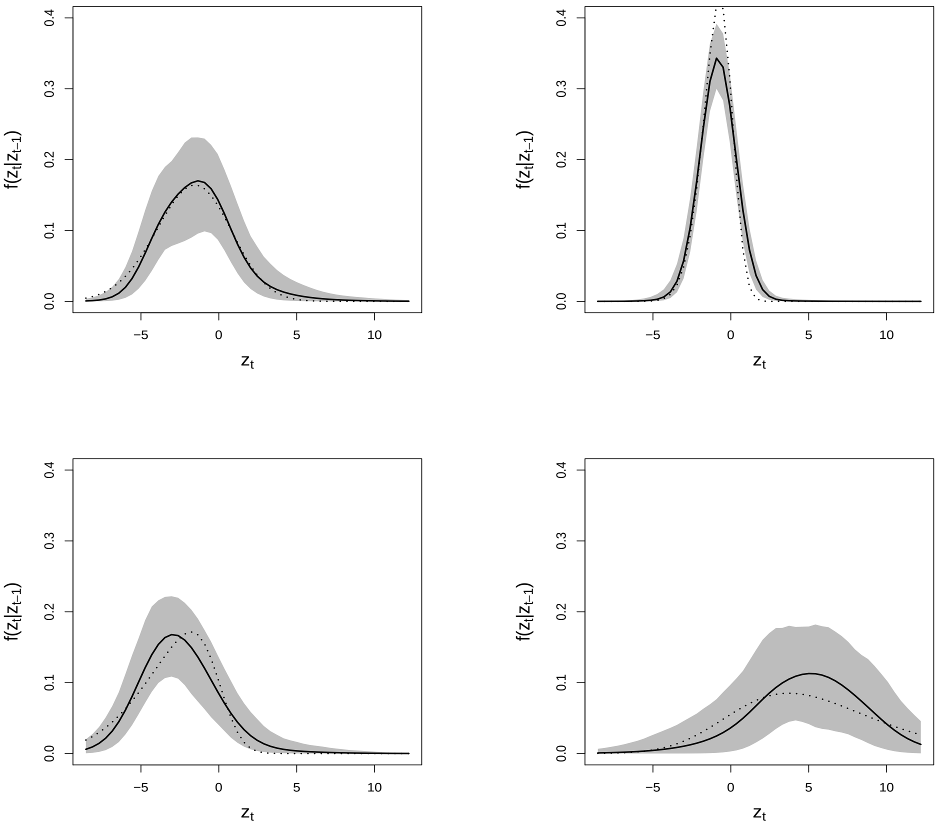

Figure 3: Skew-normal simulation. Posterior mean (solid line) and 95% credible intervals (gray shaded region) for transition densities $f(z_t \mid z_{t-1})$, for $z_{t-1} = -2.85$ (top left), $z_{t-1} = -0.5$ (top right), $z_{t-1} = 4.2$ (bottom left), and $z_{t-1} = 8.85$ (bottom right). The corresponding true densities are plotted as dotted lines.

forecast distribution for the next observation is displayed in Figure 4 (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.

Posterior predictive densities $p(z_t \mid z_{t-1}; G)$, for $t = 2, \ldots, n$, are displayed in Figure 5 (top panel). For each index t on the horizontal axis, the corresponding predictive density for z_t is plotted on the vertical axis, where darker colors represent larger values. The true predictive densities are also plotted in Figure 5 (bottom panel). In summary, all visual displays indicate that the estimation from the model is of very good quality, capturing the dynamics in the data exceedingly well.

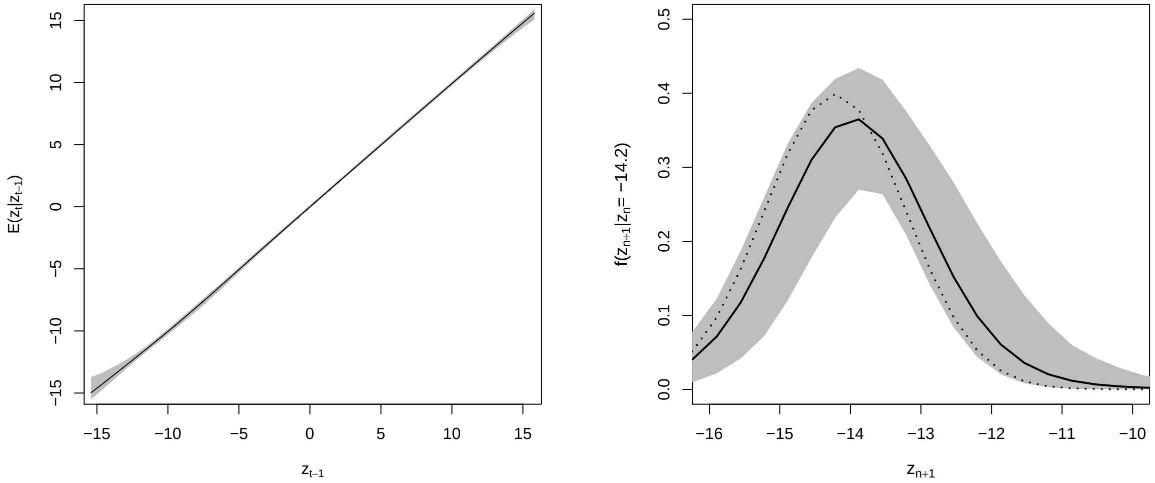

Figure 4: Brownian motion simulation. The left panel plots the posterior mean estimate (solid line) and 95% credible intervals (gray shaded region) for $\mathrm{E}(Z_t \mid Z_{t-1} = z_{t-1})$ plotted over a grid in z_{t-1}. The true expectation is indistinguishable from the model's estimate. The right 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).

3.2 Waiting Times Between Eruptions of the Old Faithful Geyser

We illustrate the proposed model on the time intervals between successive eruptions of the Old Faithful geyser, which are available through R under the dataset `faithful`. The data set 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 displayed in Figure 6 in the form of a plot of y_t versus y_{t-1}, for $t = 2, \ldots, 272$. Also plotted in Figure 6 are the posterior mean estimate and 95% credible intervals for $\mathrm{E}(Z_t \mid Z_{t-1} = z_{t-1})$.

The model required an average of 8 mixture components. Standard MCMC diagnostics suggest convergence has been reached. For instance, trace plots corresponding to $5,000$ posterior samples are shown for two quantities in Figure 7. The plot on the right panel of Figure 7 monitors the average of the standard deviations of the conditional densities $f(z_t \mid z_{t-1}; G)$, that is, $\sum_{t=2}^{n} (\delta^y_{U_t})^{0.5}/(n-1)$, while the left panel plot monitors $\sum_{t=2}^{n} \beta_{U_t}/(n-1)$, which refers to an important quantity since the β_l control the strength and direction of the autoregressions. While each β_l was centered at 0 in the prior, the posterior favors slightly

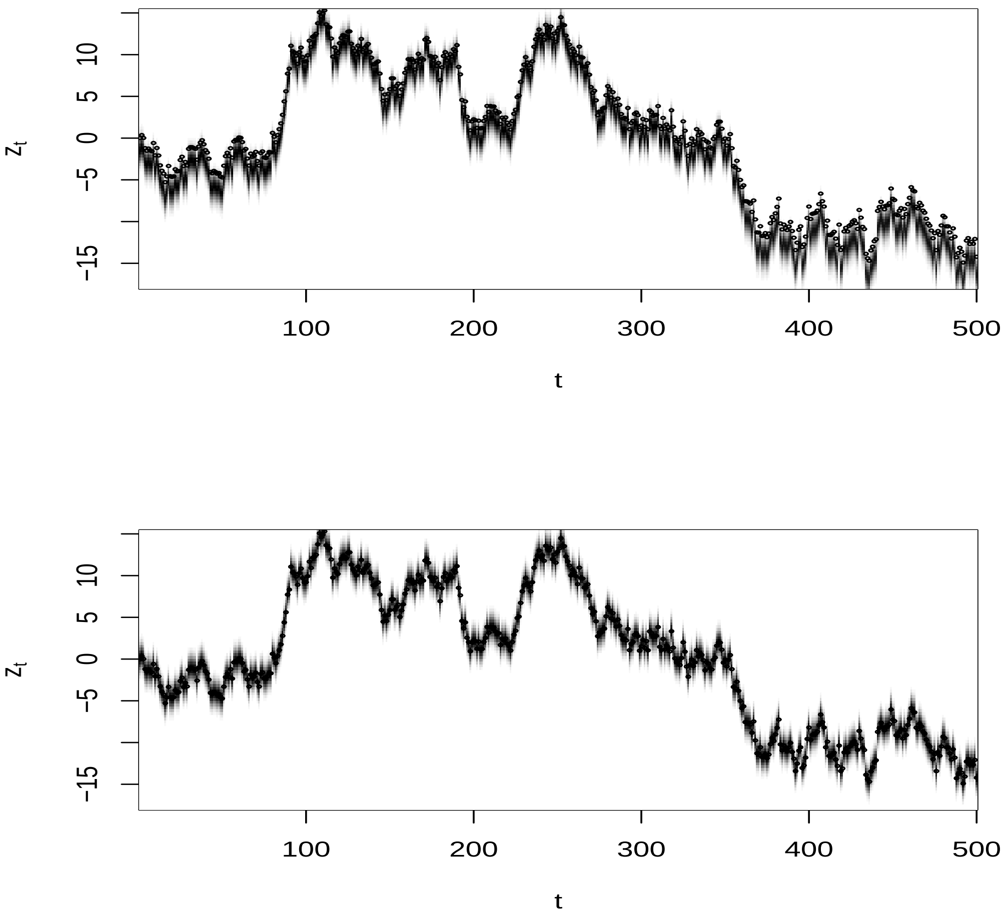

Figure 5: Brownian motion simulation. Predictive densities $p(z_t \mid z_{t-1})$ for each $t = 2, \ldots, n$. The top panel displays the estimates from the model and the bottom panel shows the true densities. Darker colors indicate larger values. The data is also included in each plot.

positive values.

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 below $z_t = 50$, 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 estimated transition densities at $f(z_t \mid z_{t-1} = 50)$ and $f(z_t \mid z_{t-1} = 80)$, which are displayed in Figure 8. One may be interested in predicting the next value z_{n+1} in the series. It is

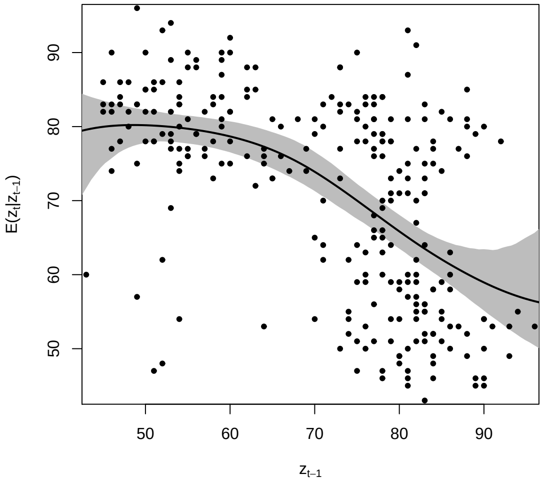

Figure 6: Old Faithful data. Posterior mean (solid line) and 95% credible intervals (gray shaded region) for $\mathrm{E}(Z_t \mid Z_{t-1} = z_{t-1})$ plotted over a grid in z_{t-1}, and overlaid on the data shown as pairs of points (z_{t-1}, z_t), for $t = 2, \ldots, 272$.

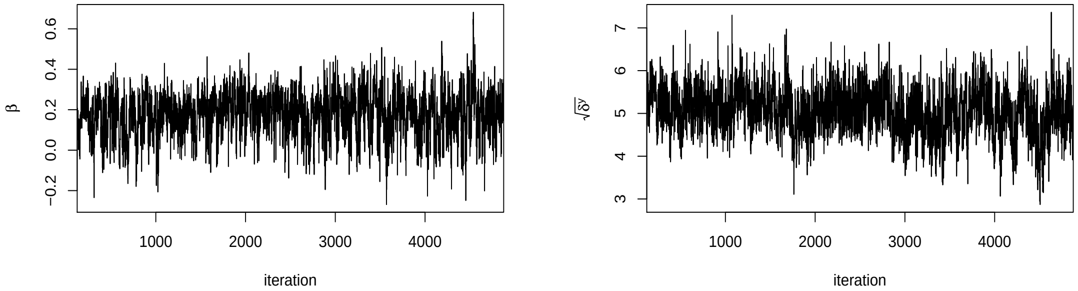

Figure 7: Old Faithful data. Trace plots of averages of $\{\beta_{U_t},\, t = 2, \ldots, 272\}$ (left panel) and $\{(\delta^y_{U_t})^{0.5},\, t = 2, \ldots, 272\}$ (right panel) over $5,000$ MCMC iterations.

important to provide estimates of uncertainty in this context. The posterior mean estimate and 95% credible intervals for the forecast density are given in Figure 9. This density is bimodal with the larger mode near 50 and another mode near 80, which is reasonable given the cross-section of data around $z_n = 74$.

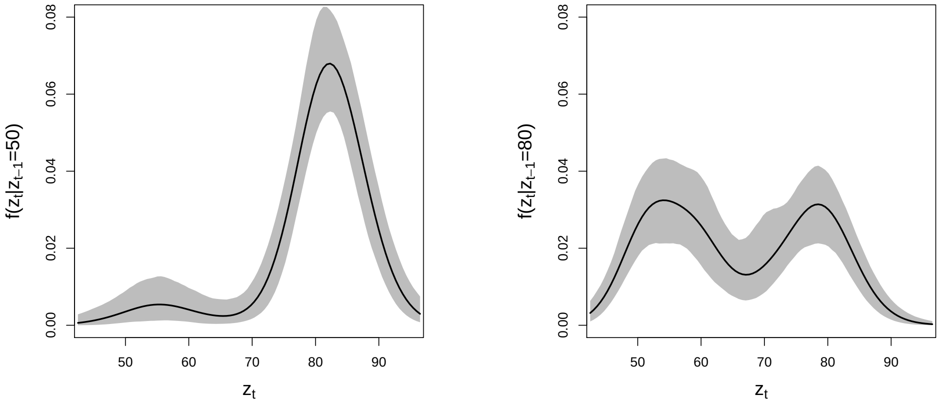

Figure 8: Old Faithful data. Posterior mean (solid line) and 95% credible intervals (gray shaded region) for transition densities $f(z_t \mid z_{t-1})$, for $z_{t-1} = 50$ (left panel) and $z_{t-1} = 80$ (right panel).

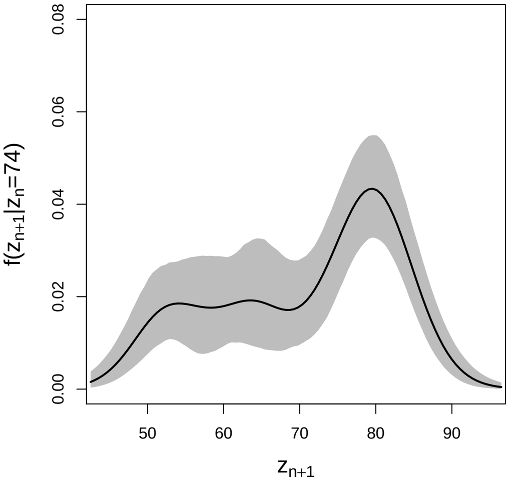

Figure 9: Old Faithful data. Posterior mean (solid line) and 95% credible intervals (gray shaded region) for the forecast density, $f(z_{n+1} \mid z_n = 74)$.

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.

Although the flexibility which is induced by the joint DP mixture modeling framework allows the first-order Markovian model to capture more complex features than many parametric models, there are scenarios in which a higher-order structure is required. The first-order model can be extended to incorporate higher order Markovian processes, by assuming $f(z_{t-r},\ldots,z_{t-1},z_t;G) \sim \int \mathrm{N}(\mu,\Sigma)\mathrm{d}G(\mu,\Sigma)$. Then $f(z_t \mid z_{t-1};G)$ is replaced by the transition density of order r, $f(z_t \mid z_{t-1},\ldots,z_{t-r};G)$. This transition density has a similar form to (2), but now the means of the normal 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 μ 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 normal kernels have the form $\mathrm{N}(z_t;\mu_l^y - \sum_{j=1}^{r}\beta_{l,(r+1,j)}(z_{t+j-r-1}-\mu_{l,j}^x),\delta_l^y)$, for $l=1,\ldots,L$. Gibbs sampling steps are thus preserved for μ_l^y and δ_l^y, as well as the last row of the matrix β. However, more care is needed in devising an MCMC algorithm to sample δ_l^x, μ_l^x (each vectors 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 flexible Markovian model. Denote the observed data by $(y_1,\ldots,y_n)$, and the underlying latent states by $(z_1,\ldots,z_n)$. Assume $y_t \mid z_t,\theta \sim f(y_t \mid z_t;\theta)$, for some parametric distribution $f(y_t \mid z_t;\theta)$, and assume the latent states evolve according to the nonparametric Markovian model for $f(z_t \mid z_{t-1};G)$ in (4). 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 nonstandard transition densities and nonlinear autoregressions. The conditional 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 structure is induced from a Dirichlet process mixture of normals specification for the joint distribution of successive observations in time. We have discussed methods for posterior inference and prior specification, and illustrated the model with synthetic and real data. 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.

References

Antoniano-Villalobos, I. and Walker, S. (2015), “A nonparametric model for stationary time series,” *Journal of Time Series Analysis*, To appear.

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*, To appear.

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. (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.