Bayesian Nonparametric Density Autoregression with Lag Selection

Matthew Heiner and Athanasios Kottas*

March 21, 2020

Abstract

We develop a Bayesian nonparametric autoregressive model applied to flexibly estimate general transition densities exhibiting nonlinear lag dependence. Our approach is related to Bayesian density regression using Dirichlet process mixtures, with the Markovian likelihood defined through the conditional distribution obtained from the mixture. This results in a Bayesian nonparametric extension of a mixtures-of-experts model formulation. We address computational challenges to posterior sampling that arise from the Markovian structure in the likelihood. The base model is illustrated with synthetic data from a classical model for population dynamics, as well as a series of waiting times between eruptions of Old Faithful Geyser. We study inferences available through the base model before extending the methodology to include automatic relevance detection among a pre-specified set of lags. Inference for global and local lag selection is explored with additional simulation studies, and the methods are illustrated through analysis of an annual time series of pink salmon abundance in a stream in Alaska. We further explore and compare transition density estimation performance for alternative configurations of the proposed model.

Keywords: Dirichlet process mixtures; Dynamical system; Local regression; Markov chain Monte Carlo; Order selection.

1 Introduction

This article is concerned with flexible transition density estimation for non-stationary, nonlinear time series. Let $\{y_t : t = 1, \ldots, T\}$ denote a univariate series governed by a time-homogeneous transition density $p(y_t \mid y_{t-1}, \ldots, y_1)$. While nonlinearity has been used to describe various qualitative characteristics of time series, we specifically refer to nonlinear dynamics, or the function mapping past observations to the present. Many existing methods for nonlinear regression have been applied to autoregressive modeling within and out of the statistical literature. Density regression has received far less attention, especially in application to transition density estimation, a crucial component of probabilistic forecasting and decision modeling. We seek to build on recent advances in transition density estimation by exploring what can be succinctly described as a Bayesian nonparametric mixture of autoregressive models. To accommodate nonlinear dependence, mixture weights are functions of lagged observations. Thus, our method is also accurately described as a locally linear autoregressive model.

Perhaps the most popular mixture modeling application to time series is the class of hidden Markov models (HMMs), which are capable of capturing nonlinear dynamics (Frühwirth-Schnatter, 2006, and references therein). Markovian dependence in a latent process, however, complicates inferences for transition densities and related functionals, especially when considering multiple lags. Our approach builds dependence in mixture weights through lagged observations directly, as with the likewise popular class of threshold autoregressive models (Tong, 1990), and mixtures-of-experts (MoE) models (Jordan and Jacobs, 1994; Peng et al., 1996; Carvalho and Tanner, 2005, 2006). In place of the parameterized link functions of lagged observations commonly employed by MoE models, we use normalized kernels for local weighting (Glasbey, 2001; Kalliovirta et al., 2015).

In contrast with the above methods, our model theoretically employs a countable mixture, bypassing the need to select the number of active components. The basic form of the transition density for observation y_t, conditional on L lags $\boldsymbol{y}_{t-1} \equiv (y_{t-1}, \ldots, y_{t-L})$, is given by $\sum_{h=1}^{\infty} q_h(\boldsymbol{y}_{t-1}) \, p_h(y_t \mid \boldsymbol{y}_{t-1})$, with component-specific normalized weight functions $q_h(\boldsymbol{y}_{t-1})$ and kernel densities $p_h(y_t \mid \boldsymbol{y}_{t-1})$. Among the various options for weight function and kernel density specifications, we focus on the flexible density regression construction of Müller et al. (1996), which builds from the conditional distribution arising from a joint modeled with a Dirichlet process mixture (DPM; Ferguson, 1973; Antoniak, 1974).

Bayesian nonparametric (BNP) extensions to popular time-series methods have seen rapid growth over the past two decades, perhaps most notably with hidden Markov (Beal et al., 2002; Taddy and Kottas, 2009; Yau et al., 2011) and dynamic linear (Rodríguez and Ter Horst, 2008; Caron et al., 2007; Fox et al., 2011) model frameworks. Dirichlet process mixtures of linear autoregressive (AR) models, which are closer to our formulation, can be viewed as a nonparametric extension of the mixture autoregressive model of Wong and Li (2000), although Müller et al. (1997) had previously employed a BNP model within a finite MoE framework using normalized weights. Lau and So (2008) proposed a DPM of autoregressive models, mixing on autoregressive coefficients, variance, and order in each kernel. Di Lucca et al. (2013) frame a similar model in the dependent Dirichlet process

(DDP; MacEachern, 2000) framework, ultimately demonstrating first-order mixtures of autoregressive models with common weights. Posterior consistency for BNP transition density estimation has been explored by Tang and Ghosal (2007a), Tang and Ghosal (2007b), and Chae and Walker (2019).

Stationarity, or time invariance of the marginal probability distribution $p(y_t)$, is not a common feature of the models described thus far. Mena and Walker (2005) and Martınez-Ovando and Walker (2011) build stationarity into their model definition, in the former case through the Gibbs construction of Pitt et al. (2002), resulting in AR-type models with BNP specifications for transition and marginal distributions. Antoniano-Villalobos and Walker (2016) build on Martınez-Ovando and Walker (2011), constructing a transition density from a mixture model on the stationary joint density of the current observation and a single lag. In contrast with Müller et al. (1996), their likelihood is based on the conditional transition density, which is a nonparametric mixture of kernels with linear autoregressive means and lag-dependent, normalized weights. Kalli and Griffin (2018) extend this framework to a stationary multivariate autoregressive model of multiple lags, although the model is implemented with a single lag. DeYoreo and Kottas (2017) use a model construction similar to Antoniano-Villalobos and Walker (2016), but do not assume stationarity. They empirically demonstrate superior flexibility over the stationary model. The model proposed in this article extends that of DeYoreo and Kottas (2017) to accommodate multiple lags and, crucially, shrink dependence to a minimally sufficient set of lags. The added modeling and computational complexity associated with high-order dependence demands that lag selection play a vital role in this work, as it affords parsimony and significantly reduces the estimation burden.

The primary contributions of this article are: 1) extension of a powerful class of non-stationary, nonlinear density autoregression models to accommodate dependence on multiple lags; 2) development of a framework for model-based selection and exploration of lag dependence; 3) investigation into the proposed model's fitness for different analysis scenarios; and, 4) demonstration of the need for lag selection in high-order density autoregression.

The rest of the article is organized as follows. In Section 2, we propose a BNP time-series model for density autoregression and present details for implementation and inference. In Section 3, we illustrate the model fit to synthetic and real data. In Section 4, we extend the model to incorporate inferences about relevant lags and demonstrate its use on data. Section 5 compares transition density estimation performance under different model configurations, using simulated nonlinear time series featuring skewness, heteroscedasticity, and different lag dependence structures. Finally, Section 6 concludes with discussion. The Supplementary Material contains details on: model modifications for stationary time series; prior specification; the Markov chain Monte Carlo (MCMC) algorithms for the base model and its extension that incorporates lag selection; and an additional simulation example.

3

2 The modeling approach

Our objective is to develop a general-purpose and fully nonparametric, time-homogeneous Markovian model that is sufficiently flexible to: 1) estimate possible non-Gaussian transition densities, dependent on lagged values, 2) capture nonlinear dynamics, and 3) select relevant lags among a pre-specified set, up to a maximal order L. The first two objectives are accomplished through a nonparametric mixture of Gaussian densities, wherein both the mixture weights and kernel means depend on lagged observations. The general model formulation for the transition density can be written as

$$f(y_t \mid \boldsymbol{y}_{t-1}) = \sum_{h=1}^{\infty} \underbrace{q_h(\boldsymbol{y}_{t-1})}_{\text{local weights}} \underbrace{\mathrm{N}(y_t \mid \mu_h(\boldsymbol{y}_{t-1}), \sigma_h^2)}_{\text{mixture kernels}}, \tag{1}$$

where $\mathrm{N}(y \mid \mu, \sigma^2)$ denotes a Gaussian density with mean μ and variance σ^2 evaluated at y, and with weight function $q_h(\boldsymbol{y}_{t-1}) \geq 0$ for all $h \in \mathbb{N}$ such that $\sum_{h=1}^{\infty} q_h(\boldsymbol{y}_{t-1}) = 1$ for all $\boldsymbol{y}_{t-1} \in \mathbb{R}^L$. The kernel mean functions $\mu_h(\boldsymbol{y}_{t-1})$ are linear in the lags, yielding a local linear model formulation. The objective of order and lag selection is accomplished through a stochastic-search prior structure.

Time homogeneity is a consequence of time invariance in the parameters governing the mixture weights and kernels. We note that this seemingly restrictive assumption is at least partially offset by the model's flexibility with respect to lagged observations. Apparently time-dependent structural changes can sometimes be attributed to heterogeneity of response across the state space. In such cases, a latent first-order Markov process governing the mixture weights may be less effective than our approach of using the lagged values directly. Nevertheless, dynamic drift or regime-switching in model structure may be more appropriate in some scenarios, for which we urge thoughtful exploration before selecting a model.

We proceed with model specification in Section 2.1, built using a covariance matrix parameterization that is useful for interpretation and implementation. Section 2.2 discusses the roles of model parameters and gives recommended prior settings. Section 2.3 briefly outlines the MCMC algorithm used for posterior inferences and addresses implementation. Finally, Section 2.4 discusses model inferences, including transition density estimation.

2.1 Model specification

One avenue to arrive at the conditional density form in (1) begins with a prior for joint density estimation. For clarity in notation, we use $y \in \mathbb{R}$ to represent the current observation and $\boldsymbol{x} \equiv \boldsymbol{y}_{t-1} \in \mathbb{R}^L$ to denote the lags. We begin as in Müller et al. (1996), who in the regression setting consider y and $\boldsymbol{x}$ to arise jointly from a Gaussian DPM. This implies the stick-breaking representation (Sethuraman, 1994) for joint density,

$$f_{YX}(y, \boldsymbol{x} \mid G) = \sum_{h=1}^{\infty} \omega_h \, \mathrm{N}\left((y, \boldsymbol{x}) \mid \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h\right), \tag{2}$$

where the $(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)$ arise i.i.d. from the Dirichlet process (DP) centering distribution G_0, and the mixture weights, ω_h, are constructed as

$$\omega_1 = v_1, \ \omega_h = v_h \prod_{j=1}^{h-1} (1 - v_j), \ \text{for } h > 2, \ \text{and } v_h \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha). \qquad (3)$$

Conditioning on $\boldsymbol{x}$, we obtain the transition density model:

$$f_{Y|X}(y \mid \boldsymbol{x}, G) = \frac{\sum_{h=1}^{\infty} \omega_h \, \mathrm{N}_{(h)}(\boldsymbol{x}) \, \mathrm{N}_{(h)}(y \mid \boldsymbol{x})}{\sum_{j=1}^{\infty} \omega_j \, \mathrm{N}_{(j)}(\boldsymbol{x})} = \sum_{h=1}^{\infty} q_h(\boldsymbol{x}) \, \mathrm{N}(y_t \mid \mu_h(\boldsymbol{x}), \sigma_h^2), \qquad (4)$$

with $q_h(\boldsymbol{x}) = \omega_h \, \mathrm{N}_{(h)}(\boldsymbol{x}) / \sum_{j=1}^{\infty} \omega_j \, \mathrm{N}_{(j)}(\boldsymbol{x})$, where $\mathrm{N}_{(h)}(\cdot)$ refers to a Gaussian density with parameters corresponding to mixture component h, and $\mathrm{N}_{(h)}(y \mid \boldsymbol{x})$ is the univariate conditional Gaussian density derived from $\mathrm{N}_{(h)}(y, \boldsymbol{x})$. The joint densities in each mixture component of the numerator of (4) have been factored into their respective marginal L-dimensional Gaussian density for $\boldsymbol{x}$ (with mean $\boldsymbol{\mu}^x$ and covariance $\boldsymbol{\Sigma}^x$) and univariate conditional Gaussian density for y (with mean $\mu(\boldsymbol{x}) \equiv \mu^y + \boldsymbol{\Sigma}^{yx}(\boldsymbol{\Sigma}^x)^{-1}(\boldsymbol{x} - \boldsymbol{\mu}^x)$ and variance $\sigma^2 \equiv (\sigma^y)^2 - \boldsymbol{\Sigma}^{yx}(\boldsymbol{\Sigma}^x)^{-1}\boldsymbol{\Sigma}^{xy})$. The second line of (4) reveals the local linear model structure with lag-dependent weights and mixture kernels with means depending linearly on $\boldsymbol{x}$.

This procedure yields a conditional density that satisfies the requirements of the proposed model (1). Specifically, since $\sum_{h=1}^{\infty} \omega_h = 1$ almost surely, so long as there exists some positive constant $c_N < +\infty$ such that $0 < \mathrm{N}_{(h)}(\boldsymbol{x}) < c_N$ for all $h \in \mathbb{N}$ and all $\boldsymbol{x} \in \mathbb{R}^L$ (which is satisfied if there exists another constant $c_\Sigma > 0$ such that $\det(\boldsymbol{\Sigma}_h^x) > c_\Sigma$ for all $h \in \mathbb{N}$), the denominator in $q_h(\boldsymbol{x})$ will be positive and finite for all $\boldsymbol{x} \in \mathbb{R}^L$.

Although $\boldsymbol{x}$ (representing $\boldsymbol{y}_{t-1}$) can legitimately be considered random in the time-series context, the Markovian likelihood requires that the conditional density (4) form the basis of the model. Furthermore, unless we impose stationarity restrictions on the process $\{y_t\}$, the joint mixture density in (2) cannot apply to consecutive length-$(L+1)$ coordinate vectors such as $(y_{t+1}, \boldsymbol{y}_t)$ and $(y_t, \boldsymbol{y}_{t-1})$. To achieve greater flexibility in transition density estimation, we elect to forego a stationarity assumption. This carries two important consequences: first, the model does not estimate a joint density for consecutive observations, so that the density implied by reversing (4) is not interpretable; second, the densities $\{\mathrm{N}_{(h)}(\boldsymbol{x})\}$ serve only to support the function of the mixture weights $\{q_h(\boldsymbol{x})\}$ and are not related to a joint probability distribution for lags. Supplement S1 includes discussion of possible mixture model formulations for the stationary case.

The model likelihood, based on (4) and conditional on the first L observations, is $\prod_{t=L+1}^{T} f_{Y|X}(y_t \mid \boldsymbol{y}_{t-1}, G)$. This is the form adopted in Antoniano-Villalobos and Walker (2016) and Kalli and Griffin (2018), who assume stationarity, and DeYoreo and Kottas (2017), who do not assume stationarity. The local re-weighting of $\{\omega_h\}$ with probability density kernels on $\boldsymbol{x}$ distinguishes our model from nonparametric extensions of MoE for regression, such as DDP-variants (Chung and Dunson, 2009; Fuentes-García et al., 2009; Barrientos et al., 2017) and kernel stick-breaking models (Park and Dunson, 2010; Reich et al., 2012). See Wade et al. (2014) and DeYoreo and Kottas (2020) for reviews of density

regression models following Müller et al. (1996), which do not pre-condition the likelihood.

2.1.1 Covariance factorization

To facilitate interpretation in our factorization of the kernels into response and lag densities, allow flexible and parsimonious covariance modeling, and to provide a vehicle for variable selection in the mixture weights, we parameterize the Gaussian covariance matrix according to the factorization $\boldsymbol{\Sigma} = \boldsymbol{B}^{-1}\boldsymbol{\Delta}(\boldsymbol{B}^{-1})'$. Here, $\boldsymbol{\Delta} = \text{diag}(\sigma^2, \delta_1^x, \ldots, \delta_L^x)$ and $\boldsymbol{B}$ is an upper unit-triangular matrix with first row $(1, \beta_1^y, \beta_2^y, \ldots, \beta_{L-1}^y, \beta_L^y)$, second row $(0, 1, \beta_{1,2}^x, \ldots, \beta_{1,L-1}^x, \beta_{1,L}^x)$, and so forth until the $(L-1)$th row $(0, \ldots, 0, 1, \beta_{L-1,L}^x)$. This factorization is equivalent to the square-root-free Cholesky decomposition employed by Daniels and Pourahmadi (2002) and Webb and Forster (2008), and in our setting by DeYoreo and Kottas (2017). This and similar decompositions have also been used for model selection (Smith and Kohn, 2002; Cai and Dunson, 2006). Our extension for lag selection in the mixture weights is discussed in Section 4.

This parameterization also yields a sequential decomposition of a joint Gaussian density for y and $\boldsymbol{x}$ into $L + 1$ univariate Gaussian densities. Specifically,

$$
\text{N}\left((y, \boldsymbol{x}) \mid \boldsymbol{\mu}, \boldsymbol{B}^{-1}\boldsymbol{\Delta}(\boldsymbol{B}^{-1})'\right) = \text{N}(x_L \mid \mu_L^x, \delta_L^x) \prod_{\ell=L-1}^{1} \text{N}\left(x_\ell \mid \mu_\ell^x - \sum_{r=\ell+1}^{L} \beta_{\ell,r}^x(x_r - \mu_r^x), \delta_\ell^x\right) \times
$$
$$
\text{N}\left(y \mid \mu^y - \sum_{\ell=1}^{L} \beta_\ell^y(x_\ell - \mu_\ell^x), \sigma^2\right). \tag{5}
$$

We construct from back (most distant lag) to front (y) so that the response density depends on the entire $\boldsymbol{x}$ vector while maintaining a consistent order convention. This fully parameterized representation of the covariance matrix is flexible, as each β parameter is unrestricted and δ parameters need only be positive, and admits control over the marginal weight density of $\boldsymbol{x}$ while preserving positive definiteness. Note also that the marginal covariance matrix of $\boldsymbol{x}$ can be constructed as $\boldsymbol{\Sigma}^x = (\boldsymbol{B}^x)^{-1}\boldsymbol{\Delta}^x((\boldsymbol{B}^x)^{-1})'$ where $\boldsymbol{B}^x$ removes the top row and first column of $\boldsymbol{B}$, and $\boldsymbol{\Delta}^x = \text{diag}(\delta_1^x, \ldots, \delta_L^x)$.

The weight kernels in $q_h(\boldsymbol{x})$ present the most obvious and pressing opportunity to improve parameter economy in the model. We therefore also consider weight kernels with local independence between elements of $\boldsymbol{x}$ (e.g., Shahbaba and Neal, 2009). This reduction is accomplished by setting all $\beta_{\ell,r}^x$, for $\ell \neq r$, equal to 0, yielding diagonal $\boldsymbol{\Sigma}^x = \boldsymbol{\Delta}^x$. We note that Gaussian mixtures with diagonal covariance can approximate general density shapes, at the cost of possibly utilizing additional mixture components to capture local behavior. The reduction becomes necessary if we include many lags, as the number of covariance parameters for *each* component h grows quadratically with L.

The final term in (5) involving μ^y and the $\{\beta_\ell^y\}$ is overparameterized if used for regression with one mixture component. However, the $\{\mu_\ell^x\}$ parameters have an integral role in the weight functions of the mixture model (4), providing (with exception of mixture label switching) at least weak identifiability. It is nevertheless preferable to monitor inferences for

component-specific intercepts $\mu^y + \sum_{\ell=1}^{L} \beta_\ell^y \mu_\ell^x$, which in our experience are far more stable than either μ^y or $\{\mu_\ell^x\}$ alone.

2.1.2 Hierarchical model formulation

To implement the model, we truncate the infinite summation needed to normalize the mixture weights $\{q_h(\boldsymbol{x})\}$, using blocked Gibbs sampling (Ishwaran and James, 2001). There are both theoretical and practical considerations when selecting the truncation level, H. Given the DP concentration parameter α, we can calculate the prior expected truncation error, $\mathrm{E}(\omega_H) = \mathrm{E}(\prod_{h=1}^{H-1}(1 - v_h)) = [\alpha/(1 + \alpha)]^{H-1}$. We can also monitor throughout MCMC sampling the last weight, ω_H, to ensure it remains small, as well as the number of occupied components to ensure that it does not approach H.

As is common with similar models, we break the mixture by introducing latent variables $\{s_t\}$ associated with each time point, such that if $s_t = h$, the observation at time t is assigned to component h. We denote all component-specific parameters as $\{\boldsymbol{\eta}_h\}_{h=1}^H$ where $\boldsymbol{\eta} \equiv \{\mu^y, \boldsymbol{\mu}^x, \boldsymbol{\beta}^y, \boldsymbol{\beta}_1^x, \ldots, \boldsymbol{\beta}_{L-2}^x, \beta_{L-1}^x, \sigma^2, \boldsymbol{\delta}^x\}$, with vectors $\boldsymbol{\beta}^y$ and $\boldsymbol{\beta}_\ell^x$ (for $\ell = 1, \ldots, L-2$), and $\beta_{L-1}^x \equiv \beta_{L-1,L}^x$ taken from the corresponding rows of $\boldsymbol{B}$, and $\boldsymbol{\delta}^x = (\delta_1^x, \ldots, \delta_L^x)$. Again, we use notation $\mathrm{N}_{(h)}(\cdot)$ to indicate that all parameters used to specify the mean and covariance are indexed by h. The hierarchical formulation of our model is given by

$$y_t \mid \boldsymbol{y}_{t-1}, s_t = h, \{\boldsymbol{\eta}\} \overset{\text{ind.}}{\sim} \mathrm{N}_{(h)} \left(y_t \mid \mu^y - \sum_{\ell=1}^{L} \beta_\ell^y (y_{t-\ell} - \mu_\ell^x), \sigma^2 \right),$$

$$\text{for } t = L+1, \ldots, T, \text{ and } h = 1, \ldots, H,$$

$$\Pr(s_t = h \mid \boldsymbol{y}_{t-1}, \{\boldsymbol{\eta}\}, \boldsymbol{\omega}) = \frac{\omega_h \, \mathrm{N}_{(h)}(\boldsymbol{y}_{t-1} \mid \boldsymbol{\mu}^x, \boldsymbol{\Sigma}^x)}{\sum_{j=1}^{H} \omega_j \, \mathrm{N}_{(j)}(\boldsymbol{y}_{t-1} \mid \boldsymbol{\mu}^x, \boldsymbol{\Sigma}^x)}, \tag{6}$$

$$\omega_1 = v_1, \ \omega_h = v_h \prod_{j=1}^{h-1}(1 - v_j), \text{ for } j = 2, \ldots, H-1, \text{ and } \omega_H = \prod_{j=1}^{H-1}(1 - v_j),$$

$$v_j \mid \alpha \overset{\text{iid}}{\sim} \mathrm{Beta}(1, \alpha), \text{ for } j = 1, \ldots, H-1,$$

$$\boldsymbol{\eta}_h \mid G_0 \overset{\text{iid}}{\sim} G_0(\boldsymbol{\eta}_h), \text{ for } h = 1, \ldots, H,$$

with $G_0(\boldsymbol{\eta}) = \mathrm{N}((\mu^y, \boldsymbol{\beta}^y) \mid \sigma^2) \times \mathrm{IG}(\sigma^2) \times \mathrm{N}(\boldsymbol{\mu}^x) \times \prod_{r=1}^{L-1} \mathrm{N}(\boldsymbol{\beta}_r^x) \times \prod_{\ell=1}^{L} \mathrm{IG}(\delta_\ell^x)$, and $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_H)$. Here $\mathrm{N}((\mu^y, \boldsymbol{\beta}^y) \mid \sigma^2)$ indicates that the prior covariance matrix for $\boldsymbol{\beta}^* \equiv (\mu^y, \boldsymbol{\beta}^y)$ is scaled by σ^2, which allows us to analytically integrate all y-indexed parameters from the full conditional for $\boldsymbol{\eta}_h$ and improve mixing in MCMC (discussed in Section 2.3).

We complete the model with a $\mathrm{Ga}(a_\alpha, b_\alpha)$ prior for α, and with conditionally conjugate priors on the parameters in G_0. Specifically, the $(L+1)$-variate Gaussian distribution for $\boldsymbol{\beta}^*$ has mean $\boldsymbol{\beta}_0^* \sim \mathrm{N}(\boldsymbol{b}_0^*, \boldsymbol{S}_0^*)$ and covariance $\sigma^2(\boldsymbol{\Lambda}_0^*)^{-1}$ with $(\boldsymbol{\Lambda}_0^*)^{-1} \sim \mathrm{IWish}(\nu^*, \nu^* \boldsymbol{\Psi}_0^*)$ (an inverse-Wishart distribution with ν^* degrees of freedom and mean $\nu^* \boldsymbol{\Psi}_0^* / [\nu^* - (L+1) - 1]$, parameterized so that $\boldsymbol{\Psi}_0^*$ is the prior harmonic mean of $(\boldsymbol{\Lambda}_0^*)^{-1}$). The inverse-gamma distribution for σ^2 has fixed shape $\nu_{\sigma^2}/2$ and scale $\nu_{\sigma^2} s_0/2$, yielding for σ^2 a prior harmonic mean of $s_0 \sim \mathrm{Ga}(a_{s_0}, b_{s_0})$ (which itself has mean a_{s_0}/b_{s_0}). The Gaussian distribution

for $\boldsymbol{\mu}^x$ has mean $\boldsymbol{\mu}_0^x \sim \mathrm{N}(\boldsymbol{m}_0^x, \boldsymbol{S}_0^{\mu_x})$ and covariance $(\boldsymbol{\Lambda}^{\mu_x})^{-1} \sim \mathrm{IWish}(\nu^{\mu_x}, \nu^{\mu_x}\boldsymbol{\Psi}_0^{\mu_x})$. The Gaussian distribution for each $\boldsymbol{\beta}_r^x$ has mean $\boldsymbol{\beta}_{0,r}^x \overset{\text{ind.}}{\sim} \mathrm{N}(\boldsymbol{b}_{0,r}^{\beta_x}, \boldsymbol{S}_{0,r}^{\beta_x})$ and covariance $(\boldsymbol{\Lambda}_{0,r}^{\beta_x})^{-1} \overset{\text{ind.}}{\sim}$ $\mathrm{IWish}(\nu_r^{\beta_x}, \nu_r^{\beta_x}\boldsymbol{\Psi}_{0,r}^{\beta_x})$, for $r = 1, \ldots, L-1$. The inverse-gamma distribution for each δ_ℓ^x has fixed shape $\nu_\ell^{\delta^x}/2$ and scale $\nu_\ell^{\delta^x} s_{0,\ell}^x/2$ with $s_{0,\ell}^x \overset{\text{ind.}}{\sim} \mathrm{Ga}(a_{s_0,\ell}^x, b_{s_0,\ell}^x)$, for $\ell = 1, \ldots, L$.

2.2 Prior settings

The priors for the hierarchical model in Section 2.1.2 are specified in generality so that the model can be fit with the time series $\{y_t\}$ at any scale and for a variety of functional characteristics. However, one may consider first removing certain known trend and cyclical behaviors, and basing hyperparameter settings on default values. Here, we recommend default values derived from the marginal center and range of the time-series data.

We first discuss the function and interpretation of model parameters. A key consideration is that model (4) is a locally weighted mixture of Gaussian linear regression models. The weight structure depends not only on $\{\omega_h\}$, which is inherited from the nonparametric prior and (for low values of α) encourages economy in clustering, but also on the Gaussian kernels for $\boldsymbol{x}$. One could imagine a normalized weight surface spanning $\mathbb{R}^L$ for each mixture component h that follows the contours of a L-variate Gaussian density weighted by ω_h. The component-specific, x-indexed parameters, $\boldsymbol{\mu}^x$ and $\boldsymbol{\Sigma}^x = (\boldsymbol{B}^x)^{-1}\boldsymbol{\Delta}^x((\boldsymbol{B}^x)^{-1})'$, determine the locations and shapes of the weight kernels. The y-indexed parameters, μ^y and $\boldsymbol{\beta}^y$, provide the component-conditional mean as a first-order linear combination of $\boldsymbol{x}$, and σ^2 provides observation error variance around the component's mean.

One primary functional of interest derived from the transition density in (4) is the conditional expectation $\mathrm{E}(y \mid \boldsymbol{x}) = \sum_h q_h(\boldsymbol{x})\,\mu_h(\boldsymbol{x})$, to which we refer as the transition mean. A modeler can encode beliefs about this functional relationship between y and $\boldsymbol{x}$ through the priors for α and parameters in the base measures for $\boldsymbol{\Sigma}^x$ and σ^2. By influencing the number of active components, α assists in controlling how many times the transition mean can change directions. To encourage smooth behavior, one may use a prior favoring relatively large variances in $\boldsymbol{\Sigma}^x$, most directly through the priors for $\{\delta_\ell^x\}$. To encourage active local behavior, including nearly discontinuous transitions, one would use small variances in $\boldsymbol{\Sigma}^x$ to allow the components to concentrate on small regions, analogous to using many knots in spline models. Supplement S2 further explores the effect of prior settings on transition means.

We recommend the following default settings for a baseline prior, which in most cases should be adjusted for the analysis at hand. We typically set a_α in the interval $[5, 15]$, depending on our prior beliefs about the degree of nonlinearity in the transition function. Setting $b_\alpha = 1$ yields a prior mean of a_α. Antoniak (1974) gives the expression $\alpha \log((\alpha + T - L)/\alpha)$ as a rough prior estimate for the number of components. While this applies in the prior joint model, the number of components in our conditional model (4) is also a function of the Gaussian weight kernels on $\boldsymbol{x}$. We set $\boldsymbol{b}_0^* = (\bar{y}, 0, \ldots, 0)$, with $\bar{y}$ representing the center of the time series, empirical or user-defined, and $\boldsymbol{S}_0^* = \mathrm{diag}([\mathrm{range}(y)/6.0]^2, 1.0, \ldots, 1.0)$, with $\mathrm{range}(y)$ representing the range of the time series, empirical or user-defined. This specifica-

tion provides reasonable flexibility for the entry in $\boldsymbol{b}_0^*$ corresponding to μ^y and encourages centering the remaining entries of $\boldsymbol{\beta}_0^*$ near 0. To avoid extreme values in $(\boldsymbol{\Lambda}_0^*)^{-1}$ (the covariance matrix for $\boldsymbol{\beta}_0^*$) occasionally encountered during MCMC, and thus promote stability and identifiability, we set $\nu^* = 50\,(L + 1 + 2)$ and $\boldsymbol{\Psi}_0^* = s_{00}^{-1}\,\text{diag}([\text{range}(y)/2.0]^2, 16.0, \ldots, 16.0)$, where s_{00} is a user-supplied prior estimate of σ^2. The prior estimate s_{00} partially compensates and controls for the fact that the covariance for $\boldsymbol{\beta}^*$ in G_0 is multiplied by σ^2. The base measure for σ^2 is largely application-specific, but we use $\nu_{\sigma^2} \in [5.0, 10]$ with $a_{s_0} = n_{s_0}\,\nu_{\sigma^2}/2$ and $b_{s_0} = n_{s_0}\,\nu_{\sigma^2}/(2\,s_{00})$, where $n_{s_0} \in [5.0, 10.0]$ is a sample-size equivalent and $s_{00} = [\text{range}(y)/6.0]^2/\mathcal{R}$ is the prior mean of s_0. The squared quantity is divided by a prior signal-to-noise ratio $\mathcal{R} > 0$ that should be set on a case-by-case basis (we typically use $\mathcal{R} \in [5.0, 10.0]$). We use $\boldsymbol{m}_0^x = \bar{y}\,\boldsymbol{1}$ and $\boldsymbol{S}_0^{\mu_x} = [\text{range}(y)/6.0]^2\,\boldsymbol{I}_L$, where $\boldsymbol{I}_k$ denotes a $k \times k$ identity matrix. We allow for variety in $\boldsymbol{\mu}^x$ by setting $\nu^{\mu_x} = 10\,(L + 2)$ and $\boldsymbol{\Psi}_0^{\mu_x} = [\text{range}(y)/2.0]^2\,\boldsymbol{I}_L$. Similarly, we set each $\boldsymbol{b}_{0,r}^{\beta_x} = \boldsymbol{0}$, each $\boldsymbol{S}_{0,r}^{\beta_x} = \boldsymbol{I}_L$, each $\nu_r^{\beta_x} = 10\,(L + 2)$ and $\boldsymbol{\Psi}_{0,r}^{\beta_x} = 2.0\,\boldsymbol{I}_{L-1-r+1}$, for $r = 1, \ldots, L - 1$. Finally, we set $\nu_\ell^{\delta_x} = 5.0$, with $a_{s_0,\ell}^x = n_{s_0,\ell}^x\,\nu_\ell^{\delta_x}/2$ and $b_{s_0,\ell}^x = n_{s_0,\ell}^x\,\nu_\ell^{\delta_x}/(2\,s_{00,\ell}^x)$, for $\ell = 1, \ldots, L$, where $n_{s_0,\ell}^x = 5.0$ and $s_{00,\ell}^x = [\text{range}(y)/8.0]^2$.

While the preceding prior settings provide a good starting point in general, they are not always appropriate. We recommend considering alternate settings, especially for α, and parameters in the base measures for $\boldsymbol{\Sigma}^x$ and σ^2, depending on prior beliefs about the functional relationship being modeled in each analysis. We further recommend checking for sensitivity of inferences for important quantities to these and other prior settings.

2.3 Computation

We briefly outline the MCMC algorithm used to obtain posterior samples from the proposed model. Further details are given in Supplement S3. We employ a Gibbs sampler with a variety of update methods for parameter blocks, which proceeds by successively sampling the parameters in the sets and manner described below.

Latent states: The latent states identifying component membership for each observation y_t are updated individually, each using a Metropolized Gibbs step (Liu, 1996) based on discrete full conditional distributions involving $\{\omega_h\}$, the weight kernel density for $\boldsymbol{y}_{t-1}$, and the kernel density for y_t.

Stick-breaking weights: The DP weights $\{\omega_h\}_{h=1}^H$ are defined through the latent $\{v_h\}_{h=1}^{H-1}$ which, conditional on component membership $\{s_t\}$, admit $H - 1$ independent beta full conditional distributions in standard DPM models (Ishwaran and James, 2001). The normalization term in each likelihood contribution of $q_h(\boldsymbol{y}_{t-1})$ complicates the full conditional distribution in our model. It is unchanged from the distribution reported in DeYoreo and Kottas (2017), with the exception that the kernels are now multivariate Gaussian on the vector $\boldsymbol{y}_{t-1}$. This adjustment yields numerical instability and poor mixing in the one-at-a-time slice sampler employed by DeYoreo and Kottas (2017). To obtain direct samples from this distribution, we instead employ the multivariate hyper-rectangle slice sampler of Neal (2003) to update all v_h, $h = 1, \ldots, H - 1$, simultaneously.

Component-specific parameters: To facilitate mixing of the y-indexed, component-

specific parameters, we partition $\boldsymbol{\eta}$ into its y and x components $\boldsymbol{\eta}^y \equiv \{\mu^y, \boldsymbol{\beta}^y, \sigma^2\}$ and $\boldsymbol{\eta}^x \equiv \{\boldsymbol{\mu}^x, \boldsymbol{\beta}_1^x, \ldots, \boldsymbol{\beta}_{L-1}^x, \boldsymbol{\delta}^x\}$, and sample $p(\boldsymbol{\eta}_h \mid \cdots) = p(\boldsymbol{\eta}_h^x \mid \cdots, -\boldsymbol{\eta}_h^y) \, p(\boldsymbol{\eta}_h^y \mid \boldsymbol{\eta}_h^x, \cdots)$, where $p(\boldsymbol{\eta}_h^x \mid \cdots, -\boldsymbol{\eta}_h^y) = \int p(\boldsymbol{\eta}_h \mid \cdots) \, \mathrm{d}\boldsymbol{\eta}_h^y$. The weight normalization terms in $q_h(\boldsymbol{y}_{t-1})$ preclude simple conjugate updates of $\boldsymbol{\eta}_h^x$, for which we employ a random-walk Metropolis step. This is then followed by an exact draw from the full conditional distribution of $\boldsymbol{\eta}_h^y$.

DP prior hyperparameters: All parameters of the DP centering distribution have conditionally conjugate updates. For computational stability, our implementation fixes, rather than updates, the parameters in G_0 associated with $\boldsymbol{\eta}^y$. Prior estimate values are used for the fixed values, e.g., $\boldsymbol{b}_0^*$ for $\boldsymbol{\beta}_0^*$. Finally, the DP concentration parameter α has a gamma posterior full conditional distribution with shape $a_\alpha + H - 1$ and rate $b_\alpha - \log(\omega_H)$.

We typically initialize MCMC chains at default prior settings such as the prior mean or applicable summary value from the next level of the hierarchy, or with draws from the prior model (usually with G_0 fixed). The primary exception is the initial allocation to components $\{s_t\}$, for which we use output from a clustering algorithm applied to $(y_t, y_{t-1}, \ldots, y_{t-L})$, for all $t = L+1, \ldots, T$. For example, we use hierarchical clustering with Euclidean distance and Ward linkage to assign the observations into H clusters. The sampler is then run for one or several rounds of tuning or adaptation, as described in Supplement S3. If adaptation is used, scaled empirical covariance matrices inform subsequent random-walk proposals. After a specified burn-in period, samples are collected for inference.

In our experience, the weakly identified $\boldsymbol{\eta}^x$ and $\boldsymbol{\omega}$ parameters present the primary mixing challenge. This appears to indicate redundancy in the weight functions, for which many configurations produce similar results. Our illustrations with the base model (i.e., without lag selection) focus on low-order dependence $L \leq 5$. Later illustrations use diagonal $\boldsymbol{\Sigma}^x = \boldsymbol{\Delta}^x$. We further aid mixing by iterating between adaptation and pre-burn-in runs before beginning a final burn-in run. We note that despite the mixing challenges, MCMC chains for parameters and functionals of interest are typically stable.

2.4 Transition density estimation

Posterior samples from the model yield rich inferences regarding the transition distribution for a time series. The three of most interest to us are the transition density, the transition mean functional, and inferences for relevant lags. We incorporate the latter in Section 4. The transition mean functional and estimates of the transition density are straightforward to compute, as the stick-breaking representation and blocked Gibbs sampler yield an approximation of the random mixing distribution G at each MCMC iteration. For any value of y and $\boldsymbol{x}$, or over a multidimensional grid of values, one can use posterior samples of parameters to calculate pointwise samples of the finite-truncated version of $f_{Y|X}$ in (4), given as $\tilde{f}_{Y|X}(y \mid \boldsymbol{x}) = \sum_{h=1}^H \tilde{q}_h(\boldsymbol{x}) \, \mathrm{N}_{(h)}(y \mid \mu(\boldsymbol{x}), \sigma^2)$, with $\tilde{q}_h(\boldsymbol{x}) = \omega_h \, \mathrm{N}_{(h)}(\boldsymbol{x}) / \sum_{j=1}^H \omega_j \, \mathrm{N}_{(j)}(\boldsymbol{x})$ and $\mu(\boldsymbol{x}) = \mu^y - \sum_{\ell=1}^L \beta_\ell^y \, (x_\ell - \mu_\ell^x)$. The samples can then be used to construct point and interval estimates for the transition density. Other functionals such as the transition mean or quantiles are similarly obtained. One can calculate the transition mean for each posterior sample with $\tilde{\mathrm{E}}_{Y|X}(y \mid \boldsymbol{x}) = \sum_{h=1}^H \tilde{q}_h(\boldsymbol{x}) \, \mu_{(h)}(\boldsymbol{x})$ over a grid of values for $\boldsymbol{x}$, yielding pointwise estimates and intervals. We obtain samples of the $u \in (0, 1)$ quantile of the transition

density by solving for the unique root of $\tilde{Q}_u(y \mid \boldsymbol{x}) = u - \sum_{h=1}^{H} \tilde{q}_h(\boldsymbol{x}) \, \Phi \left([y - \mu_{(h)}(\boldsymbol{x})]/\sigma_{(h)}\right)$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

Monte Carlo estimates of K-step-ahead forecasts can be obtained by inductively simulating $(s, y)_{T+k}$ pairs, for $k = 1, \ldots, K$, following the first two levels of the hierarchical model (6) for each posterior sample. Such samples propagate both forecast and inferential uncertainty, and can be useful for assessing model performance with validation data.

3   Data illustrations

We illustrate the proposed model with two examples. The first synthetic data example highlights some key features and potential uses of the model. The real data example illustrates the model's utility for lag-dependent density estimation. Two default prior settings were utilized in each case, with one promoting a higher signal variance through prior signal-to-noise ratio $\mathcal{R} = 8.0$ instead of the default $\mathcal{R} = 5.0$. For each model fit, multiple MCMC chains were randomly initialized using the strategy described in Section 2.3, followed by iterative tuning (no adaptation) and 300,000 burn-in samples. The next 500,000 iterations were then thinned to 5,000 for inference (plots in the following illustrations generally use 1,000 or 2,000 of these). Inferences are reported for one of the chains. These values for burn-in and thinning are fairly conservative; shorter chains often suffice.

3.1   Simulated data: Ricker model

We begin with a time series simulated from an adaptation of a classical model for population dynamics (Ricker, 1954). The series was generated from

$$y_t = y_{t-2} \exp(2.6 - y_{t-2}) + \epsilon_t, \quad \epsilon_t \overset{\text{iid}}{\sim} \mathrm{N}(0, (0.09)^2), \tag{7}$$

featuring first-order nonlinear dynamics as a function of the second lag only. We fit the model to the original real-valued time series with $L = 2$, $T = 72$ (so that 70 observations contribute to the likelihood), and $H = 40$. The $\mathcal{R} = 5.0$ fit resulted in three chains with similar traces of the log-likelihood and occupied mixture components (always at two). All traces of σ^2 for the most occupied cluster (not shown) converge to approximately 3.5 times the true value of 0.0081, due in part to the prior estimate $s_{00} = 0.119$. Flexibility and prior bias in error variance, together with low sample size, result in a transition mean fit that locally mixes two planes, capturing the general shape, but missing curvature in the region $y_{t-2} \in (0, 1)$ (not shown). Two of three chains with higher signal-to-noise ratio ($\mathcal{R} = 8.0$) use a third mixture component to better capture this curvature (although one reverts back to two components), as demonstrated for one chain in Figure 1.

The dynamics are reasonably recovered in data-rich regions of the phase space despite using an over-specified model with two lags on a short time series. We can informally assess the influence of the first lag with the second-order model by checking for sensitivity of inferences for the transition mean to different values of the first lag. For example, the left panel of Figure 1 plots estimates for the transition mean over a grid of values for the second
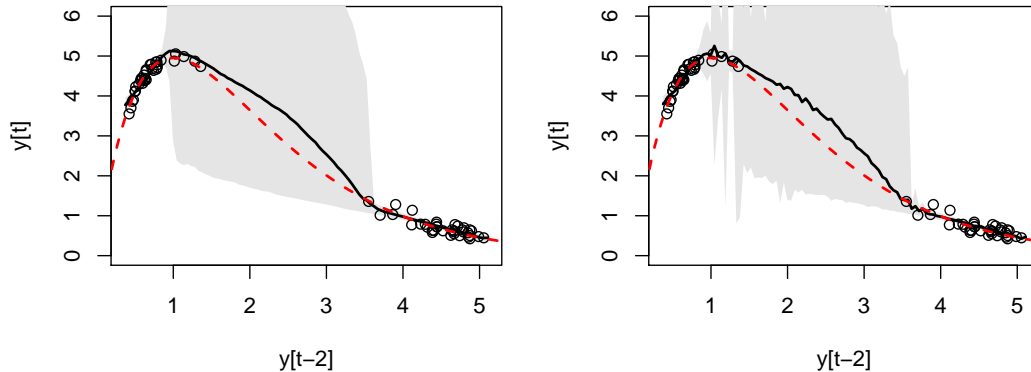
Figure 1: Model fit to the single-lag dynamical simulation with noise ($T = 72$, $L = 2$, $\mathcal{R} = 8.0$), depicting the posterior mean and 95% interval estimates for the transition mean function over a grid of values for lag 2. In the left panel, values of the first lag (y_{t-1}) were fixed at a mean value. In the right panel, values for the first lag were randomly drawn uniformly over the range of observed values. Observations are included, as well as the true transition map (dashed red line).

lag, in which all values for the first lag have been fixed at their mean. The right panel replicates this plot with grid values for the first lag drawn uniformly over the range of the data. This perturbation has minimal effect, especially where data are observed, suggesting that lag 1 is negligible in the model fit. We note particularly wide credible intervals in the data-sparse region, which approximately reach 10. This appears to stem from the weight functions concentrating locally around the data, leaving data-sparse regions to revert to an indecisive mixture of the component fits and prior.

3.2   Old Faithful data

Antoniano-Villalobos and Walker (2016) and DeYoreo and Kottas (2017) both illustrate single-lag versions of our proposed model with the well-known inter-eruption waiting times of the Old Faithful geyser in Yellowstone National Park, U.S.A. The time series has attracted attention, both for illustration and analysis from chaos (Nicholl et al., 1994) and statistical (Azzalini and Bowman, 1990) perspectives, partly due to nonlinear as well as non-Gaussian dynamics. We revisit Old Faithful using the traditional data set reported in Azzalini and Bowman (1990), consisting of 299 consecutive pairs of eruption durations and waiting times between August 1 and 15, 1985. Figure 2 shows a trace of eruption waiting times in minutes.

We fit the proposed model to the final $T = 291$ observations with $L = 2$ and $H = 40$. Likelihood traces are similar among runs under both prior signal-to-noise ratios, switching (infrequently) between values corresponding to two and three occupied mixture components. Estimated transition mean surfaces, one of which is shown in Figure 3 (left), are primarily driven by the first lag, with minor tilt along the second. The transition mean functional is less informative for values of y_{t-1} above 70 minutes, when the transition distribution becomes bimodal. In this region, estimates of transition quantiles may be more appropriate than the transition mean. Inferences for quantiles over a grid of fixed lag values are easily
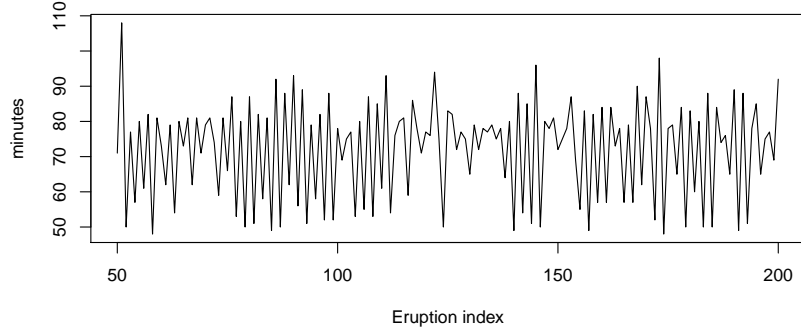
Figure 2: Trace of 150 consecutive Old Faithful eruption waiting times in minutes. This window of the middle half of the time series typifies the data, with exception of the run of long waiting times between index 120 and 140.
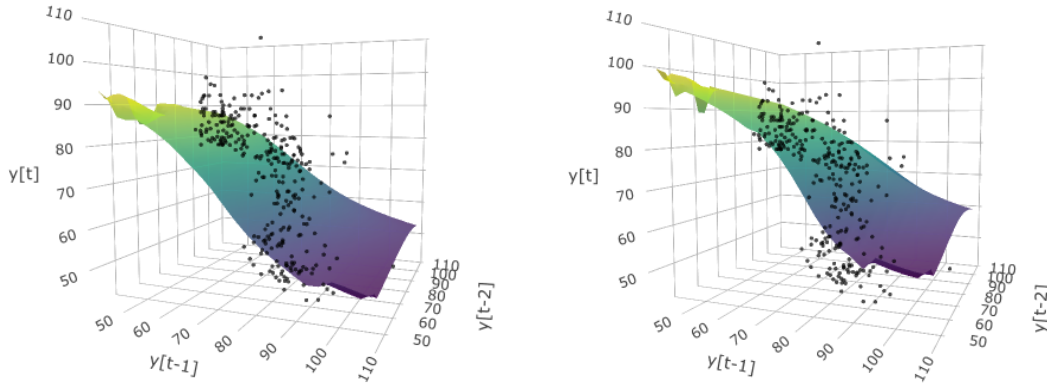

Figure 3: Nonparametric model fit to Old Faithful waiting times in minutes ($T = 291$, $L = 2$, $\mathcal{R} = 5.0$), with posterior mean estimates of the transition mean (left) and 0.8 quantile (right) surfaces. Observed transitions are included as points.

obtained from posterior samples by following the procedure described in Section 2.4. Figure 3 (right) shows a pointwise posterior mean estimate of the 0.8 quantile surface as a function of the two lags. Credible intervals for both surfaces (excluded for simplicity in the plots) are reasonable, falling within the range of the data.

Figure 4 shows estimated transition densities (posterior mean and 95% credible intervals) for three values of the two lags. These estimates demonstrate the density autoregressive feature of the model, which in this case successfully captures density dependence on lags. Interestingly, the transition density undergoes noticeable change between $y_{t-2} = 50$ and $y_{t-2} = 80$ when y_{t-1} is fixed at 80 minutes, suggesting dependence on the second lag. Other runs show similar structure. A simple analysis using a discrete-state Markov chain on a dichotomized version of the time series further supports second-order dependence.
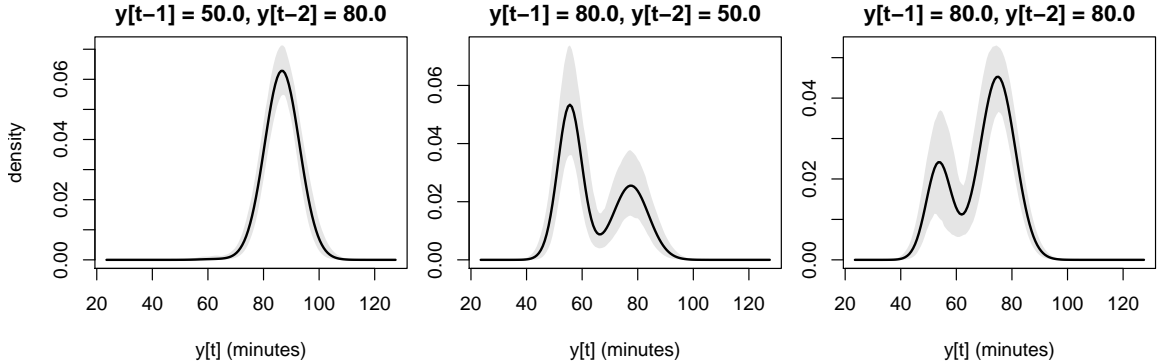
Figure 4: Posterior mean and 95% interval estimates for the transition density of Old Faithful waiting times at three pairs of fixed values of the first two lags ($T = 291$, $L = 2$, $\mathcal{R} = 5.0$).

4 Lag selection

We now discuss extending model (4) to include inferences for relevant lags. This step is important in many applications, as dependence may extend beyond the most recent lags. In some cases, not all recent lags are important. Methods for state-space reconstruction require a minimal number of lags to "unfold" an attractor, but using too many can be inefficient, or render estimation impractical. Reducing system dimensionality to the minimum necessary for fitting the data further simplifies posterior analysis and model interpretation. Our approach is to pre-specify a maximal lag horizon L, and fit an encompassing model that accommodates up to all L lags, but shrinks to select only those that significantly contribute to the transition density.

In the time series literature, autoregressive order is often assessed with standard information criteria, which can include regularization (Khalili et al., 2017). Bayesian approaches typically involve stochastic-search-type algorithms, and several are presented in Prado and West (2010, Ch. 2). In the stationary, linear case, one can use the specialized priors of Huerta and West (1999) on roots of the AR characteristic polynomial to infer order. Wood et al. (2011) employ a two-stage MCMC sampler on a time-weighted mixture of autoregressive models to infer component-specific order and perform Bayesian model averaging.

O'Hara et al. (2009) provide a review of Bayesian variable selection methods in the regression setting, including that of Kuo and Mallick (1998), which we adopt here. There is also a growing literature for variable selection in BNP regression modeling. Barcella et al. (2017) provide a review that discusses approaches for covariate-dependent DPM, DDP, and product partition models. Most approaches involve binary indicator variables associated with each covariate that either activate lag-specific kernels (as in Reich et al., 2012) or break mixtures for key parameters (i.e., regression coefficients) involving point masses at 0 (as in Chung and Dunson, 2009). Another option with DPM models is to include model order as a mixing parameter (as in Lau and So, 2008).

We propose a model extension for global lag selection in Section 4.1. Section 4.2 discusses inference, including posterior sampling and other modifications to MCMC, and sampling for functionals. Section 4.3 describes an analogous extension for local lag selection. In Section

4.4, we revisit the data illustrations from Section 3 and include two additional data sets.

4.1 Model extension for global lag selection

In model (4), both mixture kernels and weights depend on the lags, thus necessitating coordination across multiple parameters for model-based lag selection. To this end, we employ binary variables $\{\gamma_\ell\}$, for $\ell = 1, \ldots, L$, to indicate dependence of y_t on $y_{t-\ell}$ if $\gamma_\ell = 1$. The most straightforward approach to incorporating these indicators follows Kuo and Mallick (1998), wherein we replace β_ℓ^y with $\gamma_\ell \beta_\ell^y$. The modification to $\boldsymbol{\beta}^y$ controls lag dependence in the mixture kernels. Our proposed modification to the weight kernels $N_{(h)}(\boldsymbol{y}_{t-1} \mid \boldsymbol{\mu}^x, \boldsymbol{\Sigma}^x)$ totally eliminates dependence on deselected lags, and is most clearly understood in the context of the sequential construction of weight kernels given in (5). We replace $\beta_{\ell,r}^x$ with $\gamma_\ell \gamma_r \beta_{\ell,r}^x$. If $\gamma_\ell = 0$, the univariate Gaussian density associated with $y_{t-\ell}$ is replaced with 1. This is equivalent to appropriately subsetting $\{\boldsymbol{\beta}_\ell^x\}$ and $\boldsymbol{\delta}^x$ prior to constructing the covariance matrix $\boldsymbol{\Sigma}^x$, reducing the dimensionality of $N_{(h)}(\boldsymbol{y}_{t-1} \mid \boldsymbol{\mu}^x, \boldsymbol{\Sigma}^x)$ to $n_\gamma = \sum_{\ell=1}^L \gamma_\ell$. If $n_\gamma = 0$, then the weight function reduces exclusively to $\boldsymbol{\omega}$, resulting in a standard univariate Gaussian DPM model. This approach reduces computational burden and offers a clean, complete lag selection, conditional on $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_L)$.

The modification for lag selection affects the hierarchical model in (6) through: the regression mean in the mixture kernel distribution for y_t, which becomes $\mu^y - \sum_{\ell=1}^L \gamma_\ell \beta_\ell^y (y_{t-\ell} - \mu_\ell^x)$; the construction of $N_{(h)}(\boldsymbol{y}_{t-1})$ in the discrete distribution for s_t; and addition of a prior for $\{\gamma_\ell\}$. We again favor simplicity and assign independent Bernoulli(π_ℓ^γ) priors to each γ_ℓ. As a default, we use a geometrically decreasing sequence $\pi_\ell^\gamma = 0.1 + (0.4/0.5) * 0.5^\ell$, for $\ell = 1, \ldots, L$, to promote sparsity and dimension reduction.

4.2 Posterior inference

The proposed setup is minimally disruptive to the MCMC algorithm outlined in Section 2.3. Conditional on $\boldsymbol{\gamma}$, the effect of selection on the mixture kernels, and hence most of the Gibbs updates, is straightforward. We update $\boldsymbol{\gamma}$ as a block, with a Metropolis step that proposes switching a random subset of $\boldsymbol{\gamma}$ (similar to Section 3.3 of Schäfer and Chopin, 2013). Details are given in Supplement S4.1.

It is well known that variable selection methods of this type tend to result in slowly mixing MCMC algorithms (O'Hara et al., 2009). Proposed changes in $\boldsymbol{\gamma}$ are often incongruous with current-state values of model parameters, which are shared across selection configurations. Furthermore, when $\gamma_\ell = 0$, draws for the associated parameters revert to their prior distributions, which may be diffuse relative to their posterior distributions when $\gamma_\ell = 1$, producing draws that will discourage returning to $\gamma_\ell = 1$. Alternative methods such as Gibbs variable selection (Dellaportas et al., 2002) adapt the prior to improve mixing, but require tuning. We do not pursue this here, but note that despite mixing difficulties and attenuated posterior probabilities for alternate lag configurations, our experience has been that MCMC chains can provide useful inferences. We recommend running multiple MCMC chains, initialized at different selection configurations. We begin MCMC with a phase in

which $\boldsymbol{\gamma}$ is not updated, followed by iterated tuning or adaptation and burn-in phases with the full sampler, followed by a final burn-in.

Posterior inferences for relevant lags from MCMC samples are trivial, requiring only samples of $\boldsymbol{\gamma}$, which can be aggregated across iterations to obtain a posterior probability of inclusion for each lag. The full expression for the transition density, marginalizing over all 2^L possible lag configurations, is

$$\tilde{f}_{Y|X}(y \mid \boldsymbol{x}) = \sum_{\boldsymbol{\gamma} \in \{0,1\}^L} \sum_{h=1}^H \tilde{q}_h(\boldsymbol{x} \mid \boldsymbol{\gamma}) \, \mathrm{N}_{(h)}(y \mid \mu(\boldsymbol{x} \mid \boldsymbol{\gamma}), \sigma^2) \, \mathrm{Pr}(\boldsymbol{\gamma}), \qquad (8)$$

where $\mathrm{Pr}(\boldsymbol{\gamma})$ can refer to either the prior or marginal posterior of $\boldsymbol{\gamma}$. In practice, we bypass the burdensome outer summation in (8) and instead calculate the lag-conditional version of the transition density in Section 2.4 across MCMC samples, which yields the desired posterior inferences marginalized with respect to the posterior of all model parameters.

Conditional on lag selection, posterior inference for functionals proceeds as in Section 2.4, with appropriate modifications to include $\boldsymbol{\gamma}$ (see Supplement S4.2 for details). Calculation of transition density and mean estimates requires the full $\boldsymbol{x} \in \mathbb{R}^L$, regardless of inferences for $\boldsymbol{\gamma}$. However, one may be interested in inferences conditional on a certain lag configuration, or marginal inferences that in some way ignore or average over the effect of a subset of $\boldsymbol{x}$. Suppose one has fit a model with $L = 3$ and desires to examine the transition mean function of the first two lags only when $\boldsymbol{\gamma} = (1, 1, 0)$. One option would be to use only posterior samples for which this lag configuration was active (taking into account the order of full-conditional sampling), given a sufficiently long MCMC chain. They may then calculate the $\boldsymbol{\gamma}$-modified transition density using these samples for any $\boldsymbol{x}$, substituting a dummy or default value in for x_3, and examining the transition density or mean as a function of x_1 and x_2 only. If fewer than all posterior samples coincide with a particular configuration, one may proceed in the same way, substituting default (or average) values in for elements of $\boldsymbol{x}$ hypothesized to be inactive and examining inferences (calculated from posterior samples, including $\boldsymbol{\gamma}$) as a function of the subset of interest. We caution that using a subset of samples ignores posterior uncertainty, and that one should test the resulting inferences for sensitivity to the default values used for inactive x_ℓ before making conclusions. For example, one could change the default values in $\boldsymbol{x}$, or replace them with random values drawn uniformly across the range of $\{y_t\}$, as demonstrated in Section 3.

4.3   Local lag selection

Thus far, we have used a single set of global indicators, $\{\gamma_\ell\}$. If one believes that lag (variable) dependence varies across the predictor space $\mathbb{R}^L$, it is straightforward to instead use a separate set $\{\gamma_\ell^{(h)}\}$ for each mixture component $h = 1, \ldots, H$, in which case the indicators become part of $\boldsymbol{\eta}_h$. Model extensions and implementation for local lag selection require only slight modifications to the procedures in Sections 4.1 and 4.2. Mixture kernels $\mathrm{N}_{(h)}(y_t \mid \boldsymbol{y}_{t-1})$ and weight kernels $\mathrm{N}_{(h)}(\boldsymbol{y}_{t-1})$ are modified in the same manner as before, but use a unique $\boldsymbol{\gamma}_h = \left(\gamma_1^{(h)}, \ldots, \gamma_L^{(h)} \right)$ for each $h = 1, \ldots, H$. With replicates of each γ_ℓ

across components, the L independent Bernoulli priors become part of G_0, and we assign independent mixture priors for each π_ℓ^γ. Following Chung and Dunson (2009) and Lucas et al. (2006), we use $\pi_\ell^\gamma \sim (1 - \pi_\ell^\pi)\,\delta_0(\pi_\ell^\gamma) + \pi_\ell^\pi\,\text{Beta}(\pi_\ell^\gamma \mid a_\ell^\pi, b_\ell^\pi)$, for $\ell = 1, \ldots, L$, where $\pi_\ell^\pi \in (0, 1)$, and δ_0 is the Dirac delta measure centered at 0. We use $\pi_\ell^\pi = 0.1 + (0.4/0.5) * 0.5^\ell$, for $\ell = 1, \ldots, L$, $a_1^\pi = \cdots = a_L^\pi = 1.0$, and $b_1^\pi = \cdots = b_L^\pi = 0.5$ as default values.

All modifications to MCMC updates in Section 4.2 still apply, but require mixture weight and kernel calculations to reference their respective $\boldsymbol{\gamma}_h$; see Supplement S4.3 for details. Inference for global dependence can be assessed with local selection, but is more nuanced. We assess global lag dependence by monitoring the weight $\sum_h \gamma_\ell^{(h)} \sum_t 1_{(s_t=h)}/(T-L)$, which gives the proportion of observations in the time series belonging to mixture components for which lag ℓ is active. Alternatively, we can replace $\gamma_\ell^{(h)}$ in the preceding expression with $\gamma_\ell^{(h)} 1_{(|\beta_\ell^{y\,(h)}|>b_0)}$, for some small threshold $b_0 > 0$, requiring both dependence in the weights *and* a minimum contribution to the slope of the kernel for a lag to be considered active. The quantities $\sum_h \omega_h \gamma_\ell^{(h)}$ and π_ℓ^γ are also informative. Inferences for transition densities and associated functionals again follow the procedures in Section 4.2.

4.4 Data illustrations incorporating lag selection

We now revisit the analyses from Section 3 with lag selection, and include two additional examples. All models in this section utilize diagonal $\boldsymbol{\Sigma}^x$ and default prior settings. Parameters of the components of G_0 associated with $\boldsymbol{\eta}^y$ were fixed. For each example, four MCMC chains were randomly initialized using the strategy described in Section 2.3. Two chains were initialized with all lags off and two were initialized with all lags on. Tuning stages were followed by 300,000 burn-in samples. The next 500,000 iterations were then thinned to 5,000 for inference (and further thinned for computationally expensive functionals such as surfaces). Both global and local lag selection were employed and compared.

Simulated data: linear autoregression

To test the model's ability to identify simple structure, for which the proposed model is over-specified, we generated time series from a stationary, Gaussian linear autoregressive model of order two. Models with global and local lag selection perform well on time series of varying length, successfully recovering parameter values and decisively selecting the first two lags (with non-negligible inclusion of lag 3 for the longer series). Further details are given in Supplement S4.4.

4.4.1 Simulated data: Ricker model

Model runs ($T = 75$, $L = 5$, $H = 40$) fit to the nonlinear simulation from Section 3.1 consistently recover lag dependence as well as the nonlinear dynamics with both global and local selection. Specifically, lag 2 is consistently kept on for all occupied mixture components throughout the chains, and other lags are generally off, with greater mixing in the model with local selection. Inferences appear fairly robust to choice of prior signal-to-noise ratio $\mathcal{R} \in \{5.0, 8.0\}$. The estimated transition mean functional, on lag 2 only, is visually very

similar to the left panel of Figure 1. A few runs include lag 4, which is reasonable given that the data reside in two diagonal quadrants of the (y_{t-2}, y_{t-4}) lag embedding space.

4.4.2 Old Faithful data

Model runs ($T = 294$, $L = 5$, $H = 40$) fit to the Old Faithful time series have mixed results. Global lag selection runs with low prior signal-to-noise ratio $\mathcal{R} = 5.0$ all converge to lag 1 only with no mixing over other configurations. Runs with $\mathcal{R} = 8.0$ continue exploring selection configurations past the specified burn-in phase, on very long timescales. One run retains lag 3 and another uses lag 2 for part of the chain. Local selection yields similar results to global, but with exploration of lag inclusion on shorter time scales and some local inclusion of lags 2 and 3. Nevertheless, these runs do not detect the density dependence noted in Section 3.2. Additional runs with priors more favorable to higher lags and larger prior weight-kernel variance also miss dependence on lag 2.

4.4.3 Pink salmon data

We next investigate a time series of annual pink salmon abundance (escapement) in Alaska, U.S.A. from 1934 to 1963 (Alaska Fisheries Science Center, 2018). Population dynamics for pink salmon provide an opportunity to test selection in the proposed model since pink salmon have a strict two-year life cycle (Heard, 1991). Thus, we expect even lags to have the most influence in predicting the current year's population. The trace of the natural logarithm of abundance in Figure 5 suggests a comprehensive analysis might appropriately include non-stationarity with long-term trends, which we forego in favor of a simple demonstration. Repeated interventions for the struggling even-year salmon population throughout the 1950s culminated in a population transfer in 1964 that bisects the complete time series and restricts us to the first segment (Bradshaw and Heintz, 2003). Nevertheless, lag scatter plots (not shown) suggest that we should be able to detect lag dependence structure, even with as few as 30 observations.

Model runs ($T = 30$, $L = 5$, $H = 25$) fit to the pink salmon data demonstrate sensitivity to prior and model specification. Most runs with global lag selection and higher prior signal-to-noise ratio ($\mathcal{R} = 8.0$) deselect all lags, although one run has lag 2 active for many

Figure 5: Trace of the natural logarithm of pink salmon abundance in Alaska from 1934 to 1963.
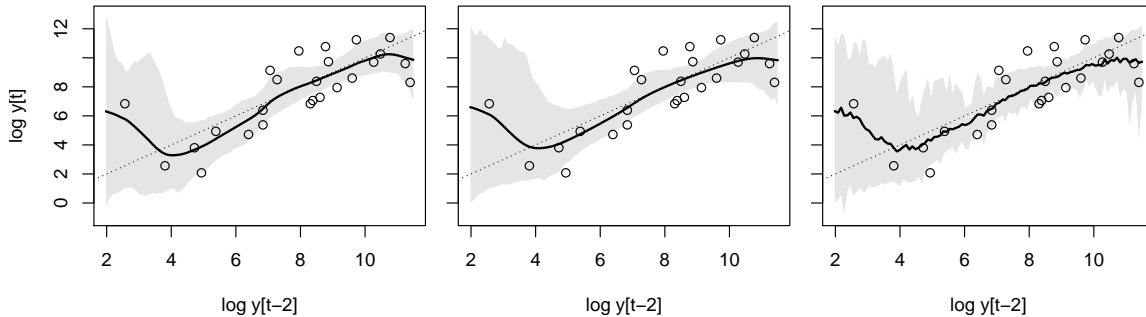
Figure 6: Model fit to the logarithm of annual pink salmon abundance with $T = 30$, $L = 5$, using global (left) and local (center and right) lag selection. In the global case, lag 2 is predominately selected, and this plot reflects only iterations of MCMC for which lag 2 is on. In the local case, other lags were fixed at the mean (center) or randomly drawn (right). The plots include pointwise posterior mean estimates and 95% credible intervals for the transition mean as a function of lag 2, together with observed two-step transitions. The dotted reference line has unit slope and passes through the origin.

inference samples. Runs with lower $\mathcal{R} = 5.0$ deselect all lags except lag 2, which is on for long periods in three of four chains. Local lag selection consistently retains lag 2 throughout most inference samples, as well as lag 4 occasionally. Increasing $\mathcal{R}$ tends to result in a higher inclusion probability for lag 4, presumably from a tendency to over-fit a transition surface informed by data only in diagonal quadrants of the $(\log(y_{t-2}), \log(y_{t-4}))$ space, similar to the Ricker model above. Figure 6 reports posterior inferences for the transition mean as a function of lag 2, under the global and local lag selection model versions. Inferences for the transition mean as a function of lag 2 only appears appropriate and insensitive to other lags.

5   Transition density estimation performance

Transition density estimation is a primary objective of the methodology. To compare density estimation across model configurations and data scenarios, we fit the model to simulated time series exhibiting various features and evaluate Monte Carlo estimates of the Kullback-Leibler (K-L) divergence between the estimated and true transition densities.

The simulated time series are variants of the Ricker-type system in (7). The first modification replaces the additive Gaussian error with multiplicative log-normal error. Specifically, transitions were generated from

$$y_t = y_{t-2} \exp(2.6 - y_{t-2} + \epsilon_t), \qquad \epsilon_t \overset{\text{iid}}{\sim} \text{N}(0, (0.09)^2), \tag{9}$$

corresponding to a log-normal transition density. This produces right skew and heteroscedasticity in the transition distribution, which continues to depend exclusively on the second lag. The lag scatter plot in Figure 7 depicts 250 transitions. We refer to this modification as the single-lag, log-normal simulation. The second modification adds dependence on the first lag through the log-scale, which is equal to $0.09\,y_{t-1}$. Thus the transition distribution is still
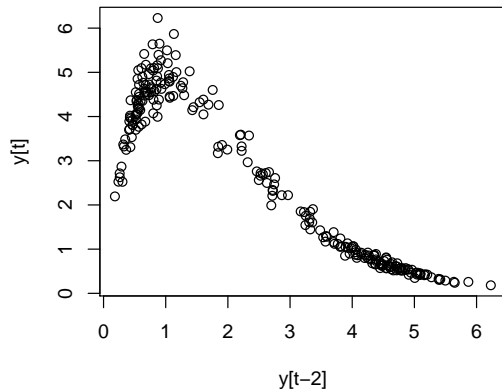
Figure 7: Lag scatter plot from the modified single-lag nonlinear simulation with log-normal transition density.
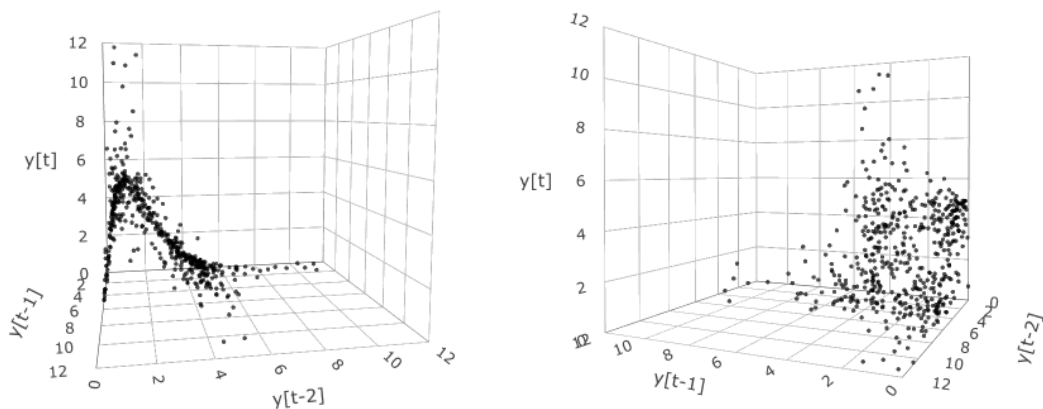

Figure 8: Lag scatter plot from the modified two-lag nonlinear simulation with log-normal transition density.

log-normal, with each parameter depending on a separate lag. The lag scatter plot in Figure 8 depicts 500 transitions, demonstrating dependence of the variance on both lags. We refer to this modification as the two-lag, log-normal simulation. In all simulations, a sequence of 1,000 observations was reserved for model fitting, and a validation set of size 1,000 was randomly sampled from the subsequent 9,000 observations. In similar data scenarios with right skew and positive-valued variables, we have previously modeled observations on the logarithmic scale. We nevertheless proceed by fitting these series directly in order to study and compare how the proposed models handle heteroscedasticity, subtle departures from Gaussianity, and subtle variation in lag dependence.

The following models were fit using default settings to all three series: the proposed model (which we denote as the BNP-WMAR model, for Bayesian nonparametric, weighted mixture of autoregressive models) with $L = 2$, full $\mathbf{\Sigma}^x$, and no lag selection; the BNP-WMAR model with $L = 5$, diagonal $\mathbf{\Sigma}^x$, and global lag selection; and finally with $L = 5$, diagonal $\mathbf{\Sigma}^x$, and local lag selection. Three chains were run for the base model that does not incorporate

20

Simulation	Model	$\hat{\mathrm{D}}_{\mathrm{KL}}$			
		70 obs.		300 obs.	
Single-lag, normal	Base model (2)	2.756	3.761	0.273	0.384
	Global selection (5)	**0.777**	0.821	**0.239**	0.252
	Local selection (5)	0.792	0.828	0.250	0.264
Single-lag, log-normal	Base model (2)	1.110	1.240	0.337	0.340
	Global selection (5)	**0.700**	0.733	0.296	0.326
	Local selection (5)	0.723	0.776	**0.296**	0.305
Two-lag, log-normal	Base model (2)	2.210	2.672	1.084	1.103
	Global selection (5)	1.429	2.096	0.966	2.002
	Local selection (5)	**1.417**	1.445	**0.948**	0.978

Table 1: Comparison of transition density estimation performance, measured by K-L divergence, from model fits to three simulations and with two sample sizes. The numbers in parentheses are L, the number of lags considered in each fit. Within each set, the minimum (left) and maximum (right) losses across runs are reported.

lag selection, and four chains were run for each model with lag selection, with two chains initialized with all lags off and the other two initialized with all lags on.

Each posterior sample was used to create density ordinates, denoted $\hat{p}(y_t \mid \boldsymbol{y}_{t-1})$ and calculated from the transition density in Section 2.4, appropriately modified by lag selection indicators. With 2,000 replicate simulation draws $\{y_j^{(i)}\}_{i=1}^{2,000}$ from the data-generating distribution (with density $p_{\mathrm{true}}(y_t \mid \boldsymbol{y}_{t-1})$) for each validation pair $\{(y_j, \boldsymbol{x}_j)\}_{j=1}^{1,000}$, we approximated the Kullback-Leibler divergence using

$$
\begin{aligned}
\mathrm{D}_{\mathrm{KL}}(p_{\mathrm{true}} \parallel \hat{p}) &\equiv \int p_{\mathrm{true}}(y \mid \boldsymbol{x}) \log \left(\frac{p_{\mathrm{true}}(y \mid \boldsymbol{x})}{\hat{p}(y \mid \boldsymbol{x})} \right) \mathrm{d}y \\
&\approx \sum_{i=1}^{2,000} \log \left(p_{\mathrm{true}}(y^{(i)} \mid \boldsymbol{x}) \right) - \log \left(\hat{p}(y^{(i)} \mid \boldsymbol{x}) \right),
\end{aligned}
\tag{10}
$$

averaged over validation observations and posterior simulations. Let $\hat{\mathrm{D}}_{\mathrm{KL}}$ denote the result. This loss metric is reported in Table 1 for two chains of each model fit to time-series of lengths $T = 75$ and $T = 305$ ($T = 72$ and 302 for the model with $L = 2$; using the same 70 and 300 observations used to fit the models with $L = 5$). The two reported runs are those producing the minimum and maximum observed K-L divergence within each set.

In the single-lag, normal scenario, the burden of fitting an unnecessary dimension of the phase space is evident, particularly with the short time series. In the $T = 302$ case with $L = 2$, two of the chains use three mixture components, whereas one uses four and performs comparably to the models with lag selection. Both global and local lag selection perform well for both sample sizes, and yield accurate inferences for lag dependence, with occasional inclusion of lag 4.

Fitting two lags again hinders the base model in the single-lag, log-normal scenario when

sample size is small. Both global and local lag selection perform well, with the best run of local selection only slightly outperforming the best run of global selection in the $T = 305$ case. In the large sample, chains of the global selection model initialized with all lags on retain both lags 2 and 4, whereas chains initialized with all lags off retain only lag 2. Despite this lack of mixing, the discrepancy in K-L loss is minimal.

In the two-lag, log-normal scenario, the base model with $L = 2$ has the advantage of being fixed at the correct lag structure. However, both models with lag selection manage superior performance. In the small sample, the base model over-fits a few points in the sparse region with y_{t-1} low and y_{t-2} high, producing inferences that fail to generalize. In contrast, the global selection model retains only lag 2 in three of four runs (selecting none in the poorly performing run), avoiding the over-fitting issue at the expense of missing density dependence on the first lag. The model compensates with a right-skewed transition density when y_{t-2} is low, irrespective of y_{t-1}. Local lag selection has similar behavior in the small sample. All models struggle in the region with small values of lag 1 and large values of lag 2, where the true density is far more concentrated than estimated.

In the large sample, global selection is inconsistent, correctly retaining both of the first two lags when initialized with all lags on, but retaining no lags and lag 2 only in respective runs initialized with no lags on. The two runs with correct lag selection yield effective density estimation, capturing variance and skew dependence on y_{t-1}. Local lag selection does the same, with consistent performance across initializations.

The dimension reduction and parsimony afforded by lag selection provide significant gains in density estimation, as measured by K-L distance, and make a strong case for the proposed model extensions. Global selection can be effective when the dependence structure is simple, but we generally recommend the more versatile local selection model.

6   Discussion

We have developed a modeling framework for fully nonparametric, nonlinear autoregressive models targeted at estimating transition densities. The model extends existing single-lag counterparts and further offers inference for lag dependence. We have demonstrated the model's utility with simulated, geological, and ecological data examples with diverse objectives. The model allows users to relax restrictive characteristics of standard models, or softly specify such through prior settings, within a single model.

Results from the base model are promising, faithfully capturing known or anticipated features in the data examples. Of course, current computing bottlenecks limit what can practically be accomplished. For example, complex dynamics call for many mixture components and high truncation level H, and computations for updating component-specific parameters are not readily distributable in our approach due to their appearance in the normalized weights.

The modeling objectives of estimating flexible transition densities, accommodating nonlinear dynamics, and selecting active lags offer many degrees of freedom that in most cases will not be entirely identified with data alone. Decisions must be made, and correspondent

behaviors encouraged through the prior settings. As such, we recommend completing a thorough exploratory analysis of data. We further advise that practitioners fit models with a variety of prior signal-to-noise ratio and flexibility (through α and possibly δ^x) settings, each with multiple MCMC chains.

Several considerations can help guide which settings are appropriate for a given scenario. One that bears on lag selection is the interplay between noise and signal. A model attempting to fit noise may erroneously reach into higher dimensions. However, in the absence of noise, finding a high-dimensional structure is an objective of techniques such as time-delay embedding. Another consideration arises from correlation among lags, which can result in multiple distinct lag configurations that each produce comparably effective forecasts. This partially motivates our recommendation of decaying inclusion probabilities in the prior.

Inference for relevant lags remains practically challenging. Our experience has been that results from the models with lag selection tend to exhibit prior sensitivity, a natural consequence of the flexibility discussed. We have also noted that in models with lag selection, mixing challenges intensify with increased time series length, which tends to sharpen posterior modes. This often manifests through kernel coefficients being estimated at small, nonzero magnitudes while corresponding lag selection indicators remain on. Local selection helps alleviate this issue by breaking up the samples informing multiple lag indicators, naturally tempering the posterior distributions and encouraging greater mixing. The cost of added versatility and improved mixing is a more intricate picture of lag importance in posterior analysis.

Although binary lag inclusion parameters are easy to interpret, they offer limited insight to relative contributions from active lags. Such contributions can be quantified for the mean transition function through functional decomposition, but this is less straightforward for transition densities. Ideally, weak dependence would manifest in the posterior probability of inclusion. Alternatively, we envision a framework that quantifies lag importance with shrinkage of continuous parameters. Continuous quantification of lag importance could in turn reduce the influence of weight kernels relative to the DP weights and thus accommodate multiple sources of influence on the mixture weights.

Notwithstanding theoretical and practical challenges, lag selection is critical for dimension reduction and is an integral part of this work. Simpler models can partially avoid some of the challenges noted, but risk failing to model, or even detect, nonlinear and/or non-Gaussian dynamics. Our proposed methods extend Bayesian nonparametric density autoregressive modeling by accommodating multiple lags and providing a framework for lag selection that works in concert with the other objectives.

Bibliography

Alaska Fisheries Science Center (2018), "AFSC/ABL: Pink salmon data collected at Sashin Creek Weir 1934-2002," URL `https://inport.nmfs.noaa.gov/inport/item/17256`.

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

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. and Bowman, A. W. (1990), "A Look at Some Data on the Old Faithful Geyser," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 39, 357–365.

Barcella, W., De Iorio, M., and Baio, G. (2017), "A comparative review of variable selection techniques for covariate dependent Dirichlet process mixture models," *Canadian Journal of Statistics*, 45, 254–273.

Barrientos, A. F., Jara, A., and Quintana, F. A. (2017), "Fully Nonparametric Regression for Bounded Data Using Dependent Bernstein Polynomials," *Journal of the American Statistical Association*, 112, 806–825.

Beal, M. J., Ghahramani, Z., and Rasmussen, C. E. (2002), "The Infinite Hidden Markov Model," in Dietterich, T. G., Becker, S., and Ghahramani, Z. (editors), *Advances in Neural Information Processing Systems*, volume 14, MIT Press.

Bradshaw, R. and Heintz, R. (2003), "SCWDATA: Pink Salmon data collected at Sashin Creek Weir 1934-2002," Database Report. Auke Bay Laboratory, Alaska Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration.

Cai, B. and Dunson, D. B. (2006), "Bayesian Covariance Selection in Generalized Linear Mixed Models," *Biometrics*, 62, 446–457.

Caron, F., Davy, M., Doucet, A., Duflos, E., and Vanheeghe, P. (2007), "Bayesian Inference for Linear Dynamic Models with Dirichlet Process Mixtures," *IEEE Transactions on Signal Processing*, 56, 71–84.

Carvalho, A. X. and Tanner, M. A. (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 of time series models," *International Journal of Mathematics and Mathematical Sciences*, 2006, 1–22.

Chae, M. and Walker, S. G. (2019), "Bayesian consistency for a nonparametric stationary Markov model," *Bernoulli*, 25, 877–901.

Chung, Y. and Dunson, D. B. (2009), "Nonparametric Bayes Conditional Distribution Modeling With Variable Selection," *Journal of the American Statistical Association*, 104, 1646–1660.

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

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2002), "On Bayesian model and variable selection using MCMC," *Statistics and Computing*, 12, 27–36.

DeYoreo, M. and Kottas, A. (2017), "A Bayesian nonparametric Markovian model for non-stationary time series," *Statistics and Computing*, 27, 1525–1538.

— (2020), "Bayesian nonparametric density regression for ordinal responses," in Fan, Y., Nott, D., Smith, M. S., and Dortet-Bernadet, J.-L. (editors), *Flexible Bayesian Regression Modelling*, Elsevier, 65–90.

Di Lucca, M. A., Guglielmi, A., Müller, P., and Quintana, F. A. (2013), "A Simple Class of Bayesian Nonparametric Autoregression Models," *Bayesian analysis*, 8, 63–88.

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

Fox, E., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. (2011), "Bayesian Nonparametric Inference of Switching Dynamic Linear Models," *IEEE Transactions on Signal Processing*, 59, 1569–1585.

Frühwirth-Schnatter, S. (2006), *Finite Mixture and Markov Switching Models*, Springer-Verlag New York.

Fuentes-García, R., Mena, R. H., and Walker, S. G. (2009), "A nonparametric dependent process for Bayesian regression," *Statistics & Probability Letters*, 79, 1112–1119.

Glasbey, C. (2001), "Non-linear autoregressive time series with multivariate Gaussian mixtures as marginal distributions," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50, 143–154.

Heard, W. R. (1991), "Life History of Pink Salmon (Oncorhynchus gorbuscha)," in Groot, C. and Margolis, L. (editors), *Pacific Salmon Life Histories*, UBC Press, 119–230.

Huerta, G. and West, M. (1999), "Priors and component structures in autoregressive time series models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61, 881–899.

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

Jordan, M. I. and Jacobs, R. A. (1994), "Hierarchical Mixtures of Experts and the EM Algorithm," *Neural computation*, 6, 181–214.

Kalli, M. and Griffin, J. E. (2018), "Bayesian nonparametric vector autoregressive models," *Journal of Econometrics*, 203, 267–282.

Kalliovirta, L., Meitz, M., and Saikkonen, P. (2015), "A Gaussian Mixture Autoregressive Model for Univariate Time Series," *Journal of Time Series Analysis*, 36, 247–266.

Khalili, A., Chen, J., and Stephens, D. A. (2017), "Regularization and selection in Gaussian mixture of autoregressive models," *Canadian Journal of Statistics*, 45, 356–374.

Kuo, L. and Mallick, B. (1998), "Variable Selection for Regression Models," *Sankhyā: The Indian Journal of Statistics, Series B*, 60, 65–81.

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

Liu, J. S. (1996), "Peskun's theorem and a modified discrete-state Gibbs sampler." *Biometrika*, 83.

Lucas, J., Carvalho, C., Wang, Q., Bild, A., Nevins, J. R., and West, M. (2006), "Sparse statistical modelling in gene expression genomics," *Bayesian inference for gene expression and proteomics*, 1, 155–176.

MacEachern, S. N. (2000), "Dependent Dirichlet Processes," Unpublished manuscript, Department of Statistics, The Ohio State University.

Martınez-Ovando, J. C. and Walker, S. G. (2011), "Time-series Modelling, Stationarity and Bayesian Nonparametric Methods," Technical report, Banco de México.

Mena, R. H. and Walker, S. G. (2005), "Stationary Autoregressive Models via a Bayesian Nonparametric Approach," *Journal of Time Series Analysis*, 26, 789–805.

Müller, P., Erkanli, A., and West, M. (1996), "Bayesian curve fitting using multivariate normal mixtures," *Biometrika*, 83, 67–79.

Müller, P., West, M., and MacEachern, S. (1997), "Bayesian Models for Non-linear Autoregressions," *Journal of Time Series Analysis*, 18, 593–614.

Neal, R. M. (2003), "Slice sampling," *The Annals of Statistics*, 31, 705–767.

Nicholl, M. J., Wheatcraft, S. W., Tyler, S. W., and Berkowitz, B. (1994), "Is Old Faithful a strange attractor?" *Journal of Geophysical Research: Solid Earth*, 99, 4495–4503.

O'Hara, R. B., Sillanpää, M. J., et al. (2009), "A review of Bayesian variable selection methods: what, how and which," *Bayesian analysis*, 4, 85–117.

Park, J.-H. and Dunson, D. B. (2010), "Bayesian Generalized Product Partition Model," *Statistica Sinica*, 20, 1203–1226.

Peng, F., Jacobs, R. A., and Tanner, M. A. (1996), "Bayesian Inference in Mixtures-of-Experts and Hierarchical Mixtures-of-Experts Models with an Application to Speech Recognition," *Journal of the American Statistical Association*, 91, 953–960.

Pitt, M. K., Chatfield, C., and Walker, S. G. (2002), "Constructing First Order Stationary Autoregressive Models via Latent Processes," *Scandinavian Journal of Statistics*, 29, 657–663.

Prado, R. and West, M. (2010), *Time Series: Modeling, Computation, and Inference*, CRC Press.

Reich, B. J., Kalendra, E., Storlie, C. B., Bondell, H. D., and Fuentes, M. (2012), "Variable selection for high dimensional Bayesian density estimation: application to human exposure simulation," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61, 47–66.

Ricker, W. E. (1954), "Stock and Recruitment," *Journal of the Fisheries Research Board of Canada*, 11, 559–623.

Robert, C. P. and Casella, G. (2004), *Monte Carlo Statistical Methods*, Springer-Verlag New York, 2nd edition.

Rodríguez, A. and Ter Horst, E. (2008), "Bayesian dynamic density estimation," *Bayesian Analysis*, 3, 339–365.

Schäfer, C. and Chopin, N. (2013), "Sequential Monte Carlo on large binary sampling spaces," *Statistics and Computing*, 23, 163–184.

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

Shahbaba, B. and Neal, R. (2009), "Nonlinear Models Using Dirichlet Process Mixtures," *Journal of Machine Learning Research*, 10, 1829–1850.

Shumway, R. H. and Stoffer, D. S. (2017), *Time Series Analysis and Its Applications: With R Examples*, Springer International Publishing, 4th edition.

Smith, M. and Kohn, R. (2002), "Parsimonious Covariance Matrix Estimation for Longitudinal Data," *Journal of the American Statistical Association*, 97, 1141–1153.

Taddy, M. A. and Kottas, A. (2009), "Markov switching Dirichlet process mixture regression," *Bayesian Analysis*, 4, 793–816.

Tang, Y. and Ghosal, S. (2007a), "A consistent nonparametric Bayesian procedure for estimating autoregressive conditional densities," *Computational Statistics & 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. (1990), *Non-linear Time Series: A Dynamical System Approach*, Oxford: Clarendon Press.

Wade, S., Walker, S. G., and Petrone, S. (2014), "A Predictive Study of Dirichlet Process Mixture Models for Curve Fitting," *Scandinavian Journal of Statistics*, 41, 580–605.

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

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

Yau, C., Papaspiliopoulos, O., Roberts, G. O., and Holmes, C. (2011), "Bayesian non-parametric hidden Markov models with applications in genomics," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73, 37–57.

S1 Model for stationary time series

One way to ensure stationarity of a process with transition kernel (4) is to espouse the joint density interpretation of the mixture (2) and constrain $\boldsymbol{\mu}_h = \mathbf{1}\mu_h$ and $\boldsymbol{\Sigma}_h$ to be positive definite Toeplitz, thereby ensuring reversibility with identical marginal distributions ($\int_{\mathcal{A}} f_{Y,X}(y,\boldsymbol{x})\,\mathrm{d}y = \int_{\mathcal{A}} f_{Y,X}(y,\boldsymbol{x})\,\mathrm{d}x_L$ for any measurable set $\mathcal{A}$). This is the approach taken by Antoniano-Villalobos and Walker (2016) and Kalli and Griffin (2018), who assume an AR(1) covariance structure for $\boldsymbol{\Sigma}_h$. Generally, each mixture component accommodates up to $L+2$ parameters. The Markovian likelihood then arising from (4) precludes closed-form Gibbs sampling.

Model-based lag selection presents the primary challenge in a stationary model with $L > 1$. An additional requirement is that lag selection is global. To see this, consider a two-component mixture of bivariate Gaussian densities, with each component using a separate lag: $f(y_t, y_{t-1}, y_{t-2}) = \omega\,\mathrm{N}_{(1)}(y_t, y_{t-1}) + (1-\omega)\,\mathrm{N}_{(2)}(y_t, y_{t-2})$. Integrating out y_t leaves a mixture of univariate Gaussian densities, whereas integrating out y_{t-2} leaves one bivariate and one univariate component, yielding two distinct marginal densities for (y_t, y_{t-1}) and (y_{t-1}, y_{t-2}). We can maintain stationarity with a joint density defined over a set of nonconsecutive lags (e.g., $f(y_t, y_{t-2})$) if we assume that conditionally independent (in the transition) subsequences of the time series follow the same stationary distribution.

Given the structural requirement on $\boldsymbol{\Sigma}_h$, the Cholesky factorization in Section 2.1.1 is not useful in the stationary case. We instead consider constructing $\boldsymbol{\Sigma}_h$ from autoregressive coefficients, which we denote $\boldsymbol{\phi}_h = (\phi_1^{(h)}, \ldots, \phi_L^{(h)})$, allowing some control over lag dependence. Given $\boldsymbol{\phi}_h$, μ_h, and innovation variance σ_h^2, we can recover $\boldsymbol{\Sigma}_h$ by solving the set of homogeneous difference equations defining the stationary autocovariance (equivalently, the Yule-Walker equations). To maintain causality (a sufficient condition for stationarity) of

each mixture component, all roots of the AR characteristic polynomial must lie outside the unit circle (Shumway and Stoffer, 2017, pp. 86, 95, 113).

We briefly note four possible approaches to modeling lag dependence with $\boldsymbol{\phi}_h$:

1. Specify a full $\boldsymbol{\phi}_h$ vector for each mixture component and use $\{\gamma_\ell \, \phi_\ell^{(h)}\}_{\ell=1}^L$ to construct $\boldsymbol{\Sigma}_h$, as in Section 4. A drawback of this approach is that the weight kernels remain L-variate Gaussian densities, dependent on all L lags.

2. To eliminate dependence on inactive lags, use the original $\boldsymbol{\phi}_h$ to construct $\boldsymbol{\Sigma}_h$ before marginalizing both mixture and weight kernels as $\int \mathrm{N}_{(h)}(y, \boldsymbol{y}_{t-1}) \, \mathrm{d}\bar{\boldsymbol{y}}_{t-1}$ and $\int \mathrm{N}_{(h)}(\boldsymbol{y}_{t-1}) \, \mathrm{d}\bar{\boldsymbol{y}}_{t-1}$, respectively, where $\bar{\boldsymbol{y}}_{t-1}$ contains the $y_{t-\ell}$ for which $\gamma_\ell = 0$. A drawback of this approach is that the model is over-parameterized, leaving unidentified parameters and possibly inflating uncertainty.

3. Combine approaches 1 and 2, removing the effect of inactive lags when constructing the transition density, with joint densities defined over possibly nonconsecutive lags. This approach seems the most promising.

4. One could define the component-specific parameter space as a union of sets of distinct AR coefficient vectors, and perform transdimensional MCMC. Beyond the computational burden this would create with H mixture components, the combinatorial complexity of considering subsets of $\{\phi_\ell^{(h)}\}_{\ell=1}^L$ makes this option unattractive.

In all cases, it would be necessary to check the roots of characteristic polynomials prior to accepting any candidate $\boldsymbol{\phi}$ or $\boldsymbol{\gamma}$ during MCMC, and mixing could suffer from inability to marginalize over mixture kernel parameters. One way to avoid checking roots is to work directly on the space of characteristic polynomial roots, as in Huerta and West (1999), although they disallow gaps in lag dependence and utilize reversible-jump algorithms to identify model order. Nevertheless, if stationarity is required in a particular modeling scenario, it may prove worthwhile to consider these approaches.

S2   Prior settings

We can visualize the effects of prior settings through prior simulation in low-dimensional models. As an example, Figure S1 depicts several realizations of the transition mean for a model with a single lag. The realizations are drawn under combinations of prior settings for α (through the shape parameter with the scale fixed at 1.0) and δ^x (through the prior mean of s_0^x). Restricting the number of components with low values of α results in transition mean functions with few change points and long stretches of near linearity, whereas allowing more components increases variability in the curve. Low values for δ^x likewise encourage rigid transition mean curves with abrupt change points. Increasing the variance in the weight kernels has a smoothing effect, as expected. Note that in regions of the lag space where multiple mixture components carry significant weight, the transition density can be multimodal, with a transition mean that does not closely follow any one of

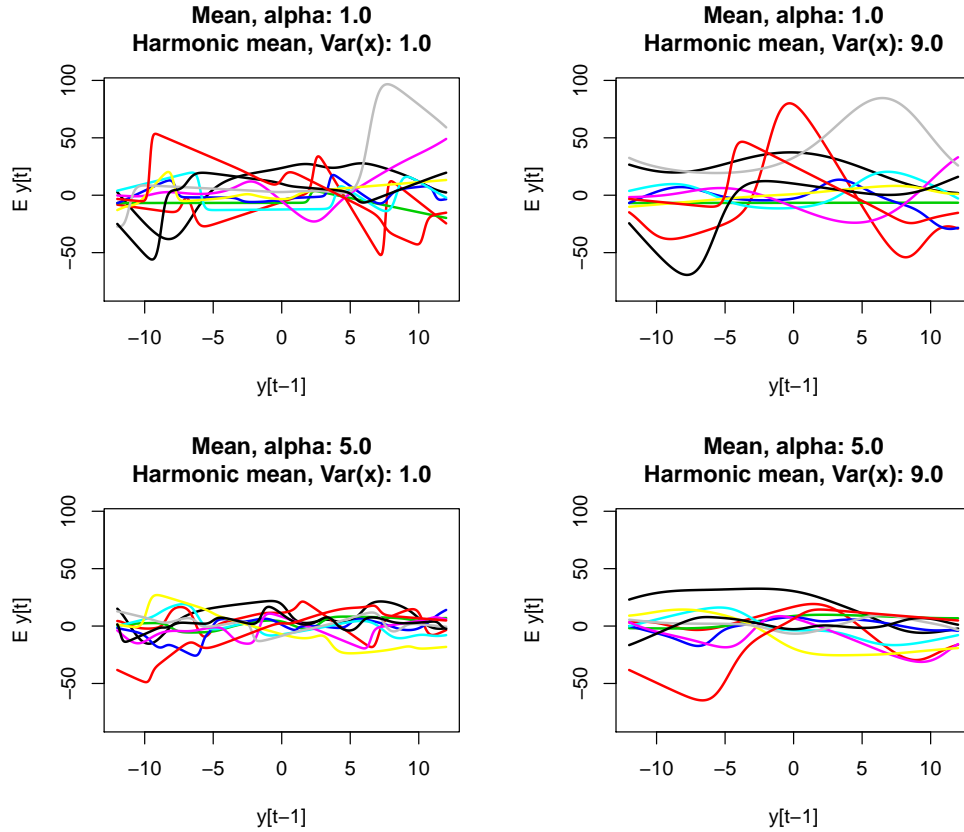Figure S1: Ten prior realizations of the transition mean for the proposed nonparametric model with a single lag, under combinations of prior settings for α and δ^x.

the component-specific lines. As with the transition mean, one can use prior simulation to elucidate the effects of prior settings for transition densities to aid practitioners in specifying desired characteristics and performing sensitivity analyses.

S3 MCMC details for the base model

If we condition on the first L observations of the time series, the hierarchical model (6) yields the full joint posterior distribution over all model parameters up to proportionality,

$$
p(\cdots \mid \{y_t\}_{t=1}^T) \propto
$$

$$
\prod_{t=L+1}^{T} \left[\frac{\omega_{s_t} \, \mathrm{N}_{(s_t)}(\boldsymbol{y}_{t-1} \mid \boldsymbol{\mu}^x, \boldsymbol{\Sigma}^x)}{\sum_{j=1}^{H} \omega_j \, \mathrm{N}_{(j)}(\boldsymbol{y}_{t-1} \mid \boldsymbol{\mu}^x, \boldsymbol{\Sigma}^x)} \, \mathrm{N}_{(s_t)} \left(y_t \mid \mu^y - \sum_{\ell=1}^{L} \beta_\ell^y (y_{t-\ell} - \mu_\ell^x), \sigma^2 \right) \right] \times
$$

$$
\prod_{h=1}^{H} \left[\mathrm{Beta}(v_h \mid 1, \alpha)^{1(h<H)} \, p(\boldsymbol{\eta}_h \mid G_0) \right] \, \mathrm{Ga}(\alpha \mid a_\alpha, b_\alpha) \times \qquad\qquad\text{(S1)}
$$

$$
\mathrm{N}(\boldsymbol{\beta}_0^*) \, \mathrm{IWish}\left((\boldsymbol{\Lambda}_0^*)^{-1}\right) \, \mathrm{Ga}(s_0) \, \mathrm{N}(\boldsymbol{\mu}_0^x) \, \mathrm{IWish}((\boldsymbol{\Lambda}^{\mu_x})^{-1}) \times
$$

$$
\prod_{r=1}^{L-1} \left[\mathrm{N}(\boldsymbol{\beta}_{0,r}^{\beta_x}) \, \mathrm{IWish}((\boldsymbol{\Lambda}_{0,r}^{\beta_x})^{-1}) \right] \, \prod_{\ell=1}^{L} \mathrm{Ga}(s_{0,\ell}^x),
$$

where

$$
p(\boldsymbol{\eta}_h \mid G_0) = \mathrm{N}\left((\mu_{(h)}^y, \boldsymbol{\beta}_{(h)}^y) \mid \boldsymbol{\beta}_0^*, \sigma_{(h)}^2 (\boldsymbol{\Lambda}_0^*)^{-1} \right) \, \mathrm{IG}\left(\sigma_{(h)}^2 \mid \frac{\nu_{\sigma^2}}{2}, \frac{\nu_{\sigma^2} s_0}{2} \right) \times
$$

$$
\mathrm{N}\left(\boldsymbol{\mu}_{(h)}^x \mid \boldsymbol{\mu}_0^x, (\boldsymbol{\Lambda}^{\mu_x})^{-1} \right) \prod_{r=1}^{L-1} \mathrm{N}\left(\boldsymbol{\beta}_{r,(h)}^x \mid \boldsymbol{\beta}_{0,r}^{\beta_x}, (\boldsymbol{\Lambda}_{0,r}^{\beta_x})^{-1} \right) \times \qquad\text{(S2)}
$$

$$
\prod_{\ell=1}^{L} \mathrm{IG}\left(\delta_{\ell,(h)}^x \mid \frac{\nu_\ell^{\delta^x}}{2}, \frac{\nu_\ell^{\delta^x} s_{0,\ell}^x}{2} \right)
$$

The Gibbs sampler proceeds by successively sampling the parameters in the sets and manner described below.

Latent states

The latent states identifying component membership for each observation y_t are updated individually, for $t = L+1, \ldots, T$, using their discrete full conditional distributions $\Pr(s_t = h \mid \cdots) \propto \omega_h \, \mathrm{N}_{(h)}(\boldsymbol{y}_{t-1} \mid \boldsymbol{\mu}^x, \boldsymbol{\Sigma}^x) \, \mathrm{N}_{(h)}\left(y_t \mid \mu^y - \sum_{\ell=1}^{L} \beta_\ell^y (y_{t-\ell} - \mu_\ell^x), \sigma^2 \right)$, for $h = 1, \ldots, H$. The step is Metropolized by first drawing a candidate with probability mass proportional to the full conditional, excluding the current state. The Metropolis acceptance ratio is then the sum over all full conditional probabilities excluding the current state, divided by the sum over all full conditional probabilities excluding the candidate state (Robert and Casella, 2004, p. 394).

Stick-breaking weights

The weights $\{\omega_h\}_{h=1}^H$ that appear in the likelihood are defined through the latent $\{v_h\}_{h=1}^{H-1}$ which, conditional on the latent states $\{s_t\}$ and absent the denominator in the first product term of (S1), admit $H-1$ independent beta full conditional distributions (Ishwaran and

James, 2001). In our model, the full conditional distributions are given as

$$p(\{v_h\} \mid \cdots) \propto \prod_{t=L+1}^{T} \left[\frac{\omega_{s_t}}{\sum_{j=1}^{H} \omega_j \, \mathrm{N}_{(j)}(\boldsymbol{y}_{t-1} \mid \boldsymbol{\mu}^x, \boldsymbol{\Sigma}^x)} \right] \prod_{h=1}^{H-1} \mathrm{Beta}(v_h \mid 1, \alpha)$$

$$\propto \frac{\prod_{h=1}^{H-1} \mathrm{Beta}(v_h \mid 1 + n_h^*, \alpha + \sum_{k=h+1}^{H} n_k^*)}{\prod_{t=L+1}^{T} \sum_{j=1}^{H} v_j \prod_{i=1}^{j-1} (1 - v_i) \, \mathrm{N}_{(j)}(\boldsymbol{y}_{t-1} \mid \boldsymbol{\mu}^x, \boldsymbol{\Sigma}^x)} ,$$

(S3)

where the $n_h^* = \sum_{t=L+1}^{T} 1_{(s_t=h)}$, for $h = 1, \ldots, H$, count membership in each of the H components. We define $v_H = 1$ for convenience in notation. To obtain direct samples from this distribution, we employ the multivariate hyperrectangle slice sampler of Neal (2003) (summarized in Figure 8 of that article) to update all v_h, $h = 1, \ldots, H - 1$, simultaneously, as follows.

Let $\boldsymbol{v} = (v_1, \ldots, v_{H-1})$ denote the vector of latent beta variables used to construct $\{\omega_h\}_{h=1}^{H}$, and let $g(\boldsymbol{v})$ denote the unnormalized posterior full conditional density (S3) evaluated at $\boldsymbol{v}$. The algorithm employs user-specified tuning parameters $\{\tau_h\}_{h=1}^{H-1}$, all of which we conservatively fix equal to 1.0 to ensure that the entire support of $\boldsymbol{v}$ (i.e., the hypercube $(0, 1)^{H-1}$) can be reached in any iteration of MCMC.

Let $\boldsymbol{v}^0$ denote the value of $\boldsymbol{v}$ from the previous iteration of MCMC, and $\boldsymbol{v}^1$ denote the output of this algorithm, which proceeds as follows (Figure 8 of Neal, 2003).

1. Define the slice.
   Draw $z \sim \mathrm{Unif}(0, g(\boldsymbol{v}_0))$.

2. Initialize the hyperrectangle.
   $\mathcal{H} = (L_1, R_1) \times \cdots \times (L_{H-1}, R_{H-1})$, where
   $L_h \leftarrow v_h^0 - \tau_h \, U_h$,
   $R_h \leftarrow L_h + \tau_h$,
   with draws $U_h \overset{\mathrm{ind.}}{\sim} \mathrm{Unif}(0, 1)$, for $h = 1, \ldots, H - 1$.

3. Propose candidates $\boldsymbol{v}^*$ and iteratively shrink $\mathcal{H}$ when points are rejected.
   Repeat the following until a candidate satisfying $z < g(\boldsymbol{v}^*)$ is found:

   (i) Draw $\tilde{U}_h \overset{\mathrm{ind.}}{\sim} \mathrm{Unif}(0, 1)$, for $h = 1, \ldots, H - 1$.
   (ii) Set candidate $v_h^* \leftarrow L_h + \tilde{U}_h \, (R_h - L_h)$, for $h = 1, \ldots, H - 1$.
   (iii) If $z < g(\boldsymbol{v}^*)$, set $\boldsymbol{v}^1 \leftarrow \boldsymbol{v}^*$ and exit the algorithm.
   (iv) If $v_h^* < v_h^0$, then set $L_h \leftarrow v_h^*$, otherwise set $R_h \leftarrow v_h^*$, for
       $h = 1, \ldots, H - 1$.

Component-specific parameters

The posterior full conditional density for each $\boldsymbol{\eta}_h$ is given by

$$p(\boldsymbol{\eta}_h \mid \cdots) \propto$$

$$\prod_{t:s_t=h} \left[\mathrm{N}_{(h)}(\boldsymbol{y}_{t-1} \mid \boldsymbol{\mu}^x, \boldsymbol{\Sigma}^x) \, \mathrm{N}_{(h)}\left(y_t \mid \mu^y - \sum_{\ell=1}^{L} \beta_\ell^y (y_{t-\ell} - \mu_\ell^x), \sigma^2 \right) \right] \times$$

$$\prod_{t=L+1}^{T} \left[\sum_{j=1}^{H} \omega_j \, \mathrm{N}_{(j)}(\boldsymbol{y}_{t-1} \mid \boldsymbol{\mu}^x, \boldsymbol{\Sigma}^x) \right]^{-1} \mathrm{N}\left((\mu_{(h)}^y, \boldsymbol{\beta}_{(h)}^y) \mid \boldsymbol{\beta}_0^*, \sigma_{(h)}^2 (\boldsymbol{\Lambda}_0^*)^{-1} \right) \times \qquad \text{(S4)}$$

$$\mathrm{IG}\left(\sigma_{(h)}^2 \mid \frac{\nu_{\sigma^2}}{2}, \frac{\nu_{\sigma^2} s_0}{2} \right) \mathrm{N}\left(\boldsymbol{\mu}_{(h)}^x \mid \boldsymbol{\mu}_0^x, (\boldsymbol{\Lambda}^{\mu_x})^{-1} \right) \times$$

$$\prod_{r=1}^{L-1} \mathrm{N}\left(\boldsymbol{\beta}_{r,(h)}^x \mid \boldsymbol{\beta}_{0,r}^{\beta_x}, (\boldsymbol{\Lambda}_{0,r}^{\beta_x})^{-1} \right) \prod_{\ell=1}^{L} \mathrm{IG}\left(\delta_{\ell,(h)}^x \mid \frac{\nu_\ell^{\delta^x}}{2}, \frac{\nu_\ell^{\delta^x} s_{0,\ell}^x}{2} \right),$$

for $h = 1, \ldots, H$. To improve mixing of the y-indexed, component-specific parameters, we partition $\boldsymbol{\eta}$ into its y and x components $\boldsymbol{\eta}^y \equiv \{\mu^y, \boldsymbol{\beta}^y, \sigma^2\}$ and $\boldsymbol{\eta}^x \equiv \{\boldsymbol{\mu}^x, \boldsymbol{\beta}_1^x, \ldots, \beta_{L-1}^x, \boldsymbol{\delta}^x\}$, and sample $p(\boldsymbol{\eta}_h \mid \cdots) = p(\boldsymbol{\eta}_h^x \mid \cdots, -\boldsymbol{\eta}_h^y) \, p(\boldsymbol{\eta}_h^y \mid \boldsymbol{\eta}_h^x, \cdots)$, where $p(\boldsymbol{\eta}_h^x \mid \cdots, -\boldsymbol{\eta}_h^y) = \int p(\boldsymbol{\eta}_h \mid \cdots) \, \mathrm{d}\boldsymbol{\eta}_h^y$. This sequential sampling scheme adds little to algorithmic complexity, as the full conditional density $p(\boldsymbol{\eta}_h^x \mid \cdots)$ already contains the mixture-weight denominator $\prod_t \sum_j \omega_j \, \mathrm{N}_{(j)}(\boldsymbol{y}_{t-1})$, precluding simple conjugate updates.

Integrating $\boldsymbol{\eta}_h^y$ from the full conditional for $\boldsymbol{\eta}_h$ yields

$$p(\boldsymbol{\eta}_h^x \mid \cdots, -\boldsymbol{\eta}_h^y) \propto \prod_{t:s_t=h} \mathrm{N}_{(h)}(\boldsymbol{y}_{t-1} \mid \boldsymbol{\mu}^x, \boldsymbol{\Sigma}^x) \prod_{t=L+1}^{T} \left[\sum_{j=1}^{H} \omega_j \, \mathrm{N}_{(j)}(\boldsymbol{y}_{t-1} \mid \boldsymbol{\mu}^x, \boldsymbol{\Sigma}^x) \right]^{-1} \times$$

$$\mathrm{N}\left(\boldsymbol{\mu}_{(h)}^x \mid \boldsymbol{\mu}_0^x, (\boldsymbol{\Lambda}^{\mu_x})^{-1} \right) \prod_{r=1}^{L-1} \mathrm{N}\left(\boldsymbol{\beta}_{r,(h)}^x \mid \boldsymbol{\beta}_{0,r}^{\beta_x}, (\boldsymbol{\Lambda}_{0,r}^{\beta_x})^{-1} \right) \times \qquad \text{(S5)}$$

$$\prod_{\ell=1}^{L} \mathrm{IG}\left(\delta_{\ell,(h)}^x \mid \frac{\nu_\ell^{\delta^x}}{2}, \frac{\nu_\ell^{\delta^x} s_{0,\ell}^x}{2} \right) \det(\boldsymbol{\Lambda}_{1,h}^*)^{-1/2} (b_{1,h}^*)^{-a_{1,h}^*},$$

where $\boldsymbol{\Lambda}_{1,h}^* = \boldsymbol{D}_h' \boldsymbol{D}_h + \boldsymbol{\Lambda}_0^*$; $b_{1,h}^* = \left[\nu_{\sigma^2} s_0 + \boldsymbol{y}_{(h)}' \boldsymbol{y}_{(h)} + (\boldsymbol{\beta}_0^*)' \boldsymbol{\Lambda}_0^* \boldsymbol{\beta}_0^* - (\boldsymbol{\beta}_{1,h}^*)' \boldsymbol{\Lambda}_{1,h}^* \boldsymbol{\beta}_{1,h}^* \right] /2$; $a_{1,h}^* = (\nu_{\sigma^2} + n_h^*)/2$; $\boldsymbol{\beta}_{1,h}^* = (\boldsymbol{\Lambda}_{1,h}^*)^{-1}(\boldsymbol{\Lambda}_0^* \boldsymbol{\beta}_0^* + \boldsymbol{D}_h' \boldsymbol{y}_{(h)})$; $\boldsymbol{y}_{(h)}$ is a n_h^*-length vector containing all y_t such that $s_t = h$; and $\boldsymbol{D}_h$ is a $n_h^* \times (L+1)$ design matrix whose rows correspond to $\boldsymbol{y}_{(h)}$ and are composed of $(1, \mu_{1,(h)}^x - y_{t-1}, \ldots, \mu_{L,(h)}^x - y_{t-L})$ for each t such that $s_t = h$. Note that proportionality in (S5) is preserved with respect to the $\{\mu_{\ell,(h)}^x\}$, which appear in the regression means for y_t. Aside from the factor containing normalizing weights in the mixture denominator, the full conditional for $\boldsymbol{\eta}_h^x$ could be factored into a series of conjugate updates that could serve as proposal distributions for a Metropolis step. This approach yields low acceptance rates in practice, and we instead utilize a random-walk Metropolis sampler with jointly Gaussian proposals for all parameters in $\boldsymbol{\eta}_h^x$ (with $\{\delta^x\}$ parameters

proposed on the logarithmic scale), which are evaluated using (S5). Proposals that produce computationally singular covariance matrices are automatically rejected.

The full conditional distribution for $\boldsymbol{\eta}_h^y$ factors as $p(\sigma_h^2 \mid \cdots, -\boldsymbol{\beta}_h^*)\, p(\boldsymbol{\beta}_h^* \mid \sigma_h^2, \cdots)$ and is drawn sequentially as $\sigma_h^2 \mid \cdots, -\boldsymbol{\beta}_h^* \sim \mathrm{IG}(a_{1,h}^*, b_{1,h}^*)$ and $\boldsymbol{\beta}_h^* \mid \sigma_h^2, \cdots \sim \mathrm{N}\left(\boldsymbol{\beta}_{1,h}^*, \sigma_h^2\, (\boldsymbol{\Lambda}_{1,h}^*)^{-1}\right)$.

Parameters in the base measure

Let $n^* = \sum_{h=1}^H 1_{(n_h^* > 0)}$ count the total number of occupied components. The posterior conditional density for $\boldsymbol{\beta}_0^*$ is proportional to
$\prod_{\{h:n_h^*>0\}} \left[\mathrm{N}(\boldsymbol{\beta}_h^* \mid \boldsymbol{\beta}_0^*, \sigma_h^2\, (\boldsymbol{\Lambda}_0^*)^{-1})\right] \mathrm{N}(\boldsymbol{\beta}_0^* \mid \boldsymbol{b}_0^*, \boldsymbol{S}_0^*)$, yielding a Gaussian update with covariance matrix $\boldsymbol{S}_1^* = \left(\sum_{\{h:n_h^*>0\}} \sigma_h^{-2}\, \boldsymbol{\Lambda}_0^* + (\boldsymbol{S}_0^*)^{-1}\right)^{-1}$ and mean
$\boldsymbol{S}_1^* \left((\boldsymbol{S}_0^*)^{-1}\, \boldsymbol{b}_0^* + \boldsymbol{\Lambda}_0^* \sum_{\{h:n_h^*>0\}} \sigma_h^{-2}\, \boldsymbol{\beta}_h^*\right)$.

The posterior conditional density for $(\boldsymbol{\Lambda}_0^*)^{-1}$ is proportional to
$\prod_{\{h:n_h^*>0\}} \left[\mathrm{N}(\boldsymbol{\beta}_h^* \mid \boldsymbol{\beta}_0^*, \sigma_h^2\, (\boldsymbol{\Lambda}_0^*)^{-1})\right] \mathrm{IWish}((\boldsymbol{\Lambda}_0^*)^{-1} \mid \nu^*, \nu^*\, \boldsymbol{\Psi}_0^*)$, yielding an inverse-Wishart update with degrees of freedom $\nu^* + n^*$ and scale matrix
$\nu^*\, \boldsymbol{\Psi}_0^* + \sum_{\{h:n_h^*>0\}} \sigma_h^{-2}\, (\boldsymbol{\beta}_h^* - \boldsymbol{\beta}_0^*)\, (\boldsymbol{\beta}_h^* - \boldsymbol{\beta}_0^*)'$.

The posterior conditional density for s_0 is proportional to
$\prod_{\{h:n_h^*>0\}} \left[\mathrm{IG}(\nu_{\sigma^2}/2, \nu_{\sigma^2}\, s_0/2)\right] \mathrm{Ga}(s_0 \mid a_{s_0}, b_{s_0})$, yielding a gamma update with shape $a_{s_0} + \nu_{\sigma^2}\, n^*/2$ and rate $b_{s_0} + \nu_{\sigma^2} \sum_{\{h:n_h^*>0\}} \sigma_h^{-2}/2$. Updates for $\{s_{0,\ell}^x\}$ are analogous, with $\delta_{\ell,(h)}^x$ replacing σ_h^2, except that all H values are required for each update.

All remaining parameters in the base measure have standard conditionally conjugate updates. Because all $\{\boldsymbol{\eta}_h^x\}$ parameters are used in the local $q_h(\boldsymbol{x})$ weights, the updates for associated G_0 parameters require all H values, rather than the n^* values associated with occupied components.

DP concentration parameter

The posterior full conditional density for the DP concentration parameter α is proportional to $\prod_{h=1}^{H-1} \left[\mathrm{Beta}(v_h \mid 1, \alpha)\right] \mathrm{Ga}(\alpha \mid a_\alpha, b_\alpha)$, yielding a gamma update with shape $a_\alpha + H - 1$ and rate $b_\alpha - \log(\omega_H)$.

Adaptation

After initialization, MCMC begins with a tuning phase for diagonal elements of the covariance matrix in the candidate-generating Gaussian proposal distribution for $\{\boldsymbol{\eta}_h^x\}_{h=1}^H$. Optionally, an adaptation phase may then be used to further tune a full candidate covariance matrix. This proceeds in four steps. In the first step, the initial covariance matrix is globally scaled to adjust acceptance rates collected over a short run. This is repeated iteratively until all acceptance rates fall within a pre-specified range (we set the range low, e.g., $[0.02, 0.20]$, to promote exploration) or a maximum number of attempts is reached. In the second step, the proposal variances are scaled locally by parameter groups corresponding to $\boldsymbol{\mu}^x$, $\{\boldsymbol{\beta}_r^x\}$, and $\boldsymbol{\delta}^x$, while preserving correlations. In the third step, empirical cross-covariance matrices

are estimated from a longer run. In the final step, these empirical covariance matrices are scaled globally until acceptance rates fall within the pre-specified range, or a maximum number of attempts is reached. At this point, adaptation ceases and the scaled empirical covariance matrices are used for subsequent random-walk proposals. We advise against adapting prematurely, which can cause the chain to develop an affinity for a local mode during burn-in.

S4   Details for lag selection

S4.1   Posterior inference with global selection

Conditional on $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_L)$, the selection effect on the mixture kernels can be passed through to the $\boldsymbol{D}_h$ matrices, for which all elements in column $\ell + 1$ are replaced with 0s if $\gamma_\ell = 0$. Because $\boldsymbol{\eta}_h^x$ is updated with a Metropolis step, one simply draws candidate values and evaluates (S5) with each $\mathrm{N}_{(h)}(\boldsymbol{y}_{t-1})$ for $h = 1, \ldots, H$, and $t = L+1, \ldots, T$, appropriately modified (with respect to $\boldsymbol{\gamma}$). The full conditional distribution for $\boldsymbol{\eta}_h^y$ is then sampled using the modified $\boldsymbol{D}_h$. All other updates proceed as before, using the appropriately modified $\mathrm{N}_{(h)}(\boldsymbol{y}_{t-1})$ and kernel means.

The posterior full conditional probability that $\gamma_\ell = 1$ is

$$\Pr(\gamma_\ell = 1 \mid \cdots) = \frac{\pi_\ell^\gamma a_\ell^\gamma}{\pi_\ell^\gamma a_\ell^\gamma + (1 - \pi_\ell^\gamma) b_\ell^\gamma}, \tag{S6}$$

where
$$a_\ell^\gamma = \prod_{t=L+1}^T \left[\mathrm{N}_{(s_t)}(\boldsymbol{y}_{t-1} \mid \gamma_\ell = 1) \, \mathrm{N}_{(s_t)}(y_t \mid \gamma_\ell = 1) \left(\sum_{j=1}^H \omega_j \, \mathrm{N}_{(j)}(\boldsymbol{y}_{t-1} \mid \gamma_\ell = 1) \right)^{-1} \right]$$
and
$$b_\ell^\gamma = \prod_{t=L+1}^T \left[\mathrm{N}_{(s_t)}(\boldsymbol{y}_{t-1} \mid \gamma_\ell = 0) \, \mathrm{N}_{(s_t)}(y_t \mid \gamma_\ell = 0) \left(\sum_{j=1}^H \omega_j \, \mathrm{N}_{(j)}(\boldsymbol{y}_{t-1} \mid \gamma_\ell = 0) \right)^{-1} \right].$$
The Gaussian kernel densities in these expressions are modified to reflect either $\gamma_\ell = 1$ or 0, and appropriately reflect all other $\{\gamma_k\}_{k \neq \ell}$.

Instead of drawing from the individual full conditionals, we update $\boldsymbol{\gamma}$ as a block using a collapsed conditional, with $\{\boldsymbol{\eta}_h^y\}$ integrated out, as in Supplement S3. First, a number of proposed switches, $k \in \{1, 2, 3\}$, is drawn from a truncated geometric distribution. Then, a uniformly drawn subset of k indices among $\{1, \ldots, L\}$, denoted $\mathcal{K}$, identifies which elements of the proposed $\boldsymbol{\gamma}^{\text{cand}}$ are switched from the current state (individually, from 0 to 1 or from 1 to 0). Symmetry of this proposal distribution (Schäfer and Chopin, 2013, Sec. 3.3) yields the Metropolis ratio, $p(\boldsymbol{\gamma}^{\text{cand}} \mid \cdots, -\{\boldsymbol{\eta}_h^y\}) / p(\boldsymbol{\gamma}^{\text{old}} \mid \cdots, -\{\boldsymbol{\eta}_h^y\})$, where

$$p(\boldsymbol{\gamma}_{\mathcal{K}} \mid \cdots, -\{\boldsymbol{\eta}_h^y\}) \propto \prod_{k \in \mathcal{K}} (\pi_k^\gamma)^{\gamma_k} (1 - \pi_k^\gamma)^{1-\gamma_k} \times \prod_{t=1}^T \left[\mathrm{N}_{(s_t)}(\boldsymbol{y}_{t-1}) \left(\sum_{j=1}^H \omega_j \, \mathrm{N}_{(j)}(\boldsymbol{y}_{t-1}) \right)^{-1} \right] \times$$
$$\prod_{h=1}^H \det(\boldsymbol{\Lambda}_{1,h}^*)^{-1/2} (b_{1,h}^*)^{-a_{1,h}^*}, \tag{S7}$$

S8

with all quantities calculated using the full $\boldsymbol{\gamma}$ vector under evaluation.

S4.2 Transition density estimation under global lag selection

For any value of y and $\boldsymbol{x}$, or over a multidimensional grid of values, samples for $f_{Y|X}$ are calculated from

$$\tilde{f}_{Y|X}(y \mid \boldsymbol{x}, \boldsymbol{\gamma}) = \sum_{h=1}^{H} \tilde{q}_h(\boldsymbol{x} \mid \boldsymbol{\gamma}) \, \mathrm{N}_{(h)}(y \mid \mu(\boldsymbol{x} \mid \boldsymbol{\gamma}), \sigma^2), \tag{S8}$$

with $\tilde{q}_h(\boldsymbol{x} \mid \boldsymbol{\gamma}) = \omega_h \, \mathrm{N}_{(h)}(\boldsymbol{x} \mid \boldsymbol{\gamma}) / \sum_{j=1}^{H} \omega_j \, \mathrm{N}_{(j)}(\boldsymbol{x} \mid \boldsymbol{\gamma})$ and $\mu(\boldsymbol{x} \mid \boldsymbol{\gamma}) = \mu^y - \sum_{\ell=1}^{L} \gamma_\ell \beta_\ell^y (x_\ell - \mu_\ell^x)$. The samples can then be used to create pointwise estimates and intervals for $\tilde{f}_{Y|X}$. The expression for the transition mean becomes $\tilde{\mathrm{E}}_{Y|X}(y \mid \boldsymbol{x}, \boldsymbol{\gamma}) = \sum_{h=1}^{H} \tilde{q}_h(\boldsymbol{x} \mid \boldsymbol{\gamma}) \mu_{(h)}(\boldsymbol{x} \mid \boldsymbol{\gamma})$. Analogous expressions include $\boldsymbol{\gamma}$ in the procedure for estimating quantiles in Section 2.4. Likewise, K-step-ahead forecasts are inductively sampled with $(s, y)_{T+k}$ pairs for $k = 1, \ldots, K$, following the first two levels of the hierarchical model (6), adjusted for $\boldsymbol{\gamma}$, for each posterior sample. While dependence on other parameters in (6) is implicit in the preceding expressions, we add explicit dependence on $\boldsymbol{\gamma}$ in order to emphasize the modifications necessary to include lag dependence.

S4.3 Posterior inference with local selection

The posterior full conditional probability that $\gamma_\ell^{(h)} = 1$ is

$$\mathrm{Pr}\left(\gamma_\ell^{(h)} = 1 \mid \cdots\right) = \frac{\pi_\ell^\gamma a_{h,\ell}^\gamma}{\pi_\ell^\gamma a_{h,\ell}^\gamma + (1 - \pi_\ell^\gamma) b_{h,\ell}^\gamma}, \tag{S9}$$

where
$a_{h,\ell}^\gamma = \prod_{\{t:s_t=h\}} \left[\mathrm{N}_{(h)}(\boldsymbol{y}_{t-1} \mid \gamma_\ell^{(h)} = 1) \, \mathrm{N}_{(h)}(y_t \mid \gamma_\ell^{(h)} = 1) \right] \prod_t \left(\sum_{j=1}^{H} \omega_j \, \mathrm{N}_{(j)}(\boldsymbol{y}_{t-1} \mid \gamma_\ell^{(h)} = 1) \right)^{-1}$
and
$b_{h,\ell}^\gamma = \prod_{\{t:s_t=h\}} \left[\mathrm{N}_{(h)}(\boldsymbol{y}_{t-1} \mid \gamma_\ell^{(h)} = 0) \, \mathrm{N}_{(h)}(y_t \mid \gamma_\ell^{(h)} = 0) \right] \prod_t \left(\sum_{j=1}^{H} \omega_j \, \mathrm{N}_{(j)}(\boldsymbol{y}_{t-1} \mid \gamma_\ell^{(h)} = 0) \right)^{-1}$.
Note that the first product in each line is over all time points allocated to mixture component h, and the second product is over all time points $t = L + 1, \ldots, T$. The Gaussian densities in these expressions are from (S1), modified to reflect either $\gamma_\ell^{(h)} = 1$ or 0, and appropriately reflecting all other $\{\gamma_k^{(i)} : i \neq h \text{ or } k \neq \ell\}$, which are held constant.

Instead of drawing from the individual full conditionals, we update each $\boldsymbol{\gamma}_h$ as a block using collapsed conditionals, with $\{\boldsymbol{\eta}_h^y\}$ integrated out, as in Supplement S3. First, a number of proposed switches, $k \in \{1, 2, 3\}$, is drawn from a truncated geometric distribution. Then, a uniformly drawn subset of k indices among $\{1, \ldots, L\}$, denoted $\mathcal{K}$, identifies which elements of the proposed $\boldsymbol{\gamma}_h^{\mathrm{cand}}$ are switched from the current state (individually, from 0 to 1 or from 1 to 0). Symmetry of this proposal distribution (Schäfer and Chopin, 2013, Sec. 3.3) yields

the Metropolis ratio, $p(\boldsymbol{\gamma}_h^{\text{cand}} \mid \cdots, -\{\boldsymbol{\eta}_h^y\})/p(\boldsymbol{\gamma}_h^{\text{old}} \mid \cdots, -\{\boldsymbol{\eta}_h^y\})$, where

$$
\begin{aligned}
p(\boldsymbol{\gamma}_{h,\mathcal{K}} \mid \cdots, -\{\boldsymbol{\eta}_h^y\}) \propto & \prod_{k \in \mathcal{K}} (\pi_k^\gamma)^{\gamma_k^{(h)}} (1 - \pi_k^\gamma)^{1-\gamma_k^{(h)}} \times \prod_{t:s_t=h} \mathrm{N}_{(h)}(\boldsymbol{y}_{t-1}) \times \\
& \prod_{t=1}^{T} \left[\left(\sum_{j=1}^{H} \omega_j \, \mathrm{N}_{(j)}(\boldsymbol{y}_{t-1}) \right)^{-1} \right] \det(\boldsymbol{\Lambda}_{1,h}^*)^{-1/2} (b_{1,h}^*)^{-a_{1,h}^*},
\end{aligned}
\tag{S10}
$$

with all quantities calculated using both the full $\boldsymbol{\gamma}_h$ vector under evaluation, and all other $\{\boldsymbol{\gamma}_j : j \neq h\}$ held constant.

Updates for π_ℓ^γ are presented in Chung and Dunson (2009), and proceed as follows. Introduce $\xi_\ell = 1_{(\pi_\ell^\gamma > 0)}$, for $\ell = 1, \ldots, L$. Then, conditional on $\xi_\ell = 1$, draw $\pi_\ell^\gamma \sim$ Beta $\left(a_\ell^\pi + \sum_{h=1}^{H} \gamma_\ell^{(h)}, b_\ell^\pi + H - \sum_{h=1}^{H} \gamma_\ell^{(h)} \right)$; otherwise, set $\pi_\ell^\gamma = 0$. Then, to update each ξ_ℓ, set $\xi_\ell = 1$ if $\sum_{h=1}^{H} \gamma_\ell^{(h)} > 0$; otherwise, the full conditional probability that $\xi_\ell = 1$ is $\alpha_\ell^\pi / (\alpha_\ell^\pi - 1 + 1/\pi_\ell^\pi)$, where $\alpha_\ell^\pi = \Gamma(b_\ell^\pi + H)\Gamma(a_\ell^\pi + b_\ell^\pi) / [\Gamma(b_\ell^\pi)\Gamma(a_\ell^\pi + b_\ell^\pi + H)]$.

S4.4 Illustration of global selection on linear autoregressive simulation

We demonstrate the model's ability to identify simple structure, for which the proposed model is over-specified. Although each of the nonlinear, non-Gaussian, and mixture capabilities are not necessary in this case, the model performs well. A stationary time series was generated from the model

$$
y_t = \mu + \phi_1(y_{t-1} - \mu) + \phi_2(y_{t-2} - \mu) + \epsilon_t, \qquad \epsilon_t \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2),
$$

with $\mu = 2.5$, $\phi_1 = 1.2$, $\phi_2 = -0.7$, and $\sigma^2 = 1.0$. We then fit the proposed nonparametric model to series of length $T = 75$ and $T = 305$ with a lag horizon of $L = 5$ (so that 70 and 300 observations contribute to the likelihood), DP truncation at $H = 25$, and default prior settings. With the short time series, both global and local lag selection recover the true structure, with all chains decisively selecting the first two lags only. With the long time series, all methods select the first two lags and include lag 3 in a non-negligible fraction of the samples (ranging from 0.2 to 0.6, with reasonable mixing in each chain). Other inferences from all fits appear to accurately recover the truth, with exception that a few runs with local selection occasionally tend to over-fit the data, adding one or two unnecessary mixture components.

Figure S2 provides trace plots for key quantities from one model fit to 70 observations with global selection, including the log-likelihood, number of occupied components, selection indicators for the first four lags, the observation (innovation) variance for the most populated component, the first three β^y coefficients for the most populated component, the center μ^y for the most populated component, and the intercept $(\mu^y + \sum_{\ell=1}^{L} \gamma_\ell \, \beta_\ell^y \, \mu_\ell^y)$ for the most populated component, thinned to 1,000 samples. The trace for the log-likelihood indicates that the chain is no longer traversing across substantially different component configurations. Most observations belong to one component throughout MCMC. Lags 1 and 2 are on for all
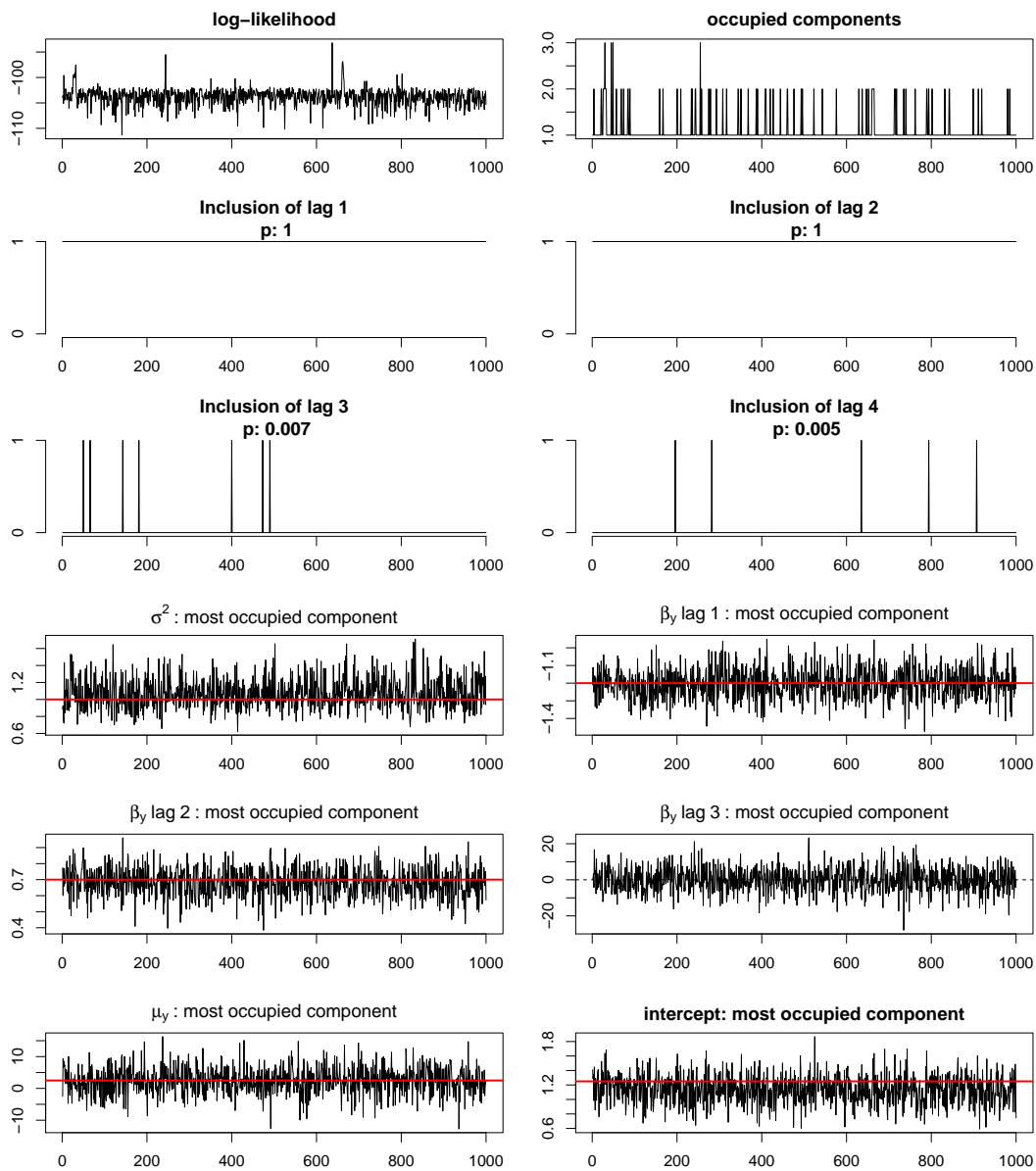
Figure S2: MCMC trace plots for the model fit to 70 observations of the simulated second-order autoregression with global lag selection. Red horizontal lines indicate true parameter values. In the lag-inclusion plots (second and third row), p refers to the Monte Carlo estimate of the posterior probability of inclusion for that lag.

inference samples, while the remaining lags are off for nearly all inference samples. Trace plots for the kernel parameters faithfully track the true values (indicated with horizontal red lines in the plots). Note that the sign is switched for the coefficients in the model formulation. The only trace without a precise marginal posterior distribution in this chain is for μ^y, which in the model is replaced by μ_ℓ^x parameters in the lag summands, and thus over-parameterized for this stationary time series. However, the intercept, which is a function of μ^y and lag coefficients is precisely identified. Trace plots for the coefficients of lags 4 and 5 are similar to that of lag 3, reflecting their prior with mean 0 in the next level of the hierarchy.

Inferences for the transition mean surface and transition densities for specific lag values (not shown), both as functions of y_{t-1} and y_{t-2}, are consistent with the data-generating mechanism. Specifically, the estimated mean surface is very close to the true plane. Posterior mean estimates of transition densities are nearly the correct Gaussian distributions. Furthermore, marginal posterior standard deviations are nearly identical to standard errors from a correctly specified linear model fit to the time series. Hence, conditional on admittedly overconfident lag inferences, the proposed model performs well in a simple scenario, with surprisingly low cost for additional flexibility.