UNIVERSITY OF CALIFORNIA
SANTA CRUZ

BAYESIAN MIXTURE MODELING AND ORDER SELECTION FOR MARKOVIAN TIME SERIES

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

STATISTICAL SCIENCE

by

Matthew J. Heiner

September 2019

The Dissertation of Matthew J. Heiner is approved:

Professor Athanasios Kottas, Chair and Primary Advisor

Professor Stephan Munch, Secondary Advisor

Professor Raquel Prado

Professor Abel Rodríguez

Quentin Williams
Acting Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Bayesian Mixture Modeling and Order Selection for Markovian Time Series

by

Matthew J. Heiner

Nonlinearity and high-order auto-dependence are common traits of univariate time series tracking successive states from multidimensional systems. Standard statistical models based on linear stochastic processes are often inadequate to capture these complex dynamics. This work contributes Bayesian statistical methodology and modeling strategies to estimate Markovian transition distributions, particularly when these distributions exhibit non-Gaussianity and/or nonlinear dependence on multiple lags. Given the challenge of modeling high-order nonlinear dynamics, we place emphasis on detecting and exploiting low-order dependence.

We propose models for both discrete and continuous state spaces with a common theme of mixture modeling. We first utilize mixtures for soft model selection. To this end, we develop two prior distributions for probability vectors which, in contrast to the popular Dirichlet distribution, retain sparsity properties in the presence of data. Both priors are tractable, allowing for efficient posterior sampling and marginalization. We derive the priors, demonstrate their properties, and employ them for lag selection in the mixture transition distribution model.

We then extend the model for estimation and selection in higher-order, discrete-state Markov chains with two primary objectives: parsimonious approximation of high-order dynamics by mixing transition models of lower order, and model selection through over-specification and shrinkage with the new priors to an identifiable and interpretable parameterization. We also extend a continuous-state version of the mixture transition distribution model by admitting nonlinear dependence in the

component distributions using Gaussian process priors. We discuss properties of the models and demonstrate their utility with simulation studies and applications to medical, geological, and ecological time series.

Finally, we propose and illustrate a Bayesian nonparametric autoregressive mixture model applied to flexibly estimate general transition densities exhibiting nonlinear lag dependence. Our approach is related to Bayesian curve fitting via joint density estimation using Dirichlet process mixtures, with the Markovian likelihood defined as the conditional distribution obtained from the mixture. We extended the model to include automatic relevance detection among a pre-specified set of lags. We illustrate the model by repeating earlier analyses.

For LauraDawn.

Acknowledgments

The text of Chapter 2 includes an adapted reprint of the following previously published article:

Heiner, M., Kottas, A., and Munch, S. (2019), "Structured priors for sparse probability vectors with application to model selection in Markov chains," *Statistics and Computing*, 29, 1077–1093, URL https://doi.org/10.1007/s11222-019-09856-2.

The co-authors listed in this publication supervised the research that forms the basis for the chapter. I also acknowledge the editors of *Statistics and Computing* and two anonymous referees for helpful suggestions. The research for Chapter 2 was supported in part by the National Science Foundation, award DMS 1310438. The research for Chapters 3, 4, and 5 was supported in part by the National Science Foundation, award SES 1631963.

Strong prior expectations accompanied our family to Santa Cruz in 2014. I have since accumulated ample data to confirm just how privileged I am to work with my advisors, Thanasis Kottas and Steve Munch. Beyond his esteemed expertise, work ethic, and exceptional attention to detail, Thanasis is a truly dedicated mentor who prioritizes his students' interests and genuinely celebrates their achievements. He has consistently been my advocate within and outside the department, dedicated time for discussions when I needed them most, and patiently listened. His ideas have repeatedly helped me branch to new solutions and resume progress when I've encountered barriers. Pursuing research with Thanasis has been a singular opportunity and a pleasure. Steve is a truly multidisciplinary scientist with contagious enthusiasm. His ideas and interests have influenced the trajectory of my studies and enriched my work. Steve's cheerful demeanor fosters camaraderie within his lab, of which I have been a grateful participant.

My research interests quickly expanded after I arrived in Santa Cruz and enrolled in Raquel Prado's course on time series analysis. Since then, Professor Prado has been like another advisor to me, consistently interested in my progress and quick to offer encouragement. Throughout my research, I have repeatedly drawn from the practical knowledge provided in just one class from Abel Rodríguez, whose expertise and projects are wide-ranging. I am grateful for the perspective and ideas he has offered as a member of my dissertation committee.

I am also grateful to the rest of the Statistics faculty at UCSC for their support. In particular, Juhee Lee was a wonderful first-year advisor. Herbie Lee generously provided funding and mentoring for a unique opportunity to design and develop my own online course. Bruno Sansó provided career opportunities and guidance. The department entrusted me with responsibility for four undergraduate courses, providing a substantial leg-up with my teaching career.

In two summers at Lawerence Livermore National Laboratory, I learned much working with Francisco Beltrán and Ana Kupresanin, who have shown continued interest in my progress. I also thank the San Francisco Bay Area Chapter of the American Statistical Association for sponsoring my travel to the Joint Statistical Meetings in 2018, as well as Adrian Raftery, who at the conference offered encouragement and suggested a connection that proved helpful.

Fellow students have been integral to my education. Many friends, including Daniel Kirsner, Kurtis Shuler, and Dan Spencer, have continuously lifted my work and my spirits. I consider Devin Francom, Robert Richardson, and Yifei Yan as both friends and mentors. Credit goes to Arthur Lui, who introduced me to the *Julia* programming language and has been my computing lifeline. These friends, and their families, have been like family to us in Santa Cruz.

Many contributed to my development as a statistician. Gilbert Fellingham, my

advisor at Brigham Young University, introduced me to Bayesian statistics and encouraged a trajectory that has not only been fun and rewarding, but a blessing to our family. I acknowledge my enthusiastic high school statistics teacher, Vicki Lyons, whose course accelerated a germinating interest. I am grateful to former university president and mission leader, Steven Bennion, who in me saw potential as a teacher.

Credit for anything I accomplish will forever be shared with my wife, LauraDawn. Our life in Santa Cruz has been a team effort, which has often required patience and sacrifice on her part. I am continuously grateful for her extraordinary, unwavering companionship. I am frequently astonished by the incredible blessing I have of being LauraDawn's husband and Lea's father. Extended family, including my parents, Joe and Kathy, and parents-in-law, Don and Adele, are also a constant strength and support to us. I credit my father for (perhaps unintentionally) instilling in me a fascination with probability and statistics in my youth, and my brother-in-law, James, for his trailblazing pursuit of that interest.

I could thank so many others. I conclude with humble acknowledgement that all these associations and circumstances are ultimately gifts from God. We have often seen His guiding hand in our life, whether in answered prayers, blessings, fortuitous meetings, or earthly angels in our path. The Gospel of His Son, Jesus Christ, is a light to us that encourages us to seek continual learning and improvement.

Chapter 1

Introduction

All data implicitly carry time stamps. Whether neglected or accounted for, time information can profoundly affect inferences drawn from observations. Time dependence in data is a consequence of our living in a dynamic world, and methods to account for this dependence feature prominently in ecology, engineering, economics, social science, medicine, and environmental science, to name a few. The rapid development of time-series analysis (and statistics, generally) over the past century runs concurrent with technological advances enabling its implementation. By time series, we refer to data collected and indexed at discrete time steps. Frequent objectives of time series analysis include forecasting, signal processing, pattern or trend recognition, inference for dynamics, and accounting for serial correlation in order to accurately infer functional, and perhaps causal, relationships. Shumway and Stoffer (2017) provide a thorough introduction to the core methods for analysis of time series data, which are typically modeled as stochastic processes, either on the time or frequency domain.

This thesis is concerned with Bayesian methods for flexible inference of transition distributions. That is, we work in the time domain and model the conditional probability distribution of the next observation, given the current and past values

of a univariate time series of categorical or continuous measurements. In many applications, forecasts of central tendency are inadequate. Financial projections are incomplete without estimates of risk, motivating models that incorporate volatility. Dynamic decision analyses for population management require probabilistic transition distributions from current and hypothetical states in order to optimize for maximum sustainable yield. These and other investigations benefit from, or require full specification of the transition distribution (Rodríguez and Ter Horst, 2008).

Many useful and widely applied methodologies target the transition mechanism, most notably the linear autoregressive (AR) model and its many variants (Box and Jenkins, 1976). AR models are stationary and linear Gaussian processes, and thus possess several appealing properties with respect to estimation, interpretation, and diagnostics. Nevertheless, real-world processes often exhibit nonlinearity and non-Gaussianity. Failure to account for these two features can severely limit the utility of standard methods, which if misapplied can produce misleading results. Modeling the important characteristics of nonlinearity and non-Gaussianity plays a central role in this work. We proceed by motivating our methodological development in Section 1.1. Section 1.2 identifies our research objectives and provides an outline of the thesis.

1.1 Modeling time dependence

Most models on the time domain incorporate serial dependence in one of two ways. First, one can model variables as a function of time directly, which is effective for filtering/smoothing noisy time series, and for capturing periodic behavior, trends, and changepoints. This approach usually ignores the directionality of time and the causal nature of sequential events. The second approach models

the present state as a function of past states, inducing memory in the system. This approach, which captures periodic behavior, trends, and changepoints more subtly, is consistent with the concept of dynamical systems, for which differential (or difference) equations dictate the trajectory of a system. This thesis focuses on methods for the second approach of modeling state dynamics.

The dynamical-system view of time series gives rise to state spaces, which one may model with observed time series directly, or infer by assuming measurement error in the observations (see Prado and West, 2010 for a review). State-space models most often refer to the latter case, in which a modeler explicitly distinguishes between measurement and dynamical error, or random perturbations at the system level. These two error types are often very challenging to separate without detailed knowledge of the measuring process or by making strong assumptions (Kantz and Schreiber, 2004, pp. 174-175). Furthermore, flexible modeling for complex dynamics that are indirectly observed poses a formidable computational challenge (Prado and West, 2010, ch. 6). In practice, measurement error is often negligible, or less consequential than dynamical error. Because we seek to capture complex dynamics and are modeling transition densities directly, we restrict our attention to modeling the observations themselves as measurements from the state space, and assume that observed noisy shocks at each observation time propagate forward with the system.

With the exception of long-memory processes, it is often assumed that influence diminishes as lag time increases. This is especially true of nonlinear systems in the presence of dynamical noise, which could mask the signal after a small number of transitions. This statistically justifies the Markovian assumption of conditional independence between the present state and distant history, given recent lags. Markovian dynamics are also directly justified in systems for which the future

trajectory only depends on the current state. The Markovian assumption is crucial to the dynamical system approach to modeling time series, and we adopt it into our methodology.

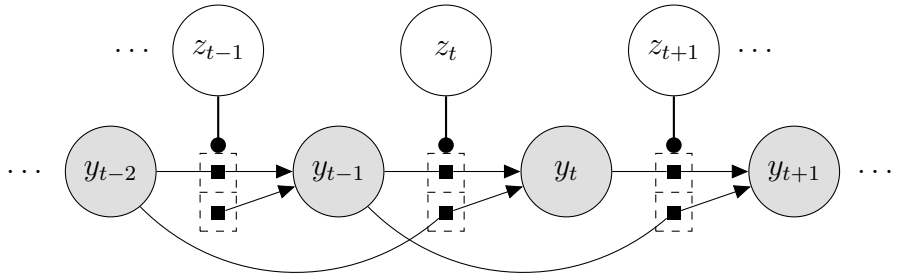
1.1.1 Mixture modeling for time series

In this thesis, we propose methods for both discrete and continuous-state Markovian time series analysis. While the methods are diverse, a recurring theme throughout is our use of discrete mixture modeling, through which a probability distribution function comprises a positively weighted sum of a finite or countable set of distribution functions. Common uses of mixture models include accounting for heterogeneity among observations in the absence of covariates or other identifying information, and approximation of complex (e.g., heavy-tailed or multimodal) distributions from simpler distributions. Frühwirth-Schnatter (2006) provide a comprehensive introduction. Our proposed methods employ mixture modeling for both uses, and further utilize mixtures for two less common purposes: 1) soft (probabilistic) selection among competing models, and 2) local selection of simple models as a vehicle to flexibly capture complex structure in a global model.

As a weighted sum of distribution functions, discrete mixtures can always be cast as hierarchical models involving random latent indicators whose probabilities are the weights. Breaking the mixture in this way can be instructive and provide an avenue for exploring dependence structures. For example, hidden Markov Models (HMMs) can be cast as dependent mixtures for which the latent states follow a discrete Markov chain.

Panel (a) in Figure 1.1 depicts the primary dependence structure utilized throughout Chapters 2, 3, and 4 with a graphical model. Shaded nodes represent observations and white nodes represent the latent variables. Conditional

(a)



(b)

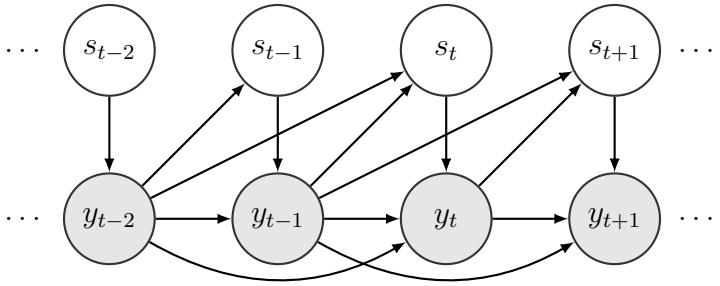


Figure 1.1: Directed graphs illustrating a second-order example of the dependence structure utilized throughout Chapters 2, 3, and 4 (a), and in Chapter 5 (b). Gray nodes represent observations; white nodes represent latent variables; solid squares represent distinct probability distributions; and dotted boxes represent gates that select, conditional on the values of associated latent variables, which distributions govern the observations. This notation follows Dietz (2010).

dependence is indicated with directed arrows. This example depicts second-order dynamics, where the latent variables control (through the dotted square gates) which lags drive the transition distributions (solid squares). This is the structure implied by using the mixture transition distribution (MTD) of Raftery (1985). Panel (b) depicts the dependence structure utilized in Chapter 5, again with a second-order example. Here, the distributions of observables *and* latent variables are influenced by lags of the observables. Note that in both panels, the latent variables do not exhibit Markov dependence. This choice stems from our objective to model transition distributions, and their functionals, directly. Additional dependence in the second level complicates this objective, especially for higher-order

hidden processes, since the transition distributions of observables are obtained by marginalizing over the latent states. Although this choice eliminates a potential source for time heterogeniety in the model, we note that inclusion of higher-order dynamics can potentially offset the drawbacks of this simplifying assumption.

1.1.2 Order and lag selection

Aside from simple constructions, such as the linear autoregressive class of models, most mainstream statistical methods for modeling Markovian dynamics assume dependence on the first lag only. While convenient and often justified, this assumption clearly over-simplifies or misspecifies the dynamics in some cases, and carries similar consequences to under-specifying models in other regression settings. As an example, the life cycle of Pink salmon in Alaska, U.S.A., which includes fresh and salt water phases before they return to the stream to spawn and die, reliably follows a two-year pattern (Heard, 1991). Thus, naive modeling of the population dynamics based on the first lag only would capture inter-population, rather than generational dynamic dependence.

Another motivation for modeling multiple lags of a time series stems from a technique called phase-space reconstruction via time-delay embedding. This relies on a theorem by Takens (1981) that justifies reconstructing the dynamics, up to topological equivalence, of a smooth-trajectory, multidimensional deterministic system using only lags of a univariate time series. The practical utility of this result is evident in fields like ecology, where full observation of all relevant variables may be impossible.

Although Takens' theorem for time-delay embedding applies to deterministic dynamics, Markovian stochastic models can approximate this method for real time series, which inevitably contain measurement and dynamical-type noise (Kantz and

Schreiber, 2004, ch. 10, 12). Nearest-neighbor or local linear regression methods are often used for nonparametric empirical forecasting in this setting, which can be useful in distinguishing deterministic chaos from stochastic processes (Sugihara et al., 1990).

Of course, including too many lags in a model leads to inefficiency, the "curse of dimensionality," and inferential difficulties associated with including multiple correlated predictors. Thus, it is also important to condition on no more lags than necessary to faithfully estimate a transition distribution. In light of these considerations, we place emphasis throughout this thesis on the problem of inferring the order of Markovian dependence, and more precisely, which lags influence the transition distribution. Within the classes of models we explore, this is intimately related to the model/variable selection problem so fundamental to regression.

1.2 Research objectives

The primary objective of this work is to contribute Bayesian statistical methodology and modeling strategies to estimate transition distributions, particularly when these distributions exhibit non-Gaussianity and/or nonlinear dependence on multiple lags. Given the challenge of modeling high-order nonlinear dynamics, we place emphasis on detecting and exploiting low-order dependence from an initial set of candidate lags.

This thesis is divided into two parts, corresponding to methodologies for discrete-state Markov chains in Chapters 2 and 3, and continuous-state series in Chapters 4 and 5. In addition to providing a natural venue for categorical time series, discrete-state Markov chains are appealing for their utility in distilling information from complex systems with a coarse representation, and that their canonical form accommodates both nonlinearity and general distributions. In

deterministic systems, generating partitions of the state space can be combined with countable lag sequences to fully represent the dynamics (Hirata et al., 2004). While we work in a stochastic setting and do not pursue estimating such partitions, this idea further motivates our study and modeling for high-order, discrete-state Markov chains.

In pursuit of a parsimonious model for high-order Markov chains, we utilize the MTD structure depicted in Panel (a) of Figure 1.1. To detect lower-order dependence, we impose sparsity with a pair of novel prior distributions for probability vectors, which are introduced and studied in Chapter 2. They are first applied to infer a single active lag in the MTD framework before we scale to inferences for multiple active lags using an extension of the MTD in Chapter 3. The primary methodological contributions of Chapter 2 are the priors themselves, initial exploration of their properties leading to recommendations for their use, and their novel application in the MTD model. The primary contributions of Chapter 3 are the higher-order extension of the MTD, application of the new priors to promote identifiability in a commonly used extension of the MTD, and simulation studies for comparison with other methods.

Although Chapter 4 applies to continuous-state time series, it utilizes the same modeling framework as the previous two chapters. We propose a semiparametric mixture model with flexible component mean functions using Gaussian process priors, and again use the priors for sparse probability vectors to identify lag dependence as in Chapter 2. Here, the mixture model has two (possibly dual) roles, providing flexibility in the transition density, and lag selection. The primary contributions of Chapter 4 are extensions, both of methods developed in previous chapters to continuous state spaces as well as existing MTD methods for continuous state spaces, and investigation of the model's intended use.

Chapter 5 is unique in methodology, complexity, and generality. The proposed model, our initial approach to the problems addressed in this thesis, employs a fully nonparametric specification based on the Dirichlet process. It is rooted in Bayesian curve fitting (Müller et al., 1996), but the likelihood is built from a conditional rather than a joint distribution. The resulting model is very flexible and well-suited for our objectives. The primary contributions of Chapter 5 are the operational extension of similar existing models to multiple lags, extension and modeling framework for exploration of lag dependence, and investigation into the model's fitness for different analysis scenarios. Results throughout this thesis owe much to the challenges encountered while pursuing this framework.

A word on notation: care has been taken to unify basic representations in the notation across chapters, especially for indexing. However, specific symbols should only be interpreted within the context of their chapter. For example, while $\{s_t\}$ universally represents a collection of discrete states indexed by time, it refers to the observed time series in earlier chapters and to a process of latent states indicating mixture component membership in Chapter 5.

Throughout the thesis, analyses were conducted and plots generated with the R statistical computing language (R Core Team, 2016). In most cases, posterior sampling via Markov chain Monte Carlo was conducted using the *Julia* scientific computing language (Bezanson et al., 2017). Multidimensional plots were generated with *Plotly* (Plotly Technologies Inc., 2015).

Chapter 2

Structured Priors for Sparse Probability Vectors

2.1 Introduction

The most common approach to Bayesian modeling of probability vectors uses the Dirichlet prior (see Agresti and Hitchcock, 2005 and references therein). This prior possesses numerous desirable features: it is conjugate in the multinomial setting, and can often be made so in more general modeling settings by introducing latent variables; the hyperparameters are interpretable; and the family is stable under aggregation and marginalization. Due to the convenience and universality of this prior, few alternatives have gained traction in the literature. One alternative, the logistic normal distribution (Atchison and Shen, 1980), relaxes the property that Dirichlet variates are always negatively correlated. More recently, Elfadaly and Garthwaite (2017) proposed a Gaussian copula-based prior which "binds" beta marginals, also allowing more general correlation structures. Agresti and Hitchcock (2005) provide background and review of the Dirichlet prior's use,

including hierarchical and mixture extensions proposed by Good (1976) and Albert and Gupta (1982) for use in contingency tables. One useful generalization of the Dirichlet distribution by Connor and Mosimann (1969), used extensively for its connection with the stick-breaking, constructive definition of the Dirichlet process (Sethuraman, 1994), has also found application in life testing (Lochner, 1975) and mixture modeling (Bouguila and Ziou, 2004).

While the Dirichlet prior shrinks proportions away from 0 and 1, one may instead seek prior models which favor sparsity. By sparsity we mean many or most of the entries of the probability vector are near 0. Sparse probability vectors for which all entries are non-zero, but most are near 0, can be modeled with a single Dirichlet distribution by lowering the shape parameter for sparse components to values below unity. If this is the case for all shape parameters, prior probability mass resides primarily in the corners of the simplex supporting the probability vector, leading to either small or large probabilities in each component. While this strategy encourages sparsity in a prior model, it fails to carry this property to the posterior since the shape parameters of the Dirichlet distribution are inseparably connected with the precision. That is, a sparse Dirichlet prior is also a low-precision Dirichlet prior, for which even small sample sizes immediately overwhelm sparsity properties.

Consider a hierarchical model which at one level employs a probability vector to mix over a discrete set. In some cases, particularly if the set consists of competing elements, the modeler may wish to encourage sparsity. For example, a time series may be approximately Markovian, but the active lag is unknown to the modeler. One alternative to fitting and comparing several models is to include lag as a component of the model to be inferred. Rather than mixing over several competing sub-models (lags) with the Dirichlet prior, a structured prior favoring sparsity (one

of the lags) would essentially select one lag. If the time series arises from a nearly deterministic dynamical system, enforcing sparsity can assist in learning both the optimal lag and transition dynamics, as we will illustrate in Section 2.3.2.

This chapter explores structured priors which are capable of retaining sparsity properties in the presence of data. This may be of interest in a multinomial setting in which a small, unknown subset of the categories have non-negligible probabilities. Similarly, we may wish to discount categories with small observed counts as anomalies. Additionally, shrinkage can be helpful when using a probability vector as a mixing distribution in which components include competing models, favoring a model-selection role over a model-averaging role. We address both of these potential uses. In Section 2.2, we develop two sparsity-inducing prior probability models and study their properties. We then demonstrate their use in a hierarchical Markov chain model with unknown active lag and sparse transition dynamics in Section 2.3, first with a simulation study (Section 2.3.2) and then with a time series of Chinook salmon abundance in Northern California, U.S.A. (Section 2.3.4). Finally, we conclude with discussion in Section 2.4.

2.2 Prior models for sparse probability vectors

A rich literature exists for both discrete and continuous shrinkage estimators/priors in linear models, the most popular being the Lasso method (Tibshirani, 1996; Park and Casella, 2008). Such models shrink unrestricted coefficients in an attempt to select among many competing predictors. The Bayesian variable selection literature also includes stochastic search or "spike-and-slab" priors, typically characterized by two-component mixture priors for coefficients allowing for "on" and "off" settings (George and McCulloch, 1993, 1997). Furthermore, regularization is a general characteristic of Bayesian models, whose priors bias or

penalize regions of the parameter/model space.

Although we draw on foundational concepts such as penalties and stochastic search, the problem of shrinkage for probability vectors requires additional care to satisfy the sum-to-one constraint. One starting point may be to consider a transformation of unrestricted latent variables and introduce traditional shrinkage priors to enforce sparsity. For example, we could normalize a set of positive random variables drawn from a distribution with tails that are heavier than the gamma (used in the constructive definition of the Dirichlet distribution). In addition to scaling issues, this approach lacks convenient computational properties such as conjugacy. We instead propose extensions of the Dirichlet and generalized Dirichlet distributions that maintain computational convenience.

In Sections 2.2.1 and 2.2.2, we propose two novel structured prior models which encourage sparsity in probability vectors. We examine some basic properties of these models in Section 2.2.3.

2.2.1 Sparse Dirichlet mixture prior

The sparse Dirichlet mixture (SDM) prior model is motivated by the idea of penalized maximum likelihood. We work directly with the conjugate Dirichlet prior by adding a multiplicative term to the prior density that favors some desired property. If $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)$ is the probability vector of interest, we can write the density as

$$p(\boldsymbol{\theta}) \propto \text{Dir}(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \times h(\boldsymbol{\theta}) \,, \tag{2.1}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_J)$ is a vector of positive real shape parameters. One choice for $h(\boldsymbol{\theta})$ which penalizes low variance (thus rewarding high variance) is $h(\boldsymbol{\theta}) = \sum_{j=1}^{J} \theta_j^2$. More generally, we might replace each θ_j^2 with $\theta_j^{\beta_j}$, $\beta_j > 1$. We propose using a

common $\beta = \beta_1, \ldots, \beta_J$, resulting in a one-parameter extension of the original Dirichlet prior. Setting $\beta = 1$, we recover the original Dirichlet prior. As the value of β increases, the model forces prior mass to the corners of the simplex S supporting $\boldsymbol{\theta}$.

Integrating to find the normalizing constant c yields the probability density function for the SDM model:

$$\begin{aligned} p_{\text{SDM}}(\boldsymbol{\theta} \mid \boldsymbol{\alpha}, \beta) &= \frac{\Gamma(\sum_{i=1}^{J} \alpha_i)}{c \prod_{r=1}^{J} \Gamma(\alpha_r)} \sum_{j=1}^{J} \prod_{h=1}^{J} \theta_h^{(\alpha_h + \beta 1_{(h=j)} - 1)} \\ &= \sum_{j=1}^{J} \frac{w_j}{\sum_{h=1}^{J} w_h} \operatorname{Dir}(\boldsymbol{\theta} \mid \boldsymbol{\alpha} + \beta \mathbf{e}_j)\,, \end{aligned} \tag{2.2}$$

where $w_j \equiv \prod_{h=1}^{J} \Gamma(\alpha_h + \beta 1_{(h=j)})$ and $\mathbf{e}_j$ is a vector of 0s with a 1 in the jth position, and $\Gamma(\cdot)$ is the gamma function. Expression (2.2) reveals that the SDM prior is a discrete mixture of Dirichlet densities in which the shape parameter of component j is increased by β. If all shape parameters are equal ($\alpha_1 = \ldots = \alpha_J = \alpha$), then the mixture weights are equal, resulting in a discrete uniform mixture of Dirichlet densities and yielding a symmetric prior model.

Because it is a fixed-weight, discrete mixture of conjugate priors, the SDM prior retains tractability under multinomial count data $\boldsymbol{n} = (n_1, \ldots, n_J)$. Exploiting the conjugacy of the Dirichlet prior in expression (2.1) immediately reveals a SDM posterior with parameters $\boldsymbol{\alpha}^* \equiv \boldsymbol{\alpha} + \boldsymbol{n}$ and β. Just as the α shape parameters are interpretable as prior pseudo-counts, β can be interpreted as the sample-size equivalent of each component's boost. Hence, an advantage of the SDM prior is that a practitioner can select a boost (or bias) factor apriori while remaining agnostic about which component should receive it. This interpretation of β, together with the mixture representation of the density (2.2), provides intuition for how the penalty function forces probability mass to the corners of the simplex. The

SDM density is a superposition of J Dirichlet densities, wherein the jth density modifies the original Dirichlet prior by forcing mass toward the $\theta_j = 1$ corner by a sample-size equivalent of β.

As the data sample size $N \equiv \sum_{j=1}^{J} n_j$ increases, the influence of a fixed β decreases and each Dirichlet component becomes similar to the same Dirichlet distribution that would result from a standard Dirichlet prior. We can also consider the posterior mean point estimator of θ_j obtained readily from (2.2),

$$\mathrm{E}(\theta_j \mid \boldsymbol{n}) = \frac{\alpha_j + n_j + u_j^* \beta}{A + N + \beta}, \tag{2.3}$$

where $A \equiv \sum_{j=1}^{J} \alpha_j$, $u_j^* \equiv w_j^* / \sum_{h=1}^{J} w_h^*$, and $w_j^* \equiv \prod_{h=1}^{J} \Gamma(\alpha_h + n_h + \beta 1_{(h=j)})$. If $\boldsymbol{\alpha}$ and β are fixed, (2.3) approximates the maximum likelihood estimator n_j/N, which converges in probability to $\boldsymbol{\theta}$ in the multinomial likelihood model.

The posterior expectation in (2.3) is plotted for a simple Bernoulli-beta scenario in Figure 2.1. The expectation for θ_1 when $\boldsymbol{n} = (0, 1)$ monotonically decreases as a function of β, while the shrinking effect of β diminishes marginally. When β or N is moderately large, the posterior weights w_j^* shift substantially in favor of the component with largest $\alpha_j^* = \alpha_j + n_j$, effectively resulting in a single Dirichlet

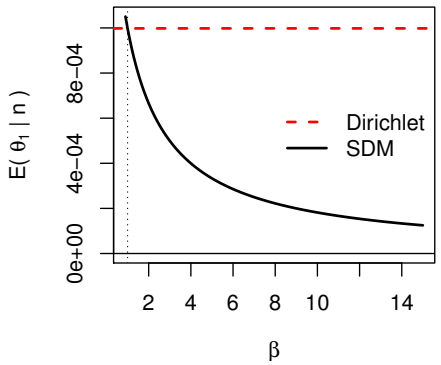


Figure 2.1: Posterior mean point estimate of θ_1 under the Dirichlet prior (dashed red) and SDM prior (black) for varying values of β. Here, the multinomial data are $\boldsymbol{n} = (0, 1)$ and $\alpha_1 = \alpha_2 = 0.001$.

distribution with a boost of size β given to the component with highest α_j^*.

The sample-size interpretation of β suggests that one may factor N into its selection in order to maintain influence. Indeed, the effects of fixed, finite-parameter priors in fixed, finite-parameter models asymptotically vanish in general. Allowing β to increase with sample size invalidates the large-sample results noted earlier. However, (2.2) guarantees that the posterior distribution will always be a convex combination of Dirichlet distributions. Because u_j^* is bounded between 0 and 1, a practitioner can use (2.3) to select β to control bounds on the posterior mean's bias from that of the baseline Dirichlet prior.

In practice, we have selected the value of β as some function of the data sample size N, typically a fixed scaling $\beta = CN$ with $C > 1/N$. To see the effect of this choice on the posterior mean of θ_j, substitute for β in (2.3). For $N \gg A$, the posterior mean is approximated by $(n_j/N + u_j^* C)/(1 + C)$. While this choice has a substantial effect in the multinomial setting (see Figure 2.1), results are relatively insensitive to the choice of β when the SDM prior is used as a mixing distribution in hierarchical models such as those in Section 2.3.

One may also consider placing a prior on β. This, however, breaks the simple conjugacy of the model, necessitating more complex estimation methods such as MCMC to estimate a parameter that is only weakly identified by the data. We therefore advocate fixing β with the preceding discussion in mind.

We note that the modified Dirichlet density in (2.1) was explored by Hjort (1996) in the context of histogram estimation. They utilized a different form of $h(\boldsymbol{\theta})$ to promote positive correlation among adjacent probabilities in $\boldsymbol{\theta}$ and achieve a smoothing effect. They reference other penalized-likelihood approaches aimed at smoothing probability vectors and reported the prior-conjugacy result for general $h(\boldsymbol{\theta})$. In contrast with this previous work, our $h(\boldsymbol{\theta})$ function promotes sparsity.

2.2.2 Stick-breaking mixture prior

The stick-breaking mixture (SBM) prior model builds the probability vector $\boldsymbol{\theta}$ through an extension of the stick-breaking construction that defines the generalized Dirichlet distribution (Connor and Mosimann, 1969). In particular,

$$\theta_1 = Z_1,\ \theta_j = Z_j \prod_{h=1}^{j-1}(1 - Z_h) \text{ for } j = 2, \ldots, J-1, \text{ and } \theta_J = \prod_{h=1}^{J-1}(1 - Z_h)\,, \quad (2.4)$$

with $Z_j \overset{\text{ind.}}{\sim} \text{Beta}(a_j, b_j)$, for $j = 1, \ldots, J-1$. The Dirichlet distribution on $\boldsymbol{\theta}$ is a special case of the generalized Dirichlet distribution. Specifically, if $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_J)$ is the Dirichlet shape parameter vector, setting $a_j = \alpha_j$ and $b_j = \sum_{h=j+1}^{J} \alpha_h$ for $j = 1, \ldots, J-1$, yields a $\text{Dir}(\boldsymbol{\alpha})$ distribution for $\boldsymbol{\theta}$ (Connor and Mosimann, 1969). Typically in practice, the Z_j are iid, resulting in a stochastically ordered (first $J-1$ elements of the) $\boldsymbol{\theta}$ vector. To allow gaps between large elements of $\boldsymbol{\theta}$ in a parsimonious way, we propose a mixture of three beta distributions,

$$Z_j \overset{\text{ind.}}{\sim} \pi_1 \,\text{Beta}(1, \eta) + \pi_2 \,\text{Beta}(\gamma_j, \delta_j) + \pi_3 \,\text{Beta}(\eta, 1)\,, \quad (2.5)$$

for $j = 1, \ldots, J-1$, where we specify π_1 and π_3 as probabilities (with $\pi_1 + \pi_3 < 1$) and $\pi_2 = 1 - \pi_1 - \pi_3$. This form (2.5) allows us to encourage sparsity by setting η large, in which case the first component corresponds to small probabilities in $\boldsymbol{\theta}$. Figure 2.2 illustrates the role of each beta distribution from (2.5), wherein the second and third components allow flexibility in modeling non-negligible probabilities. If π_1 is large, much of the original unit stick may be left unused, for which we include a third component facilitating use of the remaining stick before reaching θ_J. As an example, consider modeling a probability vector (with $J > 6$) in which θ_3, θ_4, and θ_6 are relatively large and all others are small. Then Z_j for $j = 3, 4$ could come from the second $\text{Beta}(\gamma_j, \delta_j)$ component, Z_6 could come from

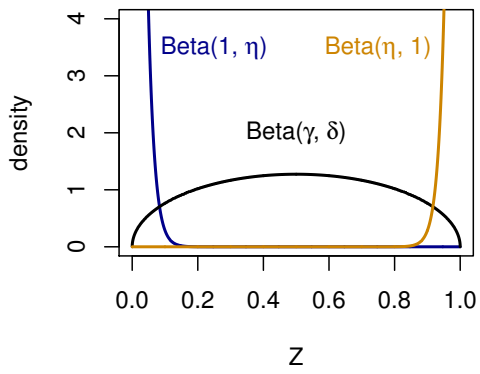


Figure 2.2: Individual beta densities comprising the three-component mixture used to draw stick-breaking latent variates Z_j from (2.5) in the SBM prior. The values used here (for illustration–not necessarily recommended) are $\eta = 50$ and $\gamma = \delta = 1.5$.

the third Beta$(\eta, 1)$ component, and all others from the Beta$(1, \eta)$ component.

One option for the second mixture component is to fix $\gamma_j = \gamma$ and $\delta_j = \delta$ for all $j = 1, \ldots, J-1$, resulting in a five-parameter prior. Alternatively, one could select $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{J-1})$ and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_{J-1})$ to mimic the Dirichlet special case, resulting in a flexible three-parameter extension of the Dirichlet prior. In that case, $\pi_2 = 1$ would yield a Dirichlet prior.

The independence of the Z_j latent variables is critical to maintaining computational simplicity. To facilitate posterior simulation, we introduce independent configuration variables ξ_j for $j = 1, \ldots, J-1$ which take on values 1 with probability π_1, 2 with probability π_2, and 3 otherwise. Conditional on ξ_j, the distribution of Z_j is Beta$(1, \eta)$ if $\xi_j = 1$, Beta(γ_j, δ_j) if $\xi_j = 2$, and Beta$(\eta, 1)$ otherwise. The joint prior distribution of $\boldsymbol{Z} = (Z_1, \ldots, Z_{J-1})$ and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_{J-1})$ is then

$$p(\boldsymbol{Z}, \boldsymbol{\xi}) = \prod_{j=1}^{J-1} \Big(\left[\pi_1 \,\text{Beta}(Z_j \mid 1, \eta)\right]^{1(\xi_j=1)} \times \left[\pi_2 \,\text{Beta}(Z_j \mid \gamma_j, \delta_j)\right]^{1(\xi_j=2)} \times \left[\pi_3 \,\text{Beta}(Z_j \mid \eta, 1)\right]^{1(\xi_j=3)} \Big).$$

If we write the likelihood for multinomial data in terms of $\boldsymbol{Z}$ using (2.4) and

integrate the resulting posterior over $\boldsymbol{Z}$, we obtain independent marginal posterior distributions for the configuration variables,

$$\Pr(\xi_j = r \mid \boldsymbol{n}) \propto \begin{cases} \pi_1 \, \eta \, \frac{\Gamma(1+n_j)\,\Gamma(\eta+\sum_{i=j+1}^{J} n_i)}{\Gamma(1+\eta+\sum_{h=j}^{J} n_h)} & \text{for } r = 1\,, \\ \pi_2 \, \frac{\Gamma(\gamma_j+\delta_j)\,\Gamma(\gamma_j+n_j)\,\Gamma(\delta_j+\sum_{i=j+1}^{J} n_i)}{\Gamma(\gamma_j)\,\Gamma(\delta_j)\,\Gamma(\gamma_j+\delta_j+\sum_{h=j}^{J} n_h)} & \text{for } r = 2\,, \\ \pi_3 \, \eta \, \frac{\Gamma(\eta+n_j)\,\Gamma(1+\sum_{i=j+1}^{J} n_i)}{\Gamma(\eta+1+\sum_{h=j}^{J} n_h)} & \text{for } r = 3\,. \end{cases} \tag{2.6}$$

Full conditional distributions for the $\{Z_j\}$ are then $p(Z_j \mid \xi_j = 1, \boldsymbol{n}) = \text{Beta}(Z_j \mid 1 + n_j, \eta + \sum_{h=j+1}^{J} n_h)$, $p(Z_j \mid \xi_j = 2, \boldsymbol{n}) = \text{Beta}(Z_j \mid \gamma_j + n_j, \delta_j + \sum_{h=j+1}^{J} n_h)$, and $p(Z_j \mid \xi_j = 3, \boldsymbol{n}) = \text{Beta}(Z_j \mid \eta + n_j, 1 + \sum_{h=j+1}^{J} n_h)$ independently for each $j = 1, \ldots, J-1$. If each of π_1, π_3, η, $\{\gamma_j\}$, and $\{\delta_j\}$ are fixed, Monte Carlo simulation from the posterior distribution of $\boldsymbol{\theta}$ proceeds by first drawing from the marginal posterior of $\boldsymbol{\xi}$, followed by the full conditional for $\boldsymbol{Z}$, and finally constructing the sample for $\boldsymbol{\theta}$ according to (2.4).

We advocate fixing the parameters in the SBM prior to produce intended behavior. The value of η controls the size of the negligible probabilities in $\boldsymbol{\theta}$. If $\gamma_j = \gamma$ and $\delta_j = \delta$ are fixed at single values (the five-parameter prior), we recommend relatively non-informative (small) values to accommodate the variety of Z_j needed to produce a range of non-negligible probabilities from the remaining stick (i.e., $\prod_{h=1}^{j-1}(1 - Z_h)$). If a baseline Dirichlet prior is used with $\gamma_j = \alpha_j$ and $\delta_j = \sum_{h=j+1}^{J} \alpha_h$ for Dirichlet shape parameter $\boldsymbol{\alpha}$ (the three-parameter extension), we recommend scaling the δ_j to reflect prior expectation of sparsity in $\boldsymbol{\theta}$. One way to do this is use the expected proportion of non-negligible entries in $\boldsymbol{\theta}$ given by

$$d = \begin{cases} \frac{\pi_1+(1-\pi_1)(1-[1-\pi_3]^J)/\pi_3}{J} & \text{if } \pi_3 > 0\,, \\ \frac{(1-\pi_1)(J-1)+1}{J} & \text{if } \pi_3 = 0\,, \end{cases} \tag{2.7}$$

so that $\delta_j = d \sum_{h=j+1}^{J} \alpha_h$. This expression (2.7) is derived in Appendix A.1 and takes into account (on average) that once a $\xi_j = 3$ is drawn, then all θ_h with $h > j$ are negligible. The scaling factor accounts for average behavior rather than adjusting δ_j conditional on $\boldsymbol{\xi}$, which would substantially complicate posterior sampling.

We recommend setting π_1 near or slightly below the intended/anticipated prevalence of sparsity, with π_3 taking a small fraction of what remains. To be more precise, one may use the expected level of sparsity in (2.7) as a guide. The values of π_1 and π_3 should also be chosen with care due to the asymmetry and truncation of the SBM prior, which we discuss in Section 2.2.3. One may consider placing a prior on $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)$ and the parameters of the second-component beta distribution. Using a Dirichlet prior on $\boldsymbol{\pi}$ results in a Dirichlet full conditional involving counts of the configuration variables $\{\xi_j\}$, and is amenable to Gibbs sampling. The other two parameters, γ and δ, have nonstandard updates, which may be sampled discretely over a grid. Experience suggests that assigning priors to these weakly identified parameters (in a simplified version of the model with π_3=0) has a minimal effect on resulting inferences and performance. Further note that interpretation of posterior distributions for the parameters is not straightforward. For example, the fraction of negligible probabilities in $\boldsymbol{\theta}$ is a function of both π_1 and π_3.

The posterior behavior of the SBM model under multinomial sampling is more complicated than for the SDM because mixtures occur at the level of the beta-distributed latent variables. However, examining the full conditional updates of the Z_j variables suggests that the posterior distributions arising from the SBM and Dirichlet priors become similar with large sample sizes. The posterior mean point estimate under the SBM also becomes similar to the maximum likelihood

estimator n_j/N for large N. Consider the posterior mean for θ_j conditional on $\boldsymbol{\xi}$,

$$\begin{aligned}
\mathrm{E}(\theta_j \mid \boldsymbol{n}, \boldsymbol{\xi}) &= \mathrm{E}[Z_j \prod_{h=1}^{j-1}(1 - Z_h) \mid \boldsymbol{n}, \boldsymbol{\xi}] \\
&= \frac{a_j + n_j}{a_j + b_j + \sum_{h=j}^{J} n_h} \prod_{h=1}^{j-1} \frac{b_h + \sum_{\ell=h+1}^{J} n_\ell}{a_h + b_h + \sum_{i=h}^{J} n_i} \\
&= \frac{a_j + n_j}{a_1 + b_1 + N} \prod_{h=1}^{j-1} \frac{b_h + \sum_{\ell=h+1}^{J} n_\ell}{a_{h+1} + b_{h+1} + \sum_{i=h+1}^{J} n_i}\,, \qquad (2.8)
\end{aligned}$$

where $(a_j, b_j) \in \{(1, \eta), (\gamma_j, \delta_j), (\eta, 1)\}$, depending on ξ_j, for each $j = 1, \ldots, J-1$. If all prior parameters are fixed, then for $\min(\{n_j\}) \gg \max(\eta, \{\gamma_j\}, \{\delta_j\})$, all terms in the product after the first approach 1. A term can fail to converge to 1 only if $n_\ell = 0$ for all $\ell > h$, in which case it approaches $b_h/(a_{h+1} + b_{h+1}) < \infty$. However, this only occurs if $n_j = 0$ also, so that the first term approaches 0 and consequently $\mathrm{E}(\theta_j \mid \boldsymbol{n}, \boldsymbol{\xi})$ approaches 0 as well. Irrespective of the values in $\boldsymbol{\xi}$, we have $\mathrm{E}(\theta_j \mid \boldsymbol{n}, \boldsymbol{\xi}) \approx n_j/N$ for large N. Thus for large samples, $\mathrm{E}[\theta_j \mid \boldsymbol{n}] = \mathrm{E}_{\boldsymbol{\xi}}[\mathrm{E}(\theta_j \mid \boldsymbol{n}, \boldsymbol{\xi}) \mid \boldsymbol{n}] \approx \mathrm{E}_{\boldsymbol{\xi}}[n_j/N \mid \boldsymbol{n}] = n_j/N$.

2.2.3 Properties of the SDM and SBM priors

The SDM and SBM prior models arise from quite different approaches. They likewise exhibit distinct properties which yield advantages and disadvantages in different modeling scenarios. One primary advantage shared by both models is that fixing the hyperparameters, as is typically done with the standard Dirichlet prior, admits direct sampling from the posterior distribution of $\boldsymbol{\theta}$.

The sparse Dirichlet mixture prior model could appropriately be called a "winner takes all" prior, as the component of (2.1) involving β forces probability mass toward the corners of the simplex supporting $\boldsymbol{\theta}$. If the Dirichlet density in (2.1) is symmetric, then the SDM prior is also symmetric. If a symmetric prior

with $\alpha < 1$ is used and there is a tie among highest multinomial counts, the posterior distribution will be multimodal. Increasing β forces prior mass deeper into the corners of the simplex, resulting in strongly biased estimates of all non-zero probabilities in $\boldsymbol{\theta}$, as posterior mass for the probability associated with the highest multinomial count will move toward 1 while the mass for all others move toward 0. This behavior may be desired when a modeler believes that only one of the multinomial components is active, or in a model or variable selection scenario where we wish to softly favor a single "selected" component. Even if there are two active components, the SDM can modestly improve estimation relative to the corresponding Dirichlet prior in small samples if $\beta < N$. Because the SDM is a fixed-weights mixture of Dirichlet densities, it retains negative correlations between all modeled probabilities.

The stick-breaking mixture prior model is based on a sequential construction mechanism (2.4) which typically precludes prior symmetry. Furthermore, inferences for $\boldsymbol{\theta}$ are not invariant to permutation of the indices j, as with the Dirichlet prior (Wong, 1998). These properties, coupled with the mixture model for the stick-breaking Z_j variables can result in non-negligible aberrations in prior (and posterior) estimates of probabilities at the end of the $\boldsymbol{\theta}$ vector. One can assess these issues using the prior expectation of $\boldsymbol{\theta}$. Consider the special case with the single γ, δ pair and $\pi_3 = 0$. Letting $\mu_Z = \mathrm{E}(Z_j) = \pi_1[1/(1+\eta)] + (1-\pi_1)[(\gamma/(\gamma+\delta)]$, independence of the Z_j yields $\mathrm{E}(\theta_j) = \mu_Z(1-\mu_Z)^{j-1}$, for $j = 1, \ldots, J-1$, and $\mathrm{E}(\theta_J) = (1-\mu_Z)^{J-1}$. A large value of π_1 leads to much of the stick remaining unbroken before reaching $\theta_J = 1 - \sum_{j=1}^{J-1} \theta_j$, resulting in a large value for θ_J. Likewise, small values of π_1 or large values of π_3 can lead to early consumption of the stick and small probabilities toward the end of the $\boldsymbol{\theta}$ vector.

These biases in the SBM reduce as J increases, in which case encouraging

sparsity makes sense. The three-parameter extension of the Dirichlet together with sparsity-corrected $\{\delta_j\}$, as well as the third component of the mixture in (2.5), can also alleviate the prior bias. Figure 2.3 demonstrates the effect of different SBM settings on components of $\boldsymbol{\theta}$ through simulation. The three-parameter extension helps restore symmetry in j. While increasing π_3 accentuates stochastic ordering in $\boldsymbol{\theta}$, it also reduces the magnitude of θ_J. We emphasize that these biases are greatly reduced with the introduction of just one data observation. Despite these artifacts of the model, our experience is that the SBM prior model produces more faithful estimates than the SDM model when $\boldsymbol{\theta}$ contains two or more non-negligible probabilities, and can outperform the standard Dirichlet prior in a variety of small-sample scenarios. We recommend using the three-parameter extension of

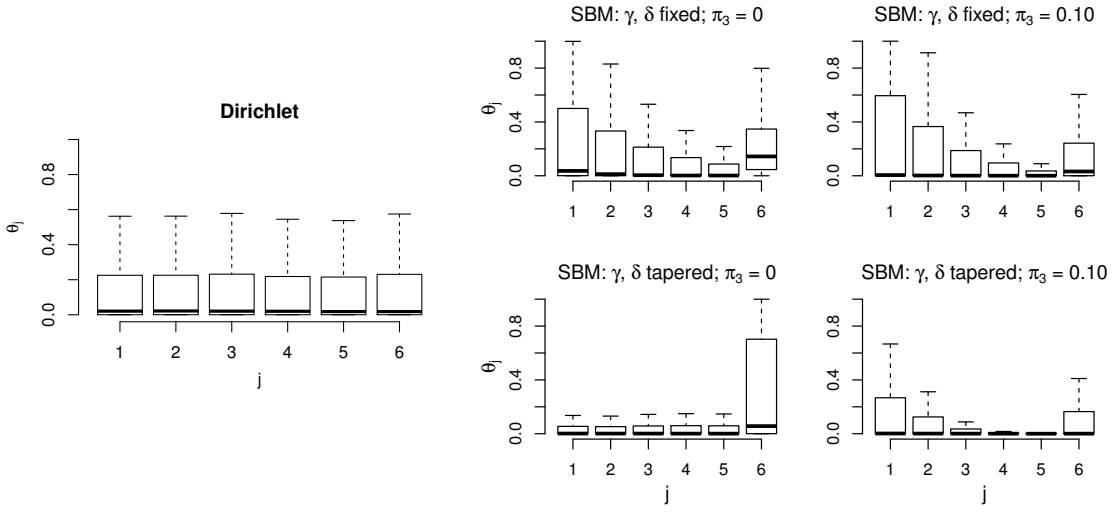


Figure 2.3: Box plots summarizing 10,000 simulated values of θ_j, for $j = 1, \ldots, J$, with $J = 6$, which demonstrate the effects of γ, δ specification and π_3. The Dirichlet prior (far left) is included for reference. Simulations on the top row are from the SBM prior with $\gamma = \delta = 1.5$ fixed, and simulations from the bottom row are from the SBM prior with Dirichlet shape parameters informing $\{\gamma_j\}, \{\delta_j\}$ and a sparsity correction to $\{\delta_j\}$ using (2.7). The left plots have $\pi_3 = 0$ and right plots have $\pi_3 = 0.1$. In all simulations, the Dirichlet shape parameters are $\alpha_j = 1/J$, $\eta = 1{,}000$, and $\pi_1 = 0.5$.

the Dirichlet with η large (we typically use $\eta = 1{,}000$), π_1 close to the anticipated level of sparsity, and $0 < \pi_3 < \pi_1$.

Under certain conditions, the generalized Dirichlet distribution admits positive correlations among elements of the modeled probability vector (Wong, 1998). Specifically, θ_i and θ_h are positively correlated for $1 < i < h \leq J$ if and only if the condition $(a_i + b_i)/(a_i + b_i + 1) > \prod_{j=1}^{i-1} b_j/(b_j + 1)$ holds, where $\{a_j\}$ and $\{b_j\}$ are the parameters of the beta distributions for the Z_j variables. Although the mixture structure of the SBM prior model adds further complication, this condition can have high posterior probability under some data scenarios if, for example, η is relatively small and $\gamma/(\gamma + \delta)$ is close to 1. Usually, the SBM distribution yields negative correlations.

To illustrate some of the properties of these models, we consider two multinomial data scenarios with $J = 3$, which allows us to visualize the posterior density for $\boldsymbol{\theta}$ as ternary plots in Figure 2.4. Two data vectors, $\boldsymbol{n}_1 = (0, 3, 3)$ and $\boldsymbol{n}_2 = (0, 3, 5)$, update a standard symmetric Dirichlet prior in addition to the two proposed models. For both the SDM and SBM models, hyperparameter settings were relatively mild, in that they did not strongly enforce sparsity. In the first scenario, two categories share the maximum count, leading the Dirichlet and SDM models to retain symmetry (and bimodality in the SDM case) in the joint posterior of θ_2 and θ_3. Although asymmetric, the posterior density under the SBM model maintains some neutrality between θ_2 and θ_3 while decisively shrinking θ_1. In the second scenario, both proposed models favor θ_3 despite only a slightly higher count in n_3.

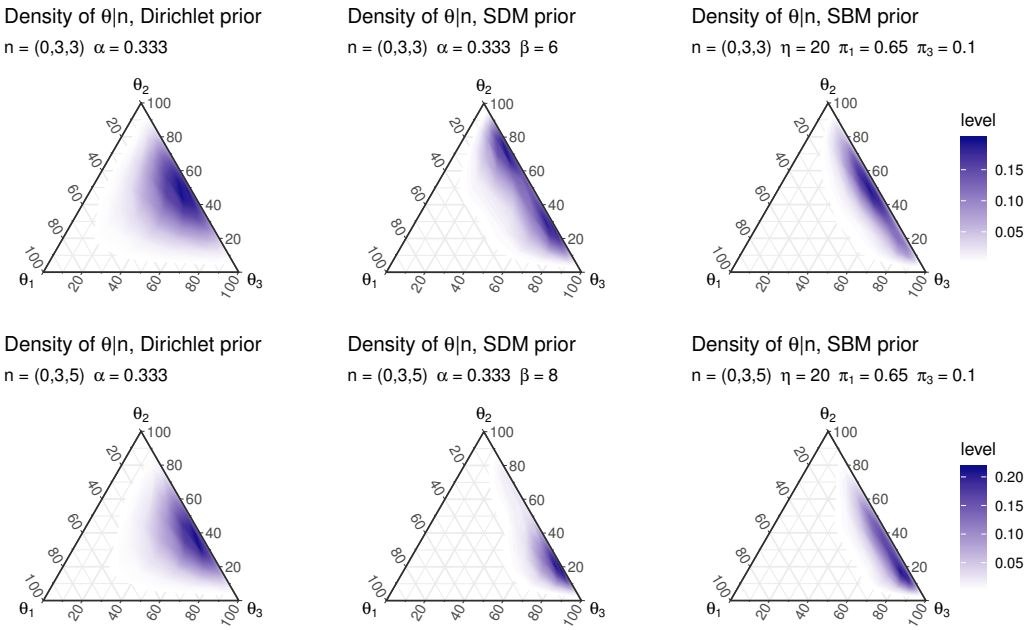


Figure 2.4: Posterior kernel density estimates for a probability vector $\boldsymbol{\theta}$ under two multinomial data scenarios ($\boldsymbol{n}_1 = (0, 3, 3)$ top, $\boldsymbol{n}_2 = (0, 3, 5)$ bottom) and three prior models: Dirichlet (left), SDM (center), and SBM (right). Fixed hyperparameter values are reported in the plots. The SBM prior is the three-parameter extension of the Dirichlet prior with sparsity correction on δ_j. Shading scales vary by plot, but are similar to those shown for the SBM model. The plots were generated using the *ggtern* package (Hamilton, 2017).

2.3 Application: Markov chains with lag uncertainty

We have illustrated the effects of the SDM and SBM prior models in the simplest setting of multinomial count data. A more interesting and potentially powerful scenario is to leverage sparsity in hierarchical models, using apriori knowledge to softly impose structure one level removed from observable quantities.

We illustrate this concept by fitting a Markov chain with unknown lag dependence. Markov chain modeling provides an appealing framework for discrete time series as well as for uncovering nonlinear dynamics. In the latter case, we may even

consider discretization of continuous time-series data. Sparsity becomes important if we consider the underlying dynamical system to be nearly deterministic. Model selection challenges arise as we consider the optimal lag in a system. We begin by introducing a hierarchical model in Section 2.3.1 which utilizes both priors to address each of these objectives. We apply this model to a simulated dynamical system in Section 2.3.2 and proceed with an illustrative example using ecological data in Section 2.3.4.

2.3.1 Bayesian lag estimation under the mixture transition distribution model

Consider a time series of nominal or ordinal values $s_t \in \{1, \ldots, K\}$ for $t = 1, \ldots, T$. Suppose our dual objectives in fitting a time-homogeneous Markov chain model to this series are estimation of the transition dynamics and selection of a *single* active lag. With Markov chains, order (or lag) is typically selected by maximizing a (possibly penalized) likelihood (as in Katz, 1981; Raftery, 1985; Prado and West, 2010), performing trans-dimensional MCMC (Green, 1995; Insua et al., 2012), using Bayes factors (Fan and Tsai, 1999; Bacallado, 2011; Zucchini and MacDonald, 2009), predictive criteria, or classical hypothesis tests (Bartlett, 1951; Besag and Mondal, 2013). Each of these approaches requires either fitting multiple models or complex estimation methods. Our approach is to build lag inference into a single model using our proposed priors applied to a popular mixture model for high-order Markov chains.

The mixture transition distribution (MTD) model, introduced by Raftery (1985) and reviewed in Berchtold and Raftery (2002), is a parsimonious model for high-order Markov chains. It approximates transition probabilities from a transition probability tensor $\boldsymbol{\Omega}$ as linear combinations of probabilities from a single

column-stochastic matrix $\boldsymbol{Q}$ and adds just one parameter for each additional lag (λ_ℓ), similar to autoregressive models. The transition probabilities in a model of order L are given as

$$\Pr(s_t = k_0 \mid s_{t-1} = k_1, \ldots, s_{t-L} = k_L) = (\boldsymbol{\Omega})_{k_0,k_1,\ldots,k_L} = \sum_{\ell=1}^{L} \lambda_\ell \, q_{k_0,k_\ell} \,, \tag{2.9}$$

where $q_{i,h} \equiv (\boldsymbol{Q})_{i,h}$, $0 \leq \lambda_\ell \leq 1$ and $\sum_{\ell=1}^{L} \lambda_\ell = 1$. This form (2.9) suggests that lags which play a prominent role in the transition probability for s_t will have relatively large λ_ℓ and lags which are not important to the transition will have λ_ℓ values near 0. Hence, inferences for $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_L)$ potentially yield information about important lags for the Markov process. It is apparent from (2.9) that $\lambda_\ell = 0$ is sufficient for conditional independence of s_t and $s_{t-\ell}$. If the columns of $\boldsymbol{Q}$ are unique, then $\lambda_\ell = 0$ is also a necessary condition for conditional independence. Inferences on $\boldsymbol{\lambda}$ have been employed to understand lag importance informally (Raftery and Tavaré, 1994), although the standard method for assessing order has been to compare BIC values (Berchtold and Raftery, 2002). Tank et al. (2017) recently targeted $\boldsymbol{\lambda}$ with a sparsity-inducing penalty to estimate Granger causality networks among multiple time series. One advantage of using (2.9) is increased flexibility over models for which lag dependence is fixed across all observations.

Here, we study the utility of the SDM and SBM priors in the context of Bayesian inference under the MTD model. Conditioning on the first L observations, the likelihood for the observed sequence is given by

$$p\left(\{s_t\}_{t=L+1}^{T} \mid \{s_t\}_{t=1}^{L}, \boldsymbol{\lambda}, \boldsymbol{Q}\right) = \prod_{t=L+1}^{T} p\left(s_t \mid \{s_{t-\ell}\}_{\ell=1}^{L}, \boldsymbol{\lambda}, \boldsymbol{Q}\right) = \prod_{t=L+1}^{T} \sum_{\ell=1}^{L} \lambda_\ell \, q_{s_t,s_{t-\ell}} \,.$$

Assuming independent Dirichlet priors on $\boldsymbol{\lambda}$ and on the columns of $\boldsymbol{Q}$, a data augmentation scheme with lag indicators yields tractable posterior conditional

distributions amenable to Gibbs sampling (Insua et al., 2012, pp. 59-60). Denote these indicators with $\{z_t\}$, for which if $z_t = \ell$, then $p(s_t \mid s_{t-1}, \ldots, s_{t-L}, \boldsymbol{Q}, z_t = \ell) = q_{s_t, s_{t-\ell}}$. While computationally convenient, this sampling scheme suffers from mixing challenges. Two alternatives are to 1) integrate $\boldsymbol{Q}$ out of the joint posterior and sample the marginal conditional posterior distributions for $\boldsymbol{\lambda}$ and $\{z_t\}$; or 2) use the likelihood without augmentation and employ Metropolis-Hastings. Both methods result in improved mixing, and we opt for the former.

In scenarios where the modeler believes one lag should dominate, but is unsure which it is, we advocate to replace the Dirichlet prior for $\boldsymbol{\lambda}$ with one favoring sparsity, particularly the SDM prior. Placing the SBM prior on $\boldsymbol{\lambda}$ is also appropriate and more in the spirit of the original MTD, as it allows two or more lags to significantly contribute to the transition distribution. If the transition probability matrix is known to be sparse, as is the case with our simulated dataset in the following section (see Figure 2.6, right panel), we propose replacing the Dirichlet priors on the columns of $\boldsymbol{Q}$ with independent SBM priors. These priors may assist to more precisely identify sparse structure than Dirichlet priors which tend to average over many components. Our proposed hierarchical model is given by:

$$
\begin{aligned}
&\Pr(s_t = k_0 \mid s_{t-1} = k_1, \ldots, s_{t-L} = k_L, \boldsymbol{Q}, z_t = \ell) = (\boldsymbol{Q})_{k_0, k_\ell}\,, \\
&\qquad \text{for } k_h = 1, \ldots, K;\ h = 0, \ldots, L;\ \ell = 1, \ldots, L; \text{ and } t = L+1, \ldots, T, \\
&\Pr(z_t = \ell \mid \boldsymbol{\lambda}) = \lambda_\ell \text{ indep. for } t = L+1, \ldots, T\,, \\
&\boldsymbol{\lambda} \sim \mathrm{SDM}(\boldsymbol{\alpha}_\lambda, \beta_\lambda)\,, \\
&(\boldsymbol{Q})_{\cdot,k} \overset{\text{ind.}}{\sim} \mathrm{SBM}(\boldsymbol{\pi}_k, \eta_k, \boldsymbol{\gamma}_k, \boldsymbol{\delta}_k)\,, \text{ for } k = 1, \ldots, K.
\end{aligned}
\tag{2.10}
$$

If we fix the hyperparameters $\boldsymbol{\alpha}_\lambda$, β_λ, $\{\boldsymbol{\pi}_k\}$, $\{\eta_k\}$, $\{\boldsymbol{\gamma}_k\}$, and $\{\boldsymbol{\delta}_k\}$, posterior Gibbs sampling can proceed with tractable conditional distributions. We improve mixing

by integrating $\boldsymbol{Q}$ from the joint posterior, sampling the marginal conditional distributions for $\boldsymbol{\lambda}$ and $\{z_t\}$. These are supported by the tractable marginal distributions derived in Appendix A.2. Additionally, to encourage occasional jumps between modes of the posterior, we include a hybrid independence Metropolis step which jointly proposes $\boldsymbol{\lambda}$ and $\{z_t\}$ from their prior every 25 iterations of MCMC.

To obtain results in Sections 2.3.3 and 2.3.4 that follow, each model was initialized with random draws from the Dirichlet prior for $\boldsymbol{\lambda}$ and discrete uniform for $\{z_t\}$. Random initialization in these models necessitated long burn-in periods, on the order of tens to hundreds of thousands of iterations. In our analyses, 500,000 burn-in iterations were followed by another one million iterations, producing convergent chains suitable for inference. Reported posterior quantities were calculated using a thinned sample retaining every 50th iteration. Full details of the MCMC algorithm are given in Appendix A.3.

2.3.2 Simulated dynamical system

We illustrate this Bayesian MTD model on applications for low-dimensional, nearly-deterministic dynamics in which only one lag dominates, and all but a few entries of the transition probability matrix are near 0. These two characteristics provide a natural setting for demonstrating the utility of our sparsity-favoring priors. We begin with a nonlinear, continuous-state system generated using a transition map adapted from a classical model for stock and recruitment of fish (Ricker, 1954), with additive Gaussian noise:

$$y_t = y_{t-2} \exp(\varphi - y_{t-2}) + \epsilon_t \,, \quad \epsilon_t \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2) \,. \tag{2.11}$$

We use $\varphi = 2.6$ and $\sigma = 0.09$. Simulated values were retained after a burn-in period of 1,000 transitions. Lag plots are shown in Figure 2.5. In order to capture

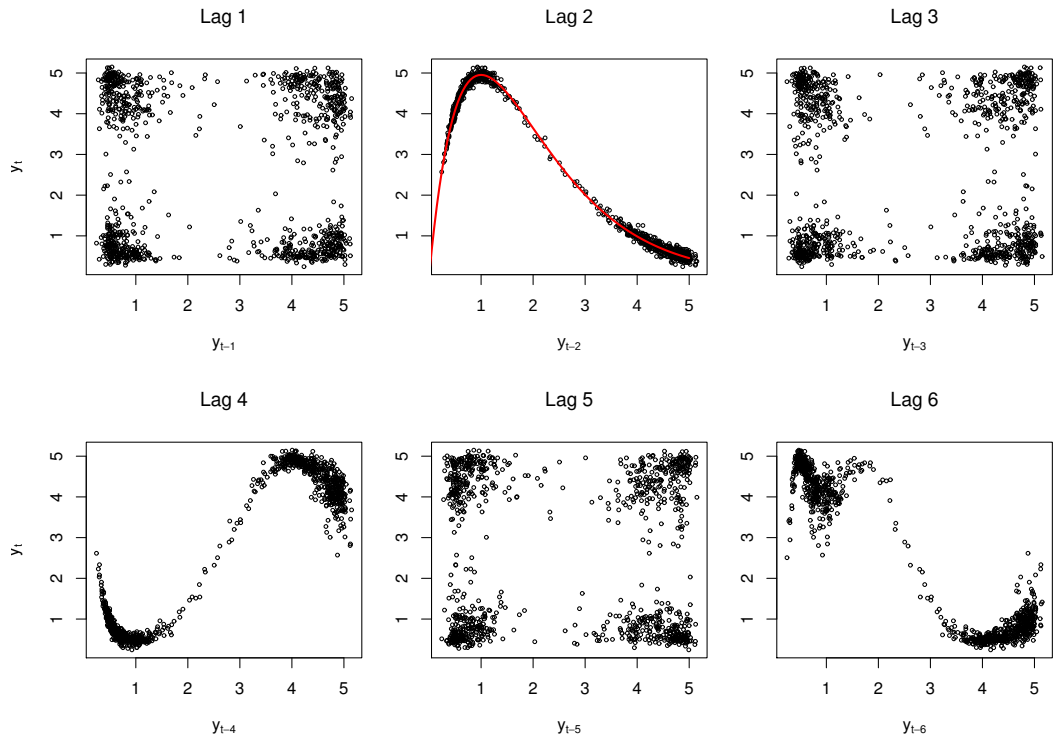
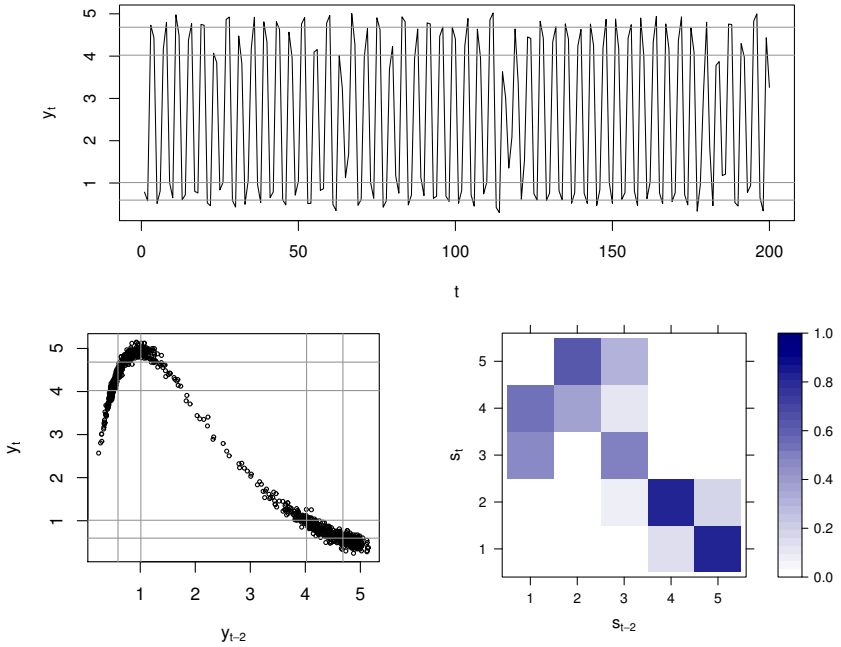


Figure 2.5: Lag scatter plots for 993 steps of the simulated dynamical model. The red curve in the lag 2 plot indicates the true mean transition function.

the nonlinear transitions in a simple and robust way, the continuous observations $\{y_t\}$ were binned into $K = 5$ and $K = 10$ ordered states $\{s_t\}$. In deterministic systems, so-called Markov and generating partitions of the state space encode dynamics with finite symbol sets. While strategies exist for estimating such partitions using noisy time series with unknown maps (Hirata et al., 2004), we take a general approach, electing to bin by quantiles (calculated from a window of 1,000 steps). For instance, $y_{550} = 0.508$. In the $K = 5$ discretization, all observations in the interval $(-\infty, 0.818]$ correspond to State 1 so that $s_{550} = 1$. In the $K = 10$ discretization, $y_{550} = 0.508 \in (0.496, 0.596]$, the interval for State 2, so that $s_{550} = 2$. The boundaries for state definitions are shown in Figure 2.6, along with an approximate transition probability matrix computed for the second lag

$K = 5$



$K = 10$

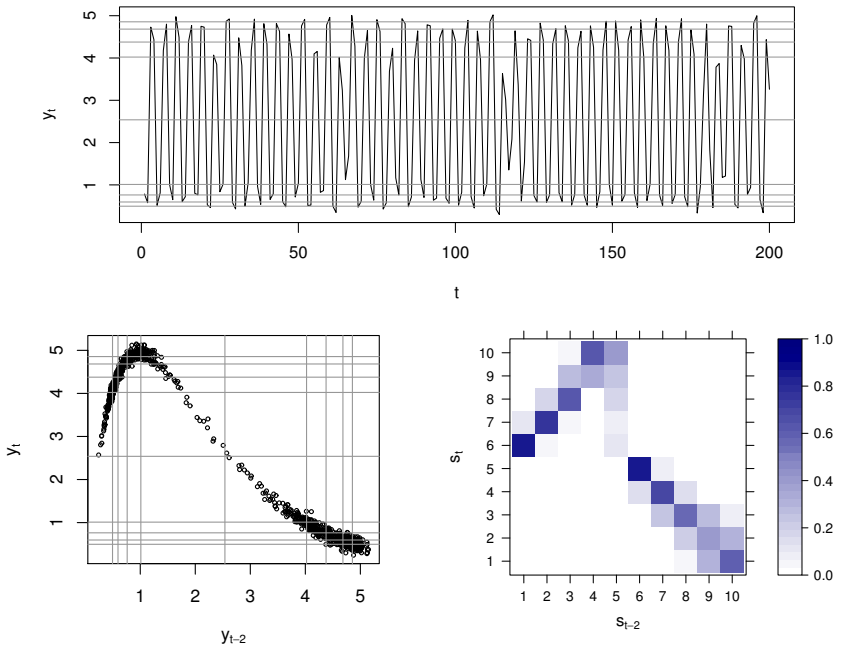


Figure 2.6: Time series, second-lag scatter plots, and approximate transition probability matrices for the $K = 5$ (upper) and $K = 10$ (lower) discretizations of $\{y_t\}$. Gray lines indicate cutoff values for state assignment in $\{s_t\}$.

only. This matrix was computed using the final 997,990 steps of the simulation after reserving 1,000 steps for model training and 10 steps for conditioning lags. Although unnecessarily large, a time series of one million steps assures numerical accuracy, allows for essentially uncorrelated validation observations, and was used to verify that key characteristics of the time series remain constant on long time scales. Due to discretization error, the time series $\{s_t\}$ is only approximately a Markov chain, whose distribution nevertheless depends primarily on the second lag.

2.3.3 Results

To understand the performance of the proposed model under different scenarios and configurations, models were fit to the simulated time series for varying series length $T \in \{50, 100, 500\}$, resolution of discretization (number of states) $K \in \{5, 10\}$, and prior combinations $p(\boldsymbol{\lambda}) \in \{\text{Dir}, \text{SDM}, \text{SBM}\}$, $p(\boldsymbol{Q}_{\cdot,k}) \in \{\text{Dir}, \text{SBM}\}$. The greatest number of lags considered was set to $L = 7$ for all models. Dirichlet priors for $\boldsymbol{\lambda}$ and columns of $\boldsymbol{Q}$ were relatively non-informative and symmetric with common shape parameter $1/L$ and $1/K$, respectively. The SDM prior for $\boldsymbol{\lambda}$ used the same shape parameter as the Dirichlet components, and added $\beta = T/4$ to strongly enforce sparsity. The SBM prior for $\boldsymbol{\lambda}$ extended the same Dirichlet prior with $\pi_1 = 0.75$, $\pi_3 = 0.10$, and $\eta = 1{,}000$, with sparsity adjustment for $\boldsymbol{\delta}$. Independent SBM priors also encouraged sparsity in $\boldsymbol{Q}$ with $\pi_1 = 0.9$ and $\pi_3 = 0.05$ for the $K = 5$ case, $\pi_1 = 0.75$ and $\pi_3 = 0.1$ for the $K = 10$ case, and $\eta = 1{,}000$, extending (with sparsity adjustment for $\boldsymbol{\delta}$) the Dirichlet priors for each of the k columns in $\boldsymbol{Q}$.

To compare competing models, we randomly sampled 2,000 observations (denoted t') from the more than 997,000 non-training steps after time T. Then

conditioning on L lags and model parameters at MCMC iteration i, we produced a model estimate of the discrete forecast distribution $p(s_{t'} \mid s_{t'-1}, \ldots, s_{t'-L}, \boldsymbol{\lambda}^{(i)}, \boldsymbol{Q}^{(i)})$ according to (2.9) and denoted $\hat{\boldsymbol{p}}_{t'}^{(i)}$. This forecast distribution was used to create a point estimate using the forecast expectation $\bar{p}_{t'}^{(i)} \equiv \sum_{k=1}^{K} k \cdot \hat{p}_{t',k}^{(i)}$, where $\hat{p}_{t',k}^{(i)}$ is the kth element of $\hat{\boldsymbol{p}}_{t'}^{(i)}$. Next, a squared-error loss was calculated as $\left(s_{t'} - \bar{p}_{t'}^{(i)}\right)^2$. This loss was averaged across the 2,000 validation times $\{t'\}$ and 2,000 randomly selected posterior samples $\{\boldsymbol{\lambda}^{(i)}, \boldsymbol{Q}^{(i)}\}$. We favor this squared-error metric because it utilizes information from the entire $\hat{\boldsymbol{p}}_{t'}^{(i)}$ vector, which is important in our applications where adjacent states are considered "closer" than non-adjacent states. Results for all combinations of simulation settings are reported in Table 2.1. The results were also verified with separate MCMC chains and different validation sets.

The most striking result from Table 2.1 is that in all simulation groups, the SDM/SBM prior combination is at or near the lowest mean forecast loss. The primary contributor to this gain in model accuracy is the SBM prior on columns of $\boldsymbol{Q}$. This is to be expected since the marginal transition map associated with the correct lag exhibits the least noise (see Figure 2.5). Adding the SDM or SBM prior on $\boldsymbol{\lambda}$ also improves accuracy in most cases, and enforcing sparsity in $\boldsymbol{Q}$ appears to assist further in the lag selection, as accuracy gains in enforcing sparsity in $\boldsymbol{\lambda}$ are often more pronounced when using the SBM prior in the $K = 10$ case.

Accuracy gains from these priors appear to be consistent across varying number of states, although they are far more pronounced in models with mores states, for which $\boldsymbol{Q}$ should be more sparse. The gains diminish as sample size increases, as expected. In every model fit, the $\boldsymbol{\lambda}$ vector heavily favors lag 2. Because of this, posterior estimates of $\boldsymbol{Q}$ resemble the validation estimate of $\boldsymbol{Q}$ in the right panel of Figure 2.6. Generally, the SDM and SBM priors on $\boldsymbol{\lambda}$ result in stronger support for lag 2 than their Dirichlet counterpart. In the $T = 50, 100$ cases, the SBM prior

			Sq.-error loss	
T	Prior $\boldsymbol{\lambda}$	Prior $\boldsymbol{Q}$	$K = 5$	$K = 10$
50	Dir	Dir	49.01	166.20
		SBM	45.27	122.67
	SDM	Dir	48.64	155.48
		SBM	**44.64**	**99.49**
	SBM	Dir	48.00	146.55
		SBM	44.68	100.04
100	Dir	Dir	39.11	80.87
		SBM	38.16	66.54
	SDM	Dir	38.90	78.95
		SBM	37.90	**61.90**
	SBM	Dir	38.60	76.72
		SBM	**37.82**	63.88
500	Dir	Dir	35.84	49.50
		SBM	35.83	48.55
	SDM	Dir	35.85	49.42
		SBM	35.83	48.46
	SBM	Dir	35.87	49.38
		SBM	35.84	**48.44**

Table 2.1: Results of the MTD model fit to the simulated dynamical system under various data and prior scenarios. The squared-error loss metric is reported as the mean across validation observations and MCMC iterations and multiplied by 100. Within groups, the lowest mean loss is highlighted with bold font.

on $\boldsymbol{Q}$ causes lags 6 and 3, and occasionally lag 5, to gather a small amount of posterior mass. While selection of lag 6 is consistent with the fact that marginal transition dynamics for lags 2 and 6 resemble one another in Figure 2.5, occasional selection of the odd lags is less intuitive and may be attributed to random noise with smaller sample sizes. Inferences for $\boldsymbol{\lambda}$ are nearly indistinguishable across models when the sample size is large ($T = 500$).

2.3.4 Chinook salmon data

The Chinook salmon data set contains 71 annual measurements of salmon abundance at the Coleman National Fish Hatchery in Anderson, California, U.S.A. from 1940 to 2010, compiled from Azat (2016) and personal correspondence. Time series and lag plots of the natural logarithm of abundance are given in Figure 2.7. Population dynamics for Chinook salmon follow a two to four-year life cycle

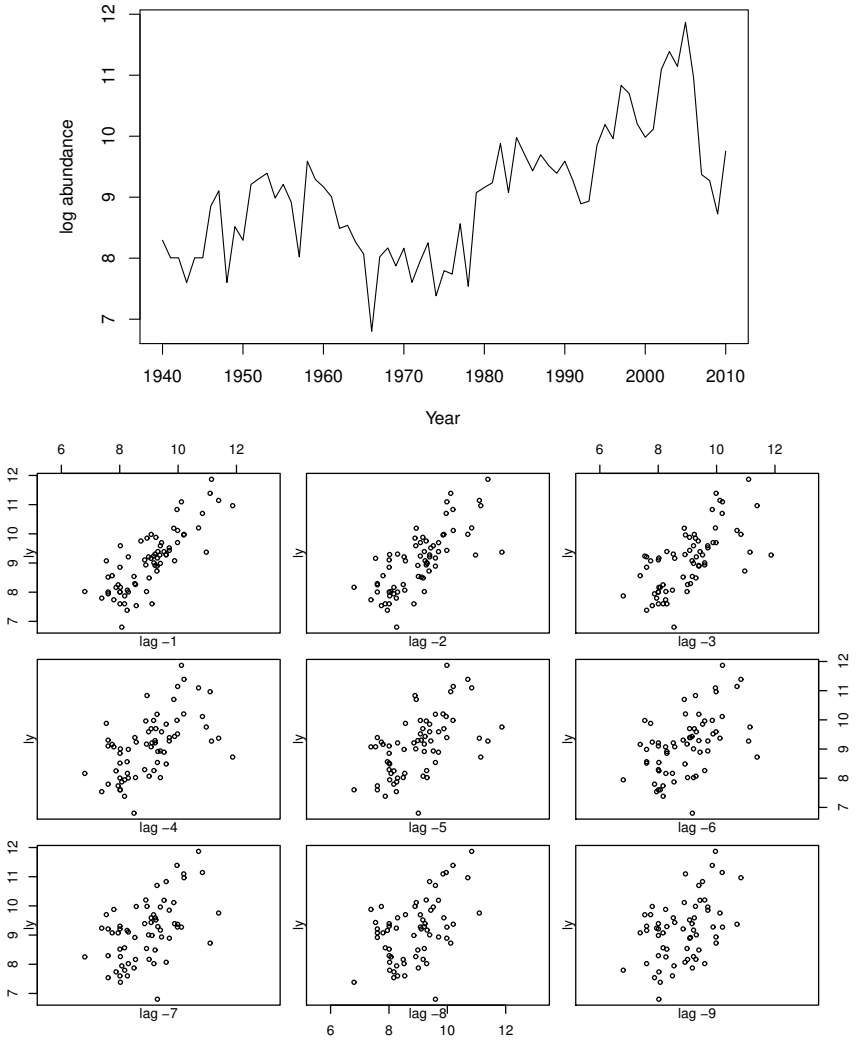


Figure 2.7: Time-series plot (above) and lag scatter plots (below) for the natural logarithm of Chinook salmon abundance from 1940 to 2010.

(Satterthwaite et al., 2017), and are commonly modeled using parametric functions similar to that of (2.11) (Quinn and Deriso, 1999, p. 89). Because North American oceanic salmon populations could be influenced by decadal-scale ocean dynamics, we consider up to $L = 10$ lags (Hare and Mantua, 2001). Conditioning on the first 10 observations and reserving the final 11 observations as a hold-out set for validation leaves 50 observations for training the models: years 1950 to 1999. After discretizing the data into sets of $K = 4$, 5, and 7 quantile-based bins using all 71 years, we fit the proposed models with the same sparsity-favoring prior settings used for the $K = 10$ run of the simulation study. Because discretization is based on quantiles, results are invariant to monotonic transformations such as the natural logarithm used for the plots in Figure 2.7.

Out-of-sample, one-step-ahead, mean forecast squared-error loss for years 2000 to 2010 is reported for the various model settings in Table 2.2. Again, we see best results from the combinations involving the new priors for finer resolution ($K = 5, 7$). In the $K = 4$ case, sparsity-favoring priors on $\boldsymbol{\lambda}$ hinder model performance. In the $K = 5$ case, the new priors hinder performance with the Dirichlet prior on $\boldsymbol{Q}$, but help when coupled with the SBM prior on $\boldsymbol{Q}$. As with

		Sq.-error loss		
Prior $\boldsymbol{\lambda}$	Prior $\boldsymbol{Q}$	$K = 4$	$K = 5$	$K = 7$
Dir	Dir	47.67	88.95	226.72
	SBM	**46.75**	94.55	222.08
SDM	Dir	59.82	100.19	189.19
	SBM	57.22	83.90	**155.81**
SBM	Dir	53.66	90.66	188.07
	SBM	49.73	**81.17**	162.98

Table 2.2: Results of the MTD model fit to the Chinook salmon data under different resolutions and prior scenarios. The squared-error loss metric is reported as the mean across validation observations and MCMC iterations and multiplied by 100. Within groups, the lowest mean loss is highlighted with bold font.

the simulation study, the largest improvement comes from replacing the Dirichlet priors on $\boldsymbol{Q}$ with the SBM.

Direct evaluation of the one-step-ahead forecast distributions in Figure 2.8 yields further insight and demonstrates the effects of the various prior configurations when $K = 7$. These distributions are given for the validation years 2000 to 2010 (again with each conditioning on the past $L = 10$ years as fixed and known) by shading the region corresponding to each state. This visualization elucidates the conclusions from Table 2.2. For example, the models employing sparse lag inference are more responsive to the switch in 2007, thus more appropriately forecasting

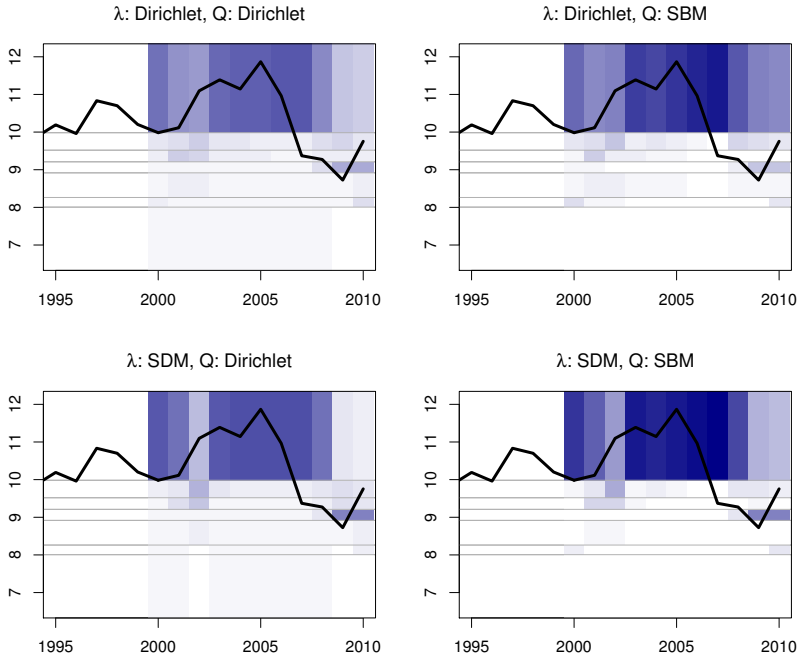


Figure 2.8: Time series of log-transformed Chinook salmon abundance with one-step-ahead forecast distributions on the holdout set of years 2000 to 2010, reported for four prior configurations. The plots for models with the SBM prior on $\boldsymbol{\lambda}$ are nearly indistinguishable from those using the SDM prior and are omitted. Shaded squares indicate posterior mean point forecast probabilities. The shading scale is similar to those of the transition matrices in Figure 2.6, with darker shades corresponding to higher probabilities. Cutoff values for the discretized states appear as horizontal lines.

2008-2010. As expected, the transition distributions resulting from the models with sparse priors produce higher estimated transition probabilities leading to more decisive forecasts.

We now turn to model inferences for $\boldsymbol{\lambda}$ and $\boldsymbol{Q}$. In the $K = 4$ case, the baseline Dirichlet/Dirichlet model fairly evenly distributes posterior mass among lags 1 through 3. Enforcing sparsity in $\boldsymbol{Q}$ only (the model supported by squared-error loss) favors lag 2 slightly more, as well as giving some support for lag 9. When sparsity is introduced into lag inference, lag 3 emerges as a preference across both priors for transition probabilities.

In the $K = 5$ case, the baseline Dirichlet/Dirichlet model mostly favors lag 1 and occasionally lag 2. Enforcing sparsity in $\boldsymbol{Q}$ spreads some of the posterior mass to lags 7 and 8. Adding sparsity in lags (the models supported by squared-error loss) returns most posterior mass to lag 1.

In the $K = 7$ case (summarized in Figure 2.9), the baseline Dirichlet/Dirichlet model fairly evenly distributes posterior mass between lags 1 and 2. Enforcing sparsity on lags only tends to favor lag 2 more. Enforcing sparsity in $\boldsymbol{Q}$ transfers some of that mass to lags 3, 5, and 8. Under the SDM prior for lags and SBM prior for transition probabilities (the model supported by squared-error loss), most posterior mass favors lag 2, with some minor support for lags 1, 5, and 8.

Across levels of resolution (K), we see that lag inferences tend to be affected by the prior on transition probabilities in addition to the prior on the lags themselves. Indeed, the SDM prior often accentuates preferences already evident, although it may assist in selecting a "winner." Consistent with the simulation study, the new priors excel in the models with finer levels of discretization. Overall, we see improved forecasting ability and potentially clearer lag effects afforded by the structured sparsity priors.

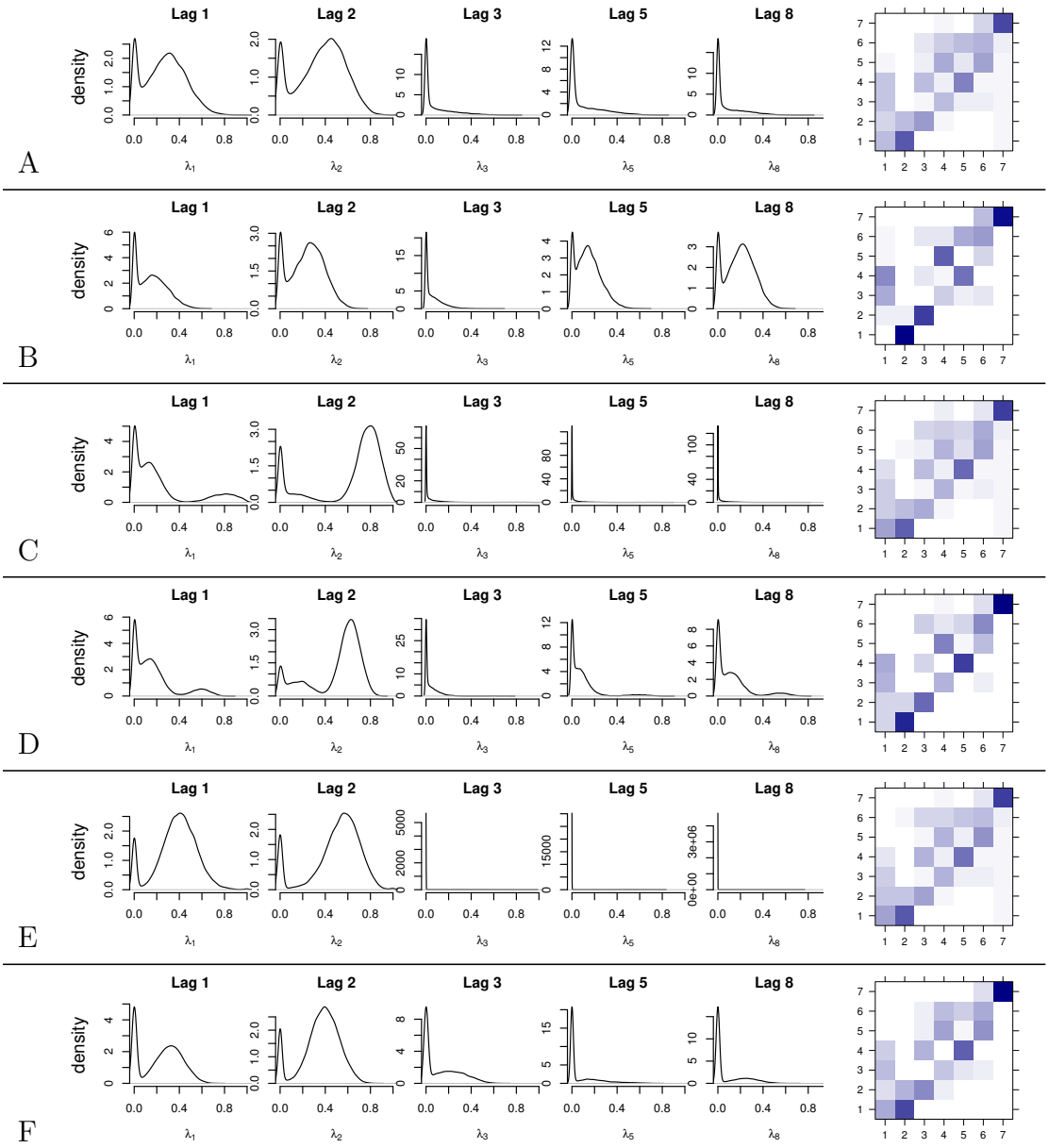


Figure 2.9: Marginal posterior density plots of selected λ_ℓ (left) and posterior mean point estimate of $\boldsymbol{Q}$ (right) for the Coleman Chinook salmon data with $K = 7$ for all prior settings: A-Dirichlet($\boldsymbol{\lambda}$)/Dirichlet($\boldsymbol{Q}$), B-Dirichlet/SBM, C-SDM/Dirichlet, D-SDM/SBM, E-SBM/Dirichlet, F-SBM/SBM. The shading scale and orientation are similar to those of the transition matrices in Figure 2.6.

2.4 Discussion

We have explored methods to model sparse probability vectors beyond the ubiquitous Dirichlet distribution, introducing two new prior models, the sparse Dirichlet mixture (SDM) and the stick-breaking mixture (SBM). We have demonstrated some properties of these models and illustrated their use in Markov chain models for time-series data. In the simulation study, the new models assisted in identifying the active lag and estimating sparse transition dynamics. In the salmon data analysis, we demonstrated how imposing structure on the model can potentially reveal additional insights into the lag dependence of a time series.

The proposed prior distributions may find utility with categorical data analyses in which there exist many categories and relatively few observations, and a modeler suspects that few categories have non-negligible probability. These methods are particularly useful for softly enforcing sparse structures in hierarchical models, levels removed from observed data. This can also be considered a model-based selection method. Indeed, the augmented MTD model in Section 2.3.1 motivated development of the new prior models as we sought to strengthen lag inferences.

As was evident in the salmon data analysis, the method of discretization has important implications for both lag and transition inferences. We have chosen to discretize by quantiles because they are invariant to monotonic transformation of the data, such as taking the logarithm. Furthermore, each state is observed with approximately equal frequency, providing comparable inferences across columns of the transition matrix as well as an approximately uniform stationary distribution, which helps justify conditioning the likelihood on the first L states. In light of our sparsity considerations, another potentially reasonable method of discretization would be to classify the points in states by clustering. This could be accomplished in data preprocessing or as part of the model, potentially with a hidden Markov

structure.

To broaden the scope of MTD models utilizing this prior structure, we consider various extensions. The model could be made more flexible if we consider mixing over a transition tensor of order higher than two and using sparsity priors to select tuples of lags, parsimoniously modeling higher-order Markov chains and simultaneously infering active lags. We pursue this in Chapter 3. As noted earlier, the proposed prior distributions can be utilized to encourage sparsity in any hierarchical model for which observations are allocated latent membership in a discrete set. Model settings beyond the ones considered here include mixture modeling and mixture deconvolution.

Chapter 3

Estimation and Selection for High-Order Markov Chains with Bayesian Mixture Transition Distribution Models

3.1 Introduction

Consider modeling a time series of nominal or ordinal values $s_t \in \{1, \ldots, K\}$ collected at equally spaced, discrete times $t = 1, \ldots, T$. A popular approach for capturing serial correlation is to assume Markovian dynamics: that the conditional probability distribution of s_t depends only on the recent past. Time homogeneity, or time invariance of the transition probabilities, is also typically assumed. These simplifying assumptions, nearly essential for inference in small or moderate sample size scenarios, are often appropriate even if the time series is not truly Markovian. Another common assumption is to condition only on the single most recent lag.

However, restricting a model to first-order dynamics, or even selecting the incorrect lag, can miss important features in the data. In this chapter, we propose Bayesian models to address two distinct objectives: estimation for the relevant time-delay coordinates, the Markovian order and important lags; and parsimonious modeling of high-order chains.

Assuming time homogeneity, a full, unrestricted first-order model requires estimation of K discrete distributions, each with $K - 1$ free parameters. A Markov chain of order L requires estimation of K^L such distributions, limiting consideration to low orders for most time series. Typically, order (or lag) is selected by maximizing a (possibly penalized) likelihood (Katz, 1981; Raftery, 1985; Prado and West, 2010), performing trans-dimensional MCMC (Green, 1995; Insua et al., 2012), using Bayes factors (Fan and Tsai, 1999; Bacallado, 2011; Zucchini and MacDonald, 2009), predictive criteria, or goodness-of-fit tests (Bartlett, 1951; Besag and Mondal, 2013). Each of these approaches requires either fitting multiple models or complex estimation methods. Our approach is to build lag inference into a single model.

Several approaches have been proposed to address exponential growth in the parameter space for higher-order transitions. Raftery (1985) introduced the mixture transition distribution (MTD), a general-purpose, parsimonious model for Markov chains. The MTD model was extended in Raftery and Tavaré (1994) and developed over the subsequent decade. Berchtold and Raftery (2002) provide a review. In the original MTD model, lags contribute to the transition probabilities by mixing over a single transition matrix. Only one new parameter is added for each additional lag. Despite its simplicity, the MTD framework can provide flexibility to capture nonstandard features, such as "outliers, bursts, and flat stretches," as demonstrated by Le et al. (1996) for a continuous-state version of the MTD.

Contemporary with the MTD model, generalized linear models for multinomial outcomes were applied to categorical time series (Liang and Zeger, 1986; Zeger and Liang, 1986; Fahrmeir and Kaufmann, 1987). These models can accommodate varying degrees of complexity by controlling the order of interactions among the linear predictors (lags), up to and including a full model with $K^L(K-1)$ parameters. These models can also account for exogenous sources of non-stationarity through covariates. However, estimation and interpretability become problematic in these models when many lags are considered.

Tree-based methods provide an alternative parsimonious approach. Variable-length Markov chains (VLMC, Ron et al., 1994; Bühlmann et al., 1999) reduce the parameter space by clustering the K^L transition distributions via recursive pruning. Sparse Markov chains (SMC, Jääskinen et al., 2014) partition the L-dimensional lag space without hierarchical constraints, resulting in greater flexibility. They also feature a prior structure which encourages low orders. Although efficient, these models lack posterior uncertainty quantification, and inferences for order and lag importance are not readily available.

More recently, Sarkar and Dunson (2016) proposed a Bayesian nonparametric model for high-order Markov chains. They model the K^L transition distributions through tensor factorization and further encourage parsimony by clustering the components of a core mixing distribution with a Dirichlet process prior (Ferguson, 1973). By allowing variable dimensions along different modes of the core mixing distribution, the model further admits inferences for lag importance. This model enjoys a fully Bayesian, albeit complicated, implementation and has been shown to perform well against the methods described above in forecasting in scenarios with up to four states and ten lags.

Our modeling strategy is to build on the simplicity and interpretability of the

MTD model. One popular extension of the MTD, referred to by Berchtold and Raftery (2002) as the MTDg model, utilizes a separate transition matrix for each lag. While this more flexible model grows linearly with each additional lag, it is not identifiable (Lèbre and Bourguignon, 2008). Recently, Tank et al. (2017) used a reparameterization to establish a unique and identifiable characterization of the MTDg model in the context of multiple time series. Using a penalized likelihood and proximal gradient optimization, they softly enforce the identifiability conditions and simultaneously select relevant series to infer Granger causality. We propose a Bayesian estimation approach to the MTDg model which utilizes the priors introduced in Chapter 2 to promote shrinkage toward the identifiability conditions of Tank et al. (2017), and to simultaneously select relevant lags. These priors were previously demonstrated to effectively select a single active lag using the original MTD model. We then propose an extension which allows for higher-order interaction between lags, as well as inference for the Markovian order (i.e., the number of active lags) up to a pre-specified maximum.

The remainder of this chapter is organized as follows. In Section 3.2, we review the MTD model and develop our proposed extensions. We outline our approach for Bayesian inference using structured priors to aid with the models' intended uses in Section 3.3. In Section 3.4, we test the models using two simulation scenarios that reflect our two objectives, demonstrating improved predictive performance over the original MTD. Section 3.5 illustrates the models through two analyses, first on a data set which appears in the preceding literature, and second on annual time series of pink salmon abundance in Alaska, U.S.A. Finally, we conclude with a summary in Section 3.6. Technical details are provided in Appendix B.

3.2 Models

In a full L-order, time-homogeneous Markov chain, the collection of all possible transition probabilities $\Pr(s_t = k_0 \mid s_{t-1} = k_1, \ldots, s_{t-L} = k_L)$, for $k_\ell \in \{1, \ldots, K\}$, $\ell \in \{1, \ldots, L\}$, $t \in \{L+1, \ldots, T\}$, can be arranged in a $(L+1)$-order tensor $(\boldsymbol{\Omega})_{k_0, k_1, \ldots, k_L}$. If we condition on the first L observations of the time series, the joint sampling distribution for the remaining sequence is given by $\Pr(\{s_t\}_{t=L+1}^{T} \mid \{s_t\}_{t=1}^{L}, \boldsymbol{\Omega}) = \prod_{t=L+1}^{T} (\boldsymbol{\Omega})_{s_t, s_{t-1}, \ldots, s_{t-L}}$, defining the conditional likelihood that we employ hereafter. We begin by specifying the original MTD model in Section 3.2.1 and motivate its extensions. In Section 3.2.2, we discuss the MTDg extension and associated identifiability results. We then introduce a Bayesian formulation for the MTDg which uses priors for sparse probability vectors in Section 3.2.3. Finally, we propose an extension to include higher-order transitions in Section 3.2.4.

3.2.1 Original mixture transition distribution

The mixture transition distribution model constructs the transition probability tensor $\boldsymbol{\Omega}$ as linear combinations of probabilities from a single column-stochastic matrix $\boldsymbol{Q}$ and adds just one parameter for each additional lag (the mixing weights, $\{\lambda_\ell\}$), similar to autoregressive models. The transition probabilities in a model of order L are given as

$$\Pr(s_t = k_0 \mid s_{t-1} = k_1, \ldots, s_{t-L} = k_L) = (\boldsymbol{\Omega})_{k_0, k_1, \ldots, k_L} = \sum_{\ell=1}^{L} \lambda_\ell \, q_{k_0, k_\ell} \,, \qquad (3.1)$$

where $q_{i,h} \equiv (\boldsymbol{Q})_{i,h}$, $0 \leq \lambda_\ell \leq 1$ and $\sum_{\ell=1}^{L} \lambda_\ell = 1$. Although the MTD model incorporates information beyond the first lag, it is restrictive in that it cannot capture nonlinear (non-additive) dynamics in more than one dimension of the lag space.

The construction in (3.1) is reminiscent of tensor factorization methods, such as the Tucker decomposition (Tucker, 1966). Yang and Dunson (2016) and Sarkar and Dunson (2016) apply a similar decomposition to probability tensors, which in the context of Markov chains, have the form

$$(\boldsymbol{\Omega})_{k_0,k_1,\ldots,k_L} = \sum_{h_1=1}^{H_1} \cdots \sum_{h_L=1}^{H_L} \omega_{k_0,h_1,\ldots,h_L} \prod_{\ell=1}^{L} \rho^{(\ell)}_{h_\ell,k_\ell}\,, \tag{3.2}$$

where all ω and ρ variates are between 0 and 1 and sum to unity over the first index. The distinction between (3.1) and (3.2) is best understood in terms of the dependence structure with latent indicator variables (illustrated for the MTD in Figure 1.1). The $\{\omega\}$ comprise a core tensor of transition probabilities that are selected exclusively with a set of L latent indicators, which in turn depend on lagged observations through the $\{\rho\}$ probabilities. The MTD analogue of the core tensor, $\boldsymbol{Q}$, provides transition probabilities over a single mode that is selected with both the lagged observations and a single latent indicator, with probabilities $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_L)$. Using multiple nodes of varying dimension (i.e., $H_1, \ldots, H_L$) and weights that depend on all lags, the model in (3.2) offers interactions between lags and greater coverage at the cost of additional complexity.

As noted earlier, we use the MTD structure for its relative simplicity, parsimony, and interpretability. Recall from Section 2.3.1 that the mixing weights in the MTD model can help identify lag dependence. Specifically, $\lambda_\ell = 0$ is sufficient for conditional independence of s_t and $s_{t-\ell}$. If the columns of $\boldsymbol{Q}$ are unique, then $\lambda_\ell = 0$ is also a necessary condition for conditional independence. As in Chapter 2, we use inferences on $\boldsymbol{\lambda}$ for insight into lag importance.

3.2.2 MTDg, identifiability, and lag selection

The MTDg model modifies (3.1) by using a distinct column-stochastic matrix $\boldsymbol{Q}^{(\ell)}$ for each lag $\ell = 1, \ldots, L$. While this increases flexibility and allows for different transition types associated with each lag, the model lacks identifiability. Tank et al. (2017) demonstrate this by introducing an intercept probability vector, $\boldsymbol{Q}^{(0)} = \left(q_1^{(0)}, \ldots, q_K^{(0)}\right)$, extending $\boldsymbol{\lambda}$ to include λ_0, and reparameterizing the transition probabilities through the products $\varphi_{k_0}^{(0)} \equiv \lambda_0 \, q_{k_0}^{(0)}$ and $\varphi_{k_0,k_\ell}^{(\ell)} \equiv \lambda_\ell \, q_{k_0,k_\ell}^{(\ell)}$, resulting in the MTDg formulation

$$\Pr(s_t = k_0 \mid s_{t-1} = k_1, \ldots, s_{t-L} = k_L) = \varphi_{k_0}^{(0)} + \sum_{\ell=1}^{L} \varphi_{k_0,k_\ell}^{(\ell)} \, . \tag{3.3}$$

One can then freely transfer probability mass by subtracting some vector $\boldsymbol{a}_\ell = \left(a_1^{(\ell)}, \ldots, a_K^{(\ell)}\right)$ from each column of $\boldsymbol{\varphi}^{(\ell)} \equiv \lambda_\ell \, \boldsymbol{Q}^{(\ell)}$ and adding it to $\boldsymbol{\varphi}^{(0)}$ while preserving all values in $\boldsymbol{\Omega}$. Selecting $a_k^{(\ell)}$ to be the minimum value in the kth row of $\boldsymbol{\varphi}^{(\ell)}$ for each $k = 1, \ldots, K$, and following the transferral procedure just described for $\ell = 1, \ldots, L$, results in a maximally reduced parameterization in the sense that the highest probability mass possible has been transferred to the intercept while maintaining non-negativity of all elements in $\boldsymbol{\varphi}^{(\ell)}$. Let $\{\tilde{\boldsymbol{\varphi}}^{(\ell)}\}_{\ell=0}^{L}$ denote the resulting parameters after the reduction procedure so that $\tilde{\boldsymbol{\varphi}}^{(0)} \equiv \boldsymbol{\varphi}^{(0)} + \sum_{\ell=1}^{L} \boldsymbol{a}_\ell$ and $\tilde{\varphi}_{i,j}^{(\ell)} \equiv \varphi_{i,j}^{(\ell)} - a_i^{(\ell)}$, for $i = 1, \ldots, K$, $j = 1, \ldots, K$, and $\ell = 1, \ldots, L$. Tank et al. (2017) show that this maximal reduction yields a unique representation for every MTDg. Furthermore, one can view the reduced model in the original parameterization using $\tilde{\boldsymbol{\lambda}} = (\tilde{\lambda}_0, \ldots, \tilde{\lambda}_L)$ with $\tilde{\lambda}_0 \equiv \sum_{k=1}^{K} \tilde{\varphi}_k^{(0)}$, and $\tilde{\lambda}_\ell \equiv \sum_{k=1}^{K} \tilde{\varphi}_{k,j}^{(\ell)} = \lambda_\ell - \sum_{k=1}^{K} a_k^{(\ell)}$, for $\ell = 1, \ldots, L$, and invariant to choice of j; probability vector $\tilde{\boldsymbol{Q}}^{(0)} \equiv (\tilde{\lambda}_0)^{-1} \, \tilde{\boldsymbol{\varphi}}^{(0)}$; and column-stochastic matrices $\tilde{\boldsymbol{Q}}^{(\ell)} \equiv (\tilde{\lambda}_\ell)^{-1} \, \tilde{\boldsymbol{\varphi}}^{(\ell)}$. Then $\tilde{\lambda}_\ell$ is interpretable as a marginal contribution of the ℓth lag to the transition distribution. Thus, with

careful construction, we may use $\tilde{\boldsymbol{\lambda}}$ to make inferences about lag importance even though the MTDg is overparameterized. Furthermore, the intercept allows us to infer the possible lack of serial dependence in a direct way.

To operationalize this reduction in an estimation procedure, Tank et al. (2017) show that solutions to a penalized likelihood based on (3.3), in which the penalty increases with respect to the absolute value of the entries in the $\{\boldsymbol{\varphi}^{(\ell)}\}_{\ell=1}^{L}$ (excluding the intercept), meet the maximal-reduction criterion. Their proposed soft penalty functions equivalently regularize λ_ℓ for $\ell = 1, \ldots, L$. Because $\lambda_\ell = 0$ is sufficient and necessary (as long as the columns of $\boldsymbol{Q}^{(\ell)}$ are distinct) for conditional independence of the current state from lag ℓ, the penalized estimate simultaneously detects lag relevance (or Granger causality in the case where ℓ indexes multiple time series).

3.2.3 Bayesian MTDg with priors for sparse probability vectors

We now present a Bayesian modeling approach to the MTDg, which admits full characterization of uncertainty. Rather than addressing the constraint that each column of $\boldsymbol{\varphi}^{(\ell)}$ sum to λ_ℓ, we work with the original $\boldsymbol{\lambda}$ and $\{\boldsymbol{Q}^{(\ell)}\}$ parameters and employ carefully chosen prior distributions that shrink toward the identifiable and interpretable $\tilde{\boldsymbol{\lambda}}$ and $\{\tilde{\boldsymbol{Q}}^{(\ell)}\}$. In this and the following section, we employ the sparse Dirichlet mixture (SDM) and stick-breaking mixture (SBM) priors, as described in Section 2.2, which go beyond the standard Dirichlet prior by enforcing sparsity in the presence of data, as well as conditional stochastic ordering. Both priors are continuous, bypassing problems that arise from the sum-to-one constraint when using priors with point masses.

As a one-parameter extension of the Dirichlet distribution, the SDM prior is a fixed-weight mixture of Dirichlet distributions, with each component featuring a

boost of equivalent sample size $\beta > 1$ in one of the categories. It can be described as a "winner-takes-all" prior in that it shifts mass toward the element with largest shape parameter. The SBM prior uses (2.4) to construct a length-J probability vector with sequentially drawn latent variables indexed by $j = 1, \ldots, J-1$, each from a mixture of three beta distributions: $\pi_1 \operatorname{Beta}(1, \eta) + \pi_2 \operatorname{Beta}(\gamma_j, \delta_j) + \pi_3 \operatorname{Beta}(\eta, 1)$, where $\pi_1 + \pi_3 < 1$, $\pi_2 = 1 - \pi_1 - \pi_3$, and η is large. To accommodate our proposed extensions of the MTD model, we adapt the SBM prior to allow the mixture weights, π_1, π_2, and π_3, to vary with j.

One important advantage of the SDM and SBM priors is their conjugacy and resulting computational tractability. If the hyperparameters of the priors are fixed, as is usually the case with Dirichlet priors, incorporating them into a hierarchical model involving multinomial counts (latent or observed) requires minimal effort since posterior Gibbs sampling proceeds with conditional distributions that can be directly sampled, allowing us to swap priors without structural changes to the updates in Appendix B.2.

If a modeler believes that exactly one lag influences the transition distribution, the SDM prior can be used in the single-$\boldsymbol{Q}$ MTD model, as demonstrated in Section 2.3. However, this is not recommended for the MTDg model. If β is not sufficiently high, the SDM prior may distribute posterior mass to a non-unique and non-interpretable configuration of $\boldsymbol{\lambda}$. We instead use a SBM prior for $\boldsymbol{\lambda}$ that favors the unique reduction and more appropriately allows for dependence on multiple lags. Here, the stick-breaking construction of the SBM provides intuition, as λ_0 is drawn first, and the rest of $\boldsymbol{\lambda}$ is broken sequentially from what remains in the unbroken stick. To avoid penalizing the intercept, we set $\pi_2 = 1$ for λ_0 only and use either $\gamma_0 = \delta_0 = 1$ (the uniform distribution) or beta shape parameters that favor large values of λ_0. The remaining beta mixtures use $\pi_1 > 0$ and $\pi_3 > 0$ to

regularize $\{\lambda_\ell\}_{\ell=1}^{L}$. Setting $\pi_1 > 0$ allows for small values of the corresponding λ_ℓ, effectively skipping the ℓth lag. Setting $\pi_3 > 0$ promotes consumption of the remaining mass before reaching λ_L. If $\gamma_\ell = \gamma$ and $\delta_\ell = \delta$ across ℓ and π_2 is relatively high, the sequential SBM construction can further regularize $\boldsymbol{\lambda}$ via stochastic ordering, consistent with the common assumption that recent lags should carry more influence.

The model is completed with prior distributions for $\{\boldsymbol{Q}^{(\ell)}\}$. The traditional choice for transition matrices is to use independent Dirichlet distributions for each column, which we adopt here. In Section 2.3, we found it advantageous to use independent SBM priors for each column of $\boldsymbol{Q}$ (in the standard MTD model) in cases of nearly deterministic dynamics. However, the MTDg model spreads estimation across multiple $\boldsymbol{Q}^{(\ell)}$ matrices, relying more heavily on the non-symmetric SBM prior. This can potentially introduce undesired artifacts to the estimated transition probabilities.

The full hierarchical model specification for the MTDg and details for posterior inference are discussed in Section 3.3 and Appendix B.2.

3.2.4 Mixtures of higher-order MTD components

The MTD and MTDg models offer parsimonious and interpretable representations for Markov chains with dependence extending beyond the most recent lag. However, these models are strictly additive in the sense that any dynamics of order higher than one (i.e., more than one active lag) are approximated with linear combinations of first-order transitions. In their survey of generalizations for the MTD, Berchtold and Raftery (2002) suggest, but do not pursue, the possibility of mixing over higher-order transition tensors. We build a Bayesian framework for such an extension to include higher-order "interactions," and we refer to the

resulting model as the mixture of mixture transition distributions (MMTD) model.

To define the MMTD model, let $R < L$ be a positive integer representing the highest-order transition tensor over which we will mix. Thus we have $\boldsymbol{\mathcal{Q}}^{(0)}$, a length-$K$ probability vector; $\boldsymbol{\mathcal{Q}}^{(1)}$, a $K \times K$ transition matrix; $\boldsymbol{\mathcal{Q}}^{(2)}$, a $K \times K \times K$ transition tensor; and so forth up to $\boldsymbol{\mathcal{Q}}^{(R)}$, a K^{R+1} transition tensor, such that $\sum_{k=1}^{K}(\boldsymbol{\mathcal{Q}}^{(R)})_{k,k_1,\ldots,k_R} = 1$ for all $(k_1, k_2, \ldots, k_R) \in \{1, \ldots, K\}^R$. Next, introduce a mixing probability vector across orders, $\boldsymbol{\Lambda} = (\Lambda_0, \Lambda_1, \ldots, \Lambda_R)$. The MMTD model for transition probabilities is then given by

$$
\begin{aligned}
\Pr(s_t = k_0 \mid s_{t-1} = k_1, \ldots, s_{t-L} = k_L) &= (\boldsymbol{\Omega})_{k_0,k_1,\ldots,k_L} \\
&= \Lambda_0 \, (\boldsymbol{\mathcal{Q}}^{(0)})_{k_0} + \Lambda_1 \sum_{\ell=1}^{L} \lambda_{\ell}^{(1)} \, (\boldsymbol{\mathcal{Q}}^{(1)})_{k_0,k_\ell} + \\
&\qquad + \Lambda_2 \sum_{1 \le \ell_1 < \ell_2 \le L} \sum \lambda_{(\ell_1,\ell_2)}^{(2)} \, (\boldsymbol{\mathcal{Q}}^{(2)})_{k_0,k_{\ell_1},k_{\ell_2}} + \ldots \\
&\qquad \ldots \; + \Lambda_R \sum_{1 \le \ell_1 < \ldots < \ell_R \le L} \cdots \sum \lambda_{(\ell_1,\ldots,\ell_R)}^{(R)} \, (\boldsymbol{\mathcal{Q}}^{(R)})_{k_0,k_{\ell_1},\ldots,k_{\ell_R}} ,
\end{aligned}
\tag{3.4}
$$

where $\boldsymbol{\lambda}^{(r)}$ is a probability vector of length $\binom{L}{r}$ for $r = 1, \ldots, R$. This mixture of mixtures is equivalent to using a single (albeit long) $\boldsymbol{\lambda}$ probability vector to mix over all possible arrangements of lags and base transition tensors $\boldsymbol{\mathcal{Q}}^{(r)}$. However, this parameterization is more informative about important orders (via inference for $\boldsymbol{\Lambda}$) in addition to lags (via inference for $\boldsymbol{\lambda}^{(r)}$). If $\Lambda_1 = 1$, we recover the original MTD model. The fully-parameterized transitions associated with $\boldsymbol{\mathcal{Q}}^{(r)}$ allow unrestricted dynamics in r dimensions of the lag space. As a discrete mixture of probability distributions, this model produces a valid probability tensor.

The model in (3.4) is clearly over-parameterized, and consequently $\boldsymbol{\Lambda}$, $\{\boldsymbol{\lambda}^{(r)}\}_{r=1}^{R}$, and $\{\boldsymbol{\mathcal{Q}}^{(r)}\}_{r=0}^{R}$ are not fully identified. Defining a reduction procedure similar to that of Tank et al. (2017) for the MMTD is more nuanced. One complication

arises because the result is dependent on the order of reduction. For example, one may first transfer probability mass to the intercept from all higher-order transition tensors, followed by transfer to the first-order transitions by defining elements of the $\boldsymbol{a}$ vector as the minima over indexes in $\boldsymbol{\mathcal{Q}}^{(r)}$, for $r \geq 2$, which correspond to a unique value of the lagged state for the lag associated with the current $\boldsymbol{\mathcal{Q}}^{(1)}$ (allowing for L such matrices). However, transferring first to the $\boldsymbol{\mathcal{Q}}^{(1)}$ associated with lag 1 and then to the $\boldsymbol{\mathcal{Q}}^{(1)}$ associated with lag 2 does not yield the same result if we reverse the order. Alternatively, one could define a reduction process in terms of projecting $\boldsymbol{\Omega}$ first onto an intercept, then projecting what remains onto the space spanned by the first-order level of the MMTD, and so forth. Thus, the intercept has the first opportunity to describe the base probabilities, then the first-order level of the model has the next opportunity to capture first-order dynamics, and each additional level fills in what lower-order levels cannot adequately model. Absent a formal procedure, we note that in estimation, this process would be implemented with regularization, for which the sequential SBM prior is well-suited. We therefore propose using a SBM prior for $\boldsymbol{\Lambda}$, similar to the one used in Section 3.2.3 for the MTDg model.

Even under regularization of model order, it is possible for a high-order $\boldsymbol{\mathcal{Q}}^{(r)}$ to mimic a lower-order tensor through repetition of transition probabilities across values of a certain lag. Rather than build complicated constraints into the model, we note that this issue can be detectable through inferences for $\{\boldsymbol{\mathcal{Q}}^{(r)}\}_{r=1}^{R}$. Specifically, a modeler may plot posterior estimates of the distributions in $\{\boldsymbol{\mathcal{Q}}^{(r)}\}_{r=1}^{R}$ and check for repeating patterns, especially patterns that coincide with a slice of the tensor (e.g., all columns equal in $\boldsymbol{\mathcal{Q}}^{(1)}$ in a $R = 1$ model would indicate that the intercept alone is adequate). We strongly recommend following this practice before interpreting model inferences for $\boldsymbol{\Lambda}$ and $\{\boldsymbol{\lambda}^{(r)}\}_{r=1}^{R}$.

We envision two primary uses for the MMTD model. The first is to uncover low-order structure from data whose practical lag dependence horizon is truly smaller than our over-specified L and the order is less than or equal to our selected R, in which case the true model is contained within the mixture framework. For example, we might postulate that a time series has second-order dependence, but we are unsure which two lags are important. Assuming a maximal lag horizon of 10, we could fit the MMTD model with $L = 10$ and $R \geq 2$. Because there is only one $\boldsymbol{\mathcal{Q}}^{(r)}$ at each level, we could use the SDM prior on each $\boldsymbol{\lambda}^{(r)}$ with a large value of β to select the appropriate lag configuration and discourage mixing lower-order transitions. If the dynamics are truly second-order, we would anticipate that Λ_2 would carry substantial posterior weight, and that inferences for $\boldsymbol{\lambda}^{(2)}$ would identify the influential lags. In this model-selection scenario, the SDM (on $\boldsymbol{\lambda}$) and SBM (on $\boldsymbol{\Lambda}$) priors play an important role in selection and interpretation.

If the true order of dependence in the time series is greater than R, our second intended use for the MMTD model is analogous to that of the MTD and MTDg, wherein we parsimoniously approximate higher-order dependence by mixing lower-order transition distributions. Adding the higher-order $\boldsymbol{\mathcal{Q}}^{(r)}$ tensors could be thought of as including interaction-like terms in the mixture. In this scenario, one may still use the SDM prior for each $\boldsymbol{\lambda}^{(r)}$, but with a lower value of β to encourage more mixing (note that $\beta = 1$ yields a Dirichlet prior). The SBM prior on $\boldsymbol{\Lambda}$ further allows mixing across orders, so that different levels of the model may mix across non-overlapping sets of lags.

As with the MTDg, we recommend using independent Dirichlet priors for each of the K^r probability distributions in $\boldsymbol{\mathcal{Q}}^{(r)}$ for $r = 0, 1, \ldots, R$. If L and R are small, T is large, and the transition probability tensor is known to be sparse, it *may potentially* be advantageous to replace these Dirichlet priors with independent

SBM priors. However, we strongly urge caution, as information from the data is spread thin and the SBM prior is not symmetric. Section 2.2.2 discusses a strategy for promoting more Dirichlet-like behavior in the SBM prior.

Our proposed model formulation requires estimation of R free Λ parameters, $\binom{L}{1}+\binom{L}{2}+\ldots+\binom{L}{R}-R$ free λ parameters, and $(K-1)+K(K-1)+K^2(K-1)+\ldots+K^R(K-1)=K^{R+1}-1$ free parameters in $\{\boldsymbol{\mathcal{Q}}^{(r)}\}_{r=0}^{R}$. The fastest-growing term in the λ parameter count increases no faster than a polynomial in L of degree $\lfloor R/2 \rfloor$ divided by $R!$, while the transition distributions grow exponentially. Table 3.1 reports the total number of parameters to estimate for different combinations of K, L, and R. Typically, K is fixed and known, and a modeler must select L and R considering parsimony, estimability for a given sample size, and computational cost. If R is much smaller than L, the MMTD substantially reduces the parameter space from the original full-order Markov chain. The parameter space is effectively

K	L	R	Λ	λ	$\mathcal{Q}$	Total	Unrestricted
2	5	2	2	13	7	22	32
2	5	4	4	26	31	61	32
2	10	2	2	53	7	62	1,024
2	10	4	4	381	31	416	1,024
5	5	2	2	13	124	139	12,500
5	5	4	4	26	3,124	3,154	12,500
5	10	2	2	53	124	179	3.91×10^7
5	10	4	4	381	3,124	3,509	3.91×10^7
7	5	2	2	13	342	357	100,842
7	5	4	4	26	16,806	16,836	100,842
7	10	2	2	53	342	397	1.69×10^9
7	10	4	4	381	16,806	17,191	1.69×10^9

Table 3.1: Free parameter count for MMTD model under different combinations of state-space size K, largest possible lag L, and largest mixing order R. The total number of parameters is the sum of the free Λ, λ, and $\mathcal{Q}$ parameters. The unrestricted total is the number of parameters required to estimate an unrestricted transition probability tensor of order L.

further reduced by the sparsity-inducing priors on $\boldsymbol{\Lambda}$ and $\{\boldsymbol{\lambda}^{(r)}\}$.

The hierarchical model specification for the MMTD and posterior inference details are discussed in Section 3.3 and Appendix B.3.

3.3 Bayesian inference and computation

We now address implementation of the MTDg and MMTD models. To obtain full posterior inference, we utilize a Gibbs sampler which alternates between collapsed and full conditional distributions made tractable by augmentation with latent configuration variables (Insua et al., 2012). As noted earlier, all inferences condition on the first L observations in the time series $\{s_t\}_{t=1}^T$.

3.3.1 MTDg model

The sampling distribution for the MTDg model is

$$p(\{s_t\}_{t=L+1}^T \mid \boldsymbol{\lambda}, \{\boldsymbol{Q}^{(\ell)}\}_{\ell=0}^L, \{s_t\}_{t=1}^L) = \prod_{t=L+1}^{T} \left[\lambda_0\, q_{s_t}^{(0)} + \sum_{\ell=1}^{L} \lambda_\ell\, q_{s_t,s_{t-\ell}}^{(\ell)}\right]. \tag{3.5}$$

We first break the mixture in (3.5) by introducing latent indicators z_t such that $\Pr(z_t = \ell \mid \boldsymbol{\lambda}) = \lambda_\ell$ for $\ell = 0, 1, \ldots, L$ independently across t. Adding the priors yields the full hierarchical model. For $t = L+1, \ldots, T$; $k = 1, \ldots, K$; $i = 1, \ldots, K$; and $\ell = 0, \ldots, L$, we have

$$\begin{aligned}
&\boldsymbol{Q}^{(0)} \sim \mathrm{Dir}(\boldsymbol{\alpha}^{(0)}), \quad (\boldsymbol{Q}^{(\ell)})_{\cdot,i} \overset{\text{ind.}}{\sim} \mathrm{Dir}(\boldsymbol{\alpha}_i^{(\ell)}) \text{ for } \ell > 0, \quad \boldsymbol{\lambda} \sim \mathrm{SBM}(\boldsymbol{\pi}_1, \boldsymbol{\pi}_3, \eta, \boldsymbol{\gamma}, \boldsymbol{\delta}), \\
&\Pr(z_t = \ell \mid \boldsymbol{\lambda}) = \lambda_\ell\,, \\
&\Pr(s_t = k \mid z_t = \ell, \{s_{t'}\}_{t'=t-L}^{t-1}, \boldsymbol{Q}^{(\ell)}) = q_k^{(0)} 1_{(\ell=0)} + q_{k,s_{t-\ell}}^{(\ell)} 1_{(\ell>0)}\,,
\end{aligned} \tag{3.6}$$

where each $\boldsymbol{\alpha}$ is a length-K vector of positive shape parameters; $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_3$ are length-L vectors containing probabilities such that $(\boldsymbol{\pi}_1)_\ell + (\boldsymbol{\pi}_3)_\ell < 1$; and $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ are length-L vectors containing positive shape parameters. We always set $(\boldsymbol{\pi}_1)_0 = 0$ to avoid penalizing λ_0, and also recommend setting $(\boldsymbol{\pi}_3)_0 = 0$.

This structure admits closed-form full conditional distributions for all of $\{z_t\}$, $\{\boldsymbol{Q}^{(\ell)}\}$, and $\boldsymbol{\lambda}$. Specifically, the update for $\boldsymbol{\lambda}$ is a conjugate SBM-multinomial update using aggregated counts of $\{z_t\}$. Each z_t can be updated with a discrete distribution involving $\boldsymbol{\lambda}$ and elements of $\{\boldsymbol{Q}^{(\ell)}\}$ as they appear in the likelihood. Given $\{z_t\}$, we can aggregate the transition counts into sufficient statistics $\boldsymbol{N}^{(0)} = (n_1^{(0)}, \ldots, n_K^{(0)})$ and $\{\boldsymbol{N}^{(\ell)}\}_{\ell=1}^L$, a set of $K \times K$ matrices. For example, if $s_t = 1$, $z_t = 2$, and $s_{t-2} = 3$, we would increment $(\boldsymbol{N}^{(2)})_{1,3}$. The full conditional distribution for the intercept $\boldsymbol{Q}^{(0)}$ is then an updated Dirichlet distribution with $\boldsymbol{N}^{(0)}$ providing the multinomial counts. Likewise, the update for $(\boldsymbol{Q}^{(\ell)})_{\cdot,i}$ involves a conjugate Dirichlet-multinomial update with its corresponding count vector $(\boldsymbol{N}^{(\ell)})_{\cdot,i}$, for each $\ell = 1, \ldots, L$ and $i = 1, \ldots, K$.

As is common with mixture models, the full joint posterior distribution is multimodal and the Gibbs sampler described above is prone to poor mixing. To improve mixing, we modify the Gibbs sampler just described in two ways. First, we integrate all $\{\boldsymbol{Q}^{(\ell)}\}_{\ell=0}^L$ parameters out of the full joint posterior. This affects only the full conditional distributions for the configuration variables $\{z_t\}$, which are drawn from a (different) discrete distribution. The second modification is an occasional (every 10 iterations) hybrid Metropolis step that jointly proposes $\boldsymbol{\lambda}$ and $\{z_t\}$ from the prior in order to encourage exploration. Ordinarily, the prior is inefficient as a proposal distribution. Although the sparse configurations proposed by the SBM prior help mitigate this issue, we still advocate running multiple long MCMC chains to ensure adequate mixing. Full details for the modified Gibbs

sampler are provided in Appendix B.2.

3.3.2 MMTD model

Our implementation for the MMTD model is analogous to the MTDg model, with a few notable extensions. As before, the sampling distribution for the time series is given as a product of transition probabilities in (3.4) across $t = L+1, \ldots, T$. We again break the mixture in (3.4) by introducing latent configuration variables Z_t such that $\Pr(Z_t = r \mid \boldsymbol{\Lambda}) = \Lambda_r$, for $r = 0, 1, \ldots, R$, independently for each observation time. Then conditional on Z_t (and for $Z_t > 0$), further introduce $\boldsymbol{z}_t$ such that $\Pr(\boldsymbol{z}_t = (\ell_1, \ldots, \ell_r) \mid Z_t = r, \boldsymbol{\lambda}^{(r)}) = \lambda^{(r)}_{(\ell_1,\ldots,\ell_r)}$, for $1 \leq \ell_1 < \ldots < \ell_r \leq L$, independently for each observation time. The hierarchical formulation for this model is given in generative order as follows. For $t = L+1, \ldots, T$; $k = 1, \ldots, K$; $k_\ell = 1, \ldots, K$; $\ell = 1, \ldots, L$, $1 \leq \ell_1 < \ldots < \ell_r \leq L$; and $r = 0, 1, \ldots, R$, we have

$$
\begin{aligned}
&\boldsymbol{\mathcal{Q}}^{(0)} \sim \mathrm{Dir}(\boldsymbol{\alpha}_{Q^{(0)}}), \quad (\boldsymbol{\mathcal{Q}}^{(r)})_{\cdot,k_1,\ldots,k_r} \overset{\text{ind.}}{\sim} \mathrm{Dir}(\boldsymbol{\alpha}_{Q^{(r)}}), \text{ for } (k_1, \ldots, k_r) \in \{1, \ldots, K\}^r\,, \\
&\boldsymbol{\Lambda} \sim \mathrm{SBM}(\boldsymbol{\pi}_1, \boldsymbol{\pi}_3, \eta, \boldsymbol{\gamma}, \boldsymbol{\delta})\,, \qquad \boldsymbol{\lambda}^{(r)} \overset{\text{ind.}}{\sim} \mathrm{SDM}(\boldsymbol{\alpha}_{\lambda^{(r)}}, \beta_{\lambda^{(r)}}), \text{ for } r = 1, \ldots, R\,, \\
&\Pr(Z_t = r \mid \boldsymbol{\Lambda}) = \Lambda_r\,, \qquad \Pr(\boldsymbol{z}_t = (\ell_1, \ldots, \ell_r) \mid Z_t = r, \boldsymbol{\lambda}^{(r)}) = \lambda^{(r)}_{(\ell_1,\ldots,\ell_r)}\,, \\
&\Pr(s_t = k \mid s_{t-1} = k_1, \ldots, s_{t-L} = k_L, Z_t = r, \boldsymbol{z}_t = (\ell_1, \ldots, \ell_r), \boldsymbol{\mathcal{Q}}^{(r)}) \\
&\quad = (\boldsymbol{\mathcal{Q}}^{(r)})_{k,k_{\ell_1},\ldots k_{\ell_r}}\,,
\end{aligned} \tag{3.7}
$$

where $\boldsymbol{\alpha}_Q$ is a length-K vector of positive shape parameters (which could potentially be separately specified for each distribution in each $\boldsymbol{\mathcal{Q}}$); $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_3$ are length-R vectors containing probabilities such that $(\boldsymbol{\pi}_1)_r + (\boldsymbol{\pi}_3)_r < 1$; $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ are length-R vectors containing positive shape parameters; $\boldsymbol{\alpha}_{\lambda^{(r)}}$ is a length-$\binom{L}{r}$ vector of positive shape parameters; and $\beta_{\lambda^{(r)}} > 1$ is the SDM sparsity parameter. We always set $(\boldsymbol{\pi}_1)_0 = 0$ to avoid penalizing Λ_0, and recommend setting $(\boldsymbol{\pi}_3)_0 = 0$ as well. Note

that all quantities in (3.7) without explicit dependence are considered independent a priori.

As with the MTDg, posterior simulation can be accomplished entirely through closed-form Gibbs sampling. To simplify computation, we uniquely map all Z_t and $\boldsymbol{z}_t$ pairs onto a single variable $\zeta_t \in \left\{0, 1, \ldots, \left[\binom{L}{1} + \binom{L}{2} + \ldots + \binom{L}{R}\right]\right\}$ whose prior probability under the model is equal to the product of the corresponding Λ and λ. Full conditional distributions for $\boldsymbol{\Lambda}$, each $\boldsymbol{\lambda}^{(r)}$, and each probability vector in $\{\boldsymbol{\mathcal{Q}}^{(r)}\}$ are exactly analogous to multinomial-SBM, multinomial-SDM, and multinomial-Dirichlet conjugate updates, respectively, where Z_t, $\boldsymbol{z}_t$, and observed data transitions supply the respective multinomial counts. Full conditional updates for Z_t and $\boldsymbol{z}_t$ (equivalently ζ_t) require calculation and sampling from a discrete distribution. Full details are given in Appendix B.3.

We again improve mixing in the sampler by integrating $\{\boldsymbol{\mathcal{Q}}^{(r)}\}_{r=0}^{R}$ from the joint posterior, sampling the collapsed conditional distributions for $\boldsymbol{\Lambda}$, each $\boldsymbol{\lambda}^{(r)}$, and $\{\zeta_t\}$. These are supported by the tractable marginal distributions reported in Appendices A.2 and B.1. Additionally, to encourage occasional jumps between modes of the posterior, we include a hybrid independence-Metropolis step which jointly proposes $\boldsymbol{\Lambda}$, each $\boldsymbol{\lambda}^{(r)}$, and $\{\zeta_t\}$ from their joint prior every 10 iterations of the MCMC algorithm.

The augmented Gibbs sampler becomes computationally demanding as R and L increase because updates for the latent configuration variables $\{\zeta_t\}$ involve calculation of $\sum_{r=0}^{R}\binom{L}{r}$ probabilities for each time point $t = L+1, \ldots, T$. Random-walk Metropolis samplers for $\boldsymbol{\Lambda}$, $\{\boldsymbol{\lambda}\}$, and $\{\boldsymbol{\mathcal{Q}}\}$ utilizing the mixture likelihood based on (3.4) may provide an alternate strategy if K is reasonably small. The logit-normal distribution (Atchison and Shen, 1980), or multivariate Gaussian random walks on the logit scale, facilitate properly constrained proposals for

probability vectors.

3.4 Simulation study

To demonstrate the effectiveness of the MMTD model for both objectives and to compare transition probability estimation performance with existing methods, we report two simulation studies. Both simulation scenarios feature time series generated from true Markov chains of differing order and lag configuration. In Simulation 1, the true generating model is a third-order chain with three states ($K = 3$) in which transition probabilities depend on lags 1, 3, and 4. In Simulation 2, the true generating model is a fifth-order binary chain ($K = 2$) for which each of the first five lags contributes to transition probabilities. In both models, each distribution in the transition tensor $\mathbf{\Omega}$ was drawn from a uniform distribution on the simplex (i.e., symmetric Dirichlet distributions with all shape parameters equal to 1). Each chain was randomly initialized and run for 1,000 steps of burn-in. The first 1,000 samples thereafter were reserved for training data and the next 1,000 for validation.

To evaluate estimation of transition probabilities, each model was fit using the prescribed number of training samples, and point estimates of the transition distributions were compared to the true transition distributions for each of the 1,000 validation points. Specifically, for validation time point t', each model produced a vector $\hat{\boldsymbol{p}}_{t'}$ to estimate each $p_{t'}^{(k)} = \Pr(s_{t'} = k \mid s_{t'-1}, \ldots, s_{t'-L}) = (\mathbf{\Omega})_{k,s_{t'-1},\ldots,s_{t'-L}}$, for $k = 1, \ldots, K$. In Bayesian models, the point estimate is the Monte Carlo-computed posterior mean of $\hat{\boldsymbol{p}}_{t'}$. In non-Bayesian models, $\hat{\boldsymbol{p}}_{t'}$ is computed from the optimized model fit. For each validation time point, we computed the $\mathcal{L}_1$ loss given by $L_{t'} = \sum_{k=1}^{K} |\hat{p}_{t'}^{(k)} - p_{t'}^{(k)}|$. The reported loss metric for model comparison is $100 \times \sum_{t'} L_{t'}/(KT')$, that is, 100 times the mean $\mathcal{L}_1$ loss across the $T' = 1{,}000$

validation points.

We fit the MTD, MTDg, and MMTD models to each training set with various settings. Implementation for the MTD model is similar to the MTDg and is described in Section 2.3.1. Let MTD(L) and MTDg(L) denote the respective model fits with user-specified maximum lag horizon L, and let MMTD(L, R) denote a model fit with user-specified maximum lag horizon L and maximum order R. All transition distributions in all $\{\boldsymbol{Q}\}$ and $\{\boldsymbol{\mathcal{Q}}\}$ in all three models utilize symmetric, unit-information Dirichlet priors (i.e., whose shape parameters all equal $1/K$ so that they sum to unity).

We use two prior settings for the MTD model. The first employs a Dirichlet prior for $\boldsymbol{\lambda}$ with all shape parameters equal to $1/L$. The second setting uses a SBM prior for $\boldsymbol{\lambda}$ with $\pi_1 = 0.5$, $\pi_3 = 0.1$, $\eta = 1{,}000$, and $\boldsymbol{\gamma}, \boldsymbol{\delta}$ selected to mimic the Dirichlet prior with shape parameters equal to $1/L$ and sparsity correction on $\boldsymbol{\delta}$ (Section 2.2.2). This prior encourages a moderate level of sparsity as well as decreasing prior probability for higher lags.

The MTDg model uses a SBM prior for $\boldsymbol{\lambda}$ with $\pi_1 = 0$ for λ_0 and $\pi_1 = 0.5$ thereafter; $\pi_3 = 0$ for λ_0 and $\pi_3 = 0.2$ thereafter; $\eta = 1{,}000$; $\gamma_0 = \delta_0 = 1$, yielding a uniform prior for λ_0, and remaining elements of $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ selected to mimic a Dirichlet prior with shape parameters equal to $1/L$ and sparsity correction on $\boldsymbol{\delta}$. This prior avoids penalizing λ_0, encourages a moderate level of sparsity in the remaining lags, and steeper decrease in prior probability for higher lags than used for the MTD.

We use two prior settings in the MMTD models. Both follow (3.7) with $\boldsymbol{\alpha}_{\boldsymbol{\lambda}^{(r)}} = \left(1/\binom{L}{r}, \ldots, 1/\binom{L}{r}\right)$, but the first setting uses $\beta_{\boldsymbol{\lambda}^{(r)}} = 1$ for all $r = 1, \ldots, R$, resulting in the symmetric, unit-information Dirichlet prior. The second setting uses $\beta_{\boldsymbol{\lambda}^{(r)}} = \sqrt{T}$ to encourage selection of a single-lag configuration within level

r. In both prior settings, we employ the SBM prior for $\mathbf{\Lambda}$ with $\pi_1 = 0$ for Λ_0 and $\pi_1 = 0.25$ thereafter; $\pi_3 = 0$ for λ_0 and $\pi_3 = 0.25$ thereafter; $\eta = 1{,}000$; and $\gamma = \delta = 1$ for all second-component beta distributions, yielding a uniform prior for Λ_0. This prior avoids penalizing Λ_0, allows for sparsity in the remaining lags, and maintains soft ordering that favors lower levels of the model. Because R is typically kept to small values, it is important that π_1 not be large and that π_3 not be too small. Otherwise, the prior can inappropriately allocate substantial mass toward large values of Λ_R. We recommend checking for this condition as part of prior sensitivity analysis.

To obtain results in Sections 3.4.1, 3.4.2, and 3.5, each model was initialized with random draws from Dirichlet distributions for $\mathbf{\Lambda}$ and each $\boldsymbol{\lambda}^{(r)}$. Random initialization in these models calls for long burn-in periods, on the order of tens to hundreds of thousands of iterations. In our analyses, 200,000 burn-in iterations were followed by another 400,000 iterations. Reported posterior quantities were calculated using a thinned sample retaining every 200th iteration. These conservative settings produced (unless otherwise noted) stable chains suitable for inference. In some cases, parallel chains sampled from MMTD models settled in neighboring modes which had minor impact on inferences and performance.

In addition to our proposed models, we fit the multinomial generalized linear models with logistic link functions to each training set using the *VGAM* package in R (Yee et al., 2010). To distinguish different settings, we denote model fit as LogitMC(L, R') with maximum lag horizon L and highest interaction order among the linear predictors R'. We also fit the variable length Markov chain models, denoted VLMC, using the *VLMC* package in R (Maechler, 2015) and employing default model settings.

3.4.1 Simulation 1 results

All models were fit to the time series from Simulation 1 for two sample sizes, $T = 200$ and $T = 500$. Here, we assume that the modeler is considering up to a horizon of six lags, which we use where possible to promote equitable comparisons. Results of the mean $\mathcal{L}_1$ loss across the 1,000 validation points are given in Table 3.2. In addition to transition probability estimation, we are interested in inferences for Markovian order and important lags afforded by the MTD, MTDg, and MMTD models. With exception of the MTD and MTDg, we see improved estimation with the larger sample size across all models.

Sample size 200

In the $T = 200$ case, the multinomial logistic models produce the best and worst results. Fitting all second-order interactions for up to six lags is cumbersome in this model, resulting in poor estimates. Fitting the full-order model to the correct lags only produces accurate estimates. However, this would require preliminary results from an iterative process which may or may not select the correct model and does not account for model uncertainty. We emphasize here that our proposed models do not require a model selection process if the modeler specifies the maximum lag horizon L and maximum order R, as order and lag inferences are built-in.

The variable length Markov chain model offers no improvement, possibly because the dynamics governing Simulation 1 skip lag 2. VLMC branches utilizing more distant lags must pass through and include lag 2. This results in a missed opportunity for greater parsimony (Jääskinen et al., 2014).

The MTD and MTDg models offer little help in this scenario because Simulation 1 is third-order with non-additive interactions. Posterior densities for $\boldsymbol{\lambda}$ (not shown) reveal that λ_3 is favored under the Dirichlet prior (in the MTD) and dominates

$T = 200$		$T = 500$	
Model	Loss	Model	Loss
LogitMC(6, 1)	18.70	LogitMC(6, 1)	16.77
LogitMC(6, 2)	37.42	LogitMC(6, 2)	18.84
LogitMC(3, 1), Lags 1, 3, 4 only	17.14	LogitMC(3, 1), Lags 1, 3, 4 only	16.39
LogitMC(3, 2), Lags 1, 3, 4 only	13.70	LogitMC(3, 2), Lags 1, 3, 4 only	10.40
LogitMC(3, 3), Lags 1, 3, 4 only	**12.64**	LogitMC(3, 3), Lags 1, 3, 4 only	7.64
VLMC	19.12	VLMC	15.26
MTD(6), Dir($\boldsymbol{\lambda}$)	17.29	MTD(6), Dir($\boldsymbol{\lambda}$)	17.27
MTD(6), SBM($\boldsymbol{\lambda}$)	17.27	MTD(6), SBM($\boldsymbol{\lambda}$)	17.21
MTDg(6)	17.30	MTDg(6)	17.05
MMTD(6, 2), Dir($\boldsymbol{\lambda}$)	14.92	MMTD(6, 2), Dir($\boldsymbol{\lambda}$)	13.77
MMTD(6, 2)	14.78	MMTD(6, 2)	13.83
MMTD(6, 3), Dir($\boldsymbol{\lambda}$)	14.65	MMTD(6, 3), Dir($\boldsymbol{\lambda}$)	7.55
MMTD(6, 3)	13.72	MMTD(6, 3)	**7.44**
MMTD(6, 4), Dir($\boldsymbol{\lambda}$)	14.93	MMTD(6, 4), Dir($\boldsymbol{\lambda}$)	7.58
MMTD(6, 4)	14.24	MMTD(6, 4)	7.48
MMTD(6, 5), Dir($\boldsymbol{\lambda}$)	15.13	MMTD(6, 5), Dir($\boldsymbol{\lambda}$)	7.56
MMTD(6, 5)	14.40	MMTD(6, 5)	7.47

Table 3.2: Simulation 1 ($K = 3$ states for a third-order chain with active lags 1, 3, and 4). Results for transition probability estimation under various models and model settings using two sample sizes, $T = 200$ and $T = 500$. The reported loss is 100 times the mean $\mathcal{L}_1$ loss, computed across 1,000 validation time points. Within each sample size group, the lowest mean loss is highlighted with bold font.

with the SBM prior (in both the MTD and MTDg). The latter effectively produces a first order Markov chain dependent on the third lag.

Several MMTD models were fit with increasing maximum order R ranging from 2 to 5. The second-order model provides a substantial improvement over mixing first-order transitions and fits nearly as well as the correctly specified third-order model. As expected, estimation performance stops improving when R exceeds the true order of three. Using SDM priors on $\boldsymbol{\lambda}$ parameters improves estimation, but more so when the correct model is contained in the specified model.

In the $R = 2$ model, posterior inference supports second-order dynamics with the lag 3,4 combination receiving most posterior weight. Adding the SDM priors on $\{\boldsymbol{\lambda}^{(r)}\}$ results in stronger support for the same conclusion. In the $R = 3$ model, posterior inferences support second or third-order dynamics with lags 3 and 4 receiving most posterior weight. Adding the SDM priors led to weakly favoring order 3 (selecting lags 1, 3 and 4) in one model run. The $R = 4$ model most often supports second-order dynamics (lags 3 and 4) under both prior scenarios. Results from the $R = 5$ model are similar to the $R = 4$ model.

It appears that the signal associated with lag 1 is relatively weak when $T = 200$. The SBM prior on $\boldsymbol{\Lambda}$ shrinks inferences toward second-order, but not decidedly away from third-order dynamics. Overall, the MMTD consistently produces the most faithful estimates of transition probabilities from a single model without requiring iterative model selection.

Sample size 500

In the $T = 500$ case, even the multinomial logistic models fit directly to the correct lags only fail to outperform the MMTD with $R \geq 3$. With a larger sample size, the VLMC model is more competitive, but the MTD and MTDg remain insufficiently flexible to capture the structure. The MTD model mixes over lags 2, 4, and 5 with a Dirichlet prior on $\boldsymbol{\lambda}$, and primarily over lags 4 and 5 with the SDM prior. The MTDg concentrates some mass on lags 3 and 4.

In this large-sample scenario, MMTD performance improves substantially when the specified mixture model contains the true model structure (i.e., $R \geq 3$), although the second-order MMTD again improves over the first-order additive MTD and MTDg. In the MMTD models with $R \geq 3$, posterior mass concentrates on the correct order and lag configuration. Furthermore, we see little drop in

performance when the maximal order is over-specified. The SDM priors on $\{\boldsymbol{\lambda}^{(r)}\}$ appear not to significantly improve estimation, and inferences are qualitatively similar. The inferences from the $R = 2$ model are similar to those of the $R = 2$ models fit to $T = 200$ observations. Among the models considered in this simulation scenario, the MMTD consistently produces the most faithful estimates of transition probabilities.

3.4.2 Simulation 2 results

All models were fit to the time series from Simulation 2 for three sample sizes: $T = 100$, $T = 200$ and $T = 500$. Here, we assume that the modeler is considering up to a horizon of seven lags, which we use where possible to promote equitable comparisons. Results of the mean $\mathcal{L}_1$ loss across the 1,000 validation points are given in Table 3.3. Again, we examine order and lag inferences from the MTD, MTDg, and MMTD models in addition to estimation performance.

Sample size 100

In the $T = 100$ case, high order interactions are not estimable in the multinomial logistic model. The VLMC model performs best in this scenario, presumably because Simulation 2 features no gap in relevant lags.

Because the simulation uses lags 1 through 5, the MTD(7), MTDg(7), and MMTD(7, 4) models are under-specified and must rely on a lower-order sub-model and/or mixing across lags to approximate the fifth-order dynamics. The MTD models mix primarily over lags 2 and 4, while the MTDg model concentrates on lag 2. The MMTD(7, 4) models mix primarily over low orders, with slight preference for lag 2. The over-specified MMTD(7, 7) models do not outperform the $R = 4$ models, and produce qualitatively equivalent inferences for $\boldsymbol{\Lambda}$ and $\{\boldsymbol{\lambda}^{(r)}\}$.

$T = 100$		$T = 200$		$T = 500$	
Model	Loss	Model	Loss	Model	Loss
LogitMC(7, 1)	24.30	LogitMC(7, 1)	20.03	LogitMC(7, 1)	18.53
LogitMC(7, 2)	26.15	LogitMC(7, 2)	16.26	LogitMC(7, 2)	14.66
LogitMC(7, 3)	n/a	LogitMC(7, 3)	18.70	LogitMC(7, 3)	13.67
LogitMC(5, 1)	24.49	LogitMC(5, 1)	20.25	LogitMC(5, 1)	18.90
LogitMC(5, 2)	20.86	LogitMC(5, 2)	16.24	LogitMC(5, 2)	15.35
LogitMC(5, 3)	n/a	LogitMC(5, 3)	**11.26**	LogitMC(5, 3)	8.29
LogitMC(5, 4)	n/a	LogitMC(5, 4)	n/a	LogitMC(5, 4)	7.79
VLMC	**20.52**	VLMC	15.45	VLMC	12.13
MTD(7) with Dir($\boldsymbol{\lambda}$)	24.71	MTD(7) with Dir($\boldsymbol{\lambda}$)	22.47	MTD(7) with Dir($\boldsymbol{\lambda}$)	19.82
MTD(7) with SBM($\boldsymbol{\lambda}$)	24.50	MTD(7) with SBM($\boldsymbol{\lambda}$)	23.31	MTD(7) with SBM($\boldsymbol{\lambda}$)	19.59
MTDg(7)	24.01	MTDg(7)	23.93	MTDg(7)	19.68
MMTD(7, 4) with Dir($\boldsymbol{\lambda}$)	23.68	MMTD(7, 4) with Dir($\boldsymbol{\lambda}$)	15.26	MMTD(7, 4) with Dir($\boldsymbol{\lambda}$)	12.15
MMTD(7, 4)	23.21	MMTD(7, 4)	15.38	MMTD(7, 4)	14.70
MMTD(7, 7) with Dir($\boldsymbol{\lambda}$)	23.68	MMTD(7, 7) with Dir($\boldsymbol{\lambda}$)	14.13	MMTD(7, 7) with Dir($\boldsymbol{\lambda}$)	7.59
MMTD(7, 7)	23.33	MMTD(7, 7)	13.93	MMTD(7, 7)	**7.38**

Table 3.3: Simulation 2 ($K = 2$ states for a fifth-order chain with five active lags). Results for transition probability estimation under various models and model settings using three sample sizes: $T = 100$, $T = 200$ and $T = 500$. The reported loss is 100 times the mean $\mathcal{L}_1$ loss, computed across 1,000 validation time points. Within each sample size group, the lowest mean loss is highlighted with bold font.

It is apparent that the small sample size is insufficient to capture the fifth-order structure.

Sample size 200

With $T = 200$, the time series is long enough to include third-order interactions in the multinomial logistic model, which performs well. The VLMC model is again competitive with the MMTD and generally outperforms the logistic models.

As before, the MTD and MTDg models are unable to leverage increased sample size to the extent that the other models can. The MTD models mix primarily over lags 1 and 5, while the MTDg model mixes primarily on lag 1 (due to the ordered prior on $\boldsymbol{\lambda}$).

The higher-order interactions allowed by the MMTD become advantageous with $T = 200$, making this model competitive. The MMTD(7, 4) models concentrate posterior mass on order 4 and lags 1, 2, 3, and 5. Posterior mass in the MMTD(7, 7) model is split between order 4 and 5, again demonstrating the shrinking effect of the SBM prior on $\boldsymbol{\Lambda}$. Different runs of the MCMC chain favor lag configurations $(1, 2, 3, 5)$ and $(1, 2, 3, 4, 5)$. SDM priors on $\{\boldsymbol{\lambda}^{(r)}\}$ had a minor concentrating effect on the posterior densities.

We note that the over-specified MMTD(7, 7) with a less strictly ordered prior on $\boldsymbol{\Lambda}$ (such as the SDM) can outperform the correctly specified logistic model in this scenario. However, inferences from such models can be suspect, as they do not shrink toward the "reduced" and identifiable parameterization. While we favor reliably interpretable inferences, one may consider modifying or replacing the SBM prior on $\boldsymbol{\Lambda}$ to improve predictive performance.

Sample size 500

The fifth-order binary chain in Simulation 2 has 32 total (univariate) transition distributions which are easily estimated with 500 samples. Therefore, the multinomial logistic models with high-order interactions approach the performance of the over-specified MMTD models. The VLMC is also competitive. Again, the MTD(7) and MTDg(7) models lag noticeably behind in estimation performance, although both attempt to mix over multiple lags.

The MMTD(7, 4) with a Dirichlet prior on $\{\boldsymbol{\lambda}^{(r)}\}$ again concentrates on order

4 and lags 1, 2, 3, and 5. The same model with the SDM prior selects lags 3, 4, 5, and 6, and performs noticeably worse (loss of 14.70). A second MCMC run places most weight on orders 3 and 4, and lags 1, 3, 4, and 5, resulting in average $\mathcal{L}_1$ loss of 12.65. This highlights multimodality of the posterior and the need for replicate MCMC runs. The MMTD(7, 7) decisively identifies the correct order and lag structure resulting in the best estimation performance. We conclude that the MMTD consistently produces the most faithful estimates of transition probabilities from a single model without requiring iterative model selection.

3.5 Data illustrations

We now apply the MTDg and MMTD models to two data analyses. The first data example was studied with the original MTD and in the subsequent literature. The second is a novel analysis of pink salmon population dynamics in Alaska, U.S.A. during the twentieth century. We illustrate the use of inferences on order and lag importance available from the models.

3.5.1 Seizure data

Berchtold and Raftery (2002) demonstrate the MTD model using a binarized time series adapted from MacDonald and Zucchini (1997), which reports the occurrence of at least one epileptic seizure for a patient on each of 204 consecutive days. Berchtold and Raftery (2002) fit several Markov chain and MTD models, using the Bayesian information criterion (BIC) to ultimately select a MTD with eight lags. They report that λ_8 has the greatest magnitude. Note that the MTD model used in Berchtold and Raftery (2002) allows negative values in $\boldsymbol{\lambda}$, which requires a complex set of constraints for estimation. The seizure time series was

revisited using the methods of Sarkar and Dunson (2016), who report a model of maximal order 8 (with three active lags in the posterior mode), with lag 8 having the highest posterior inclusion probability. Although coefficient magnitudes and lag-inclusion probabilities are not necessarily commensurate, the two models agree on the maximal order and most important lag. They appear to differ, however, on the relative importance of other lags. In Berchtold and Raftery (2002), λ_4 is a distant second in magnitude, whereas Sarkar and Dunson (2016) report that lag 1 has much higher posterior inclusion probability than lag 4.

In light of these two analyses, we fit the MTDg and MMTD to the seizure data with $L = 10$ and $R = 4$, each with prior settings identical to those used in the simulation studies. Trace plots (not shown) indicate that the marginal posterior distributions over $\boldsymbol{\Lambda}$ and each $\boldsymbol{\lambda}^{(r)}$ are multimodal, suggesting that more than one combination of lags could model the dynamics with similar accuracy. We note also that the assumption of time-homogeneity is questionable, as no seizures were reported in the last 29 days.

The MTDg model concentrates most posterior weight on lag 8, followed distantly by lags 4 and 9. The transition matrix for lag 8 suggests that the status eight days prior is most often replicated in the present (seizure or no seizure). This transition pattern is repeated for lags 4 and 9, producing a compounding effect for repeated seizures in the model, which effect we should emphasize is additive only. That this model clearly selects lag 8 demonstrates the utility of the SBM prior for the MTDg. The prior simultaneously shrinks parameters toward the identifiable model and maintains a conditional stochastic ordering on the lags while maintaining flexibility to select distant lags when this is supported by the data.

The MMTD(10, 4) model with Dirichlet priors on lag configurations mixes primarily over orders 1 and 2. The standard model (with SDM priors on lag

configuration weights) shifts more posterior weight to higher orders, with Λ_2 and Λ_3 edging one another in separate MCMC runs. Without clear selection of order and lags, we discourage over-interpretation of the transition probabilities in $\{\boldsymbol{\mathcal{Q}}^{(r)}\}$. However, it is clear that the estimated probabilities favor persisting in previous states. For example, the posterior means for $(\boldsymbol{\mathcal{Q}}^{(2)})_{1,1,1}$ and $(\boldsymbol{\mathcal{Q}}^{(2)})_{2,2,2}$ are 0.78 and 0.73 respectively (posterior medians are 0.92 and 0.85). That is, no occurrence of seizure in recent days yields a high probability for no seizure on the current day, and repeated occurrence of seizures on multiple past days yields a high probability of seizure on the current day.

We can more comprehensively assess lag importance by computing a lag inclusion index as the sum of all products $\Lambda_r \times \lambda^{(r)}_{(\boldsymbol{z}_j)}$ across $j = 1, \ldots, \binom{L}{r}$ and $r = 1, \ldots, R$ for which lag ℓ appears in the lag configuration $\boldsymbol{z}_j$. We compute this for each lag at each MCMC sample. Inference for Λ_0 is included as lag 0 for reference. Due to the shrinking SBM prior for $\boldsymbol{\Lambda}$, a high inclusion index for lag 0 should not be interpreted as a lack of Markovian dependence (unless it is near 1 with high confidence). However, a low inclusion index for lag 0 relative to other lags can indicate strong Markovian dependence. We summarize the inclusion index for the models fit to the seizure data in Figure 3.1, with bars reporting the posterior mean and whiskers reaching to the ends of 95% posterior credible intervals. Note the large uncertainty for this inclusion index for all lags except lag 8 in the model with SDM priors on $\{\boldsymbol{\lambda}^{(r)}\}$. We further note that the posterior inclusion pattern across lags (associated with the SDM priors for lag configuration weights) resembles a plot with similar interpretation in Figure 6 (e) of Sarkar and Dunson (2016). The most notable exception is that lag 1 does not feature prominently in our inferences. All analyses, including our three, agree that lag 8 is the most important in determining the transition probability.

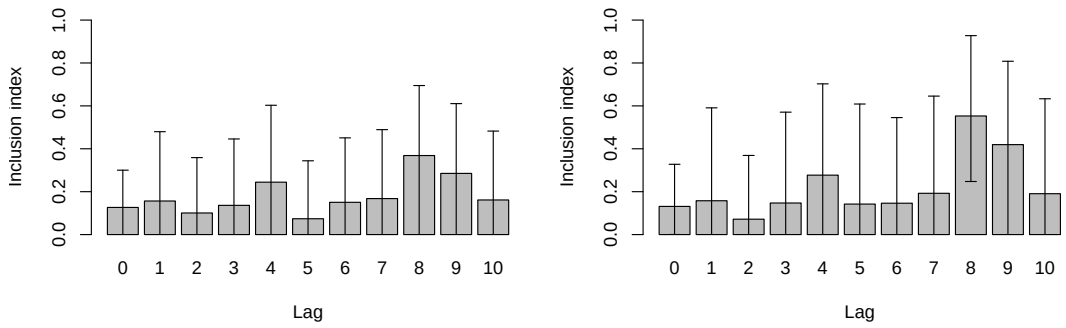


Figure 3.1: Posterior mean (with 95% credible interval) inclusion index for each lag in the seizure analysis, under the MMTD(10,4) model with the Dirichlet priors (left) and SDM priors (right) on lag configuration weights $\{\boldsymbol{\lambda}^{(r)}\}$.

3.5.2 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 a testing opportunity for our model because 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. A time-series plot of the natural logarithm of abundance is given in Figure 3.2 together with bivariate lag scatter plots. In this scenario, we might expect non-stationarity with long-term trends. It appears from the time series that the even-year population began to struggle in the late 1940s. Repeated interventions 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, the lag scatter plots suggest that we should be able to detect lag dependence structure, even with as few as 30 observations. After discretizing the data into sets of $K = 4$ quantile-based bins using all 30 years, we fit the proposed models with the same prior settings used for the simulation studies. Because discretization is based on quantiles, results are invariant to monotonic

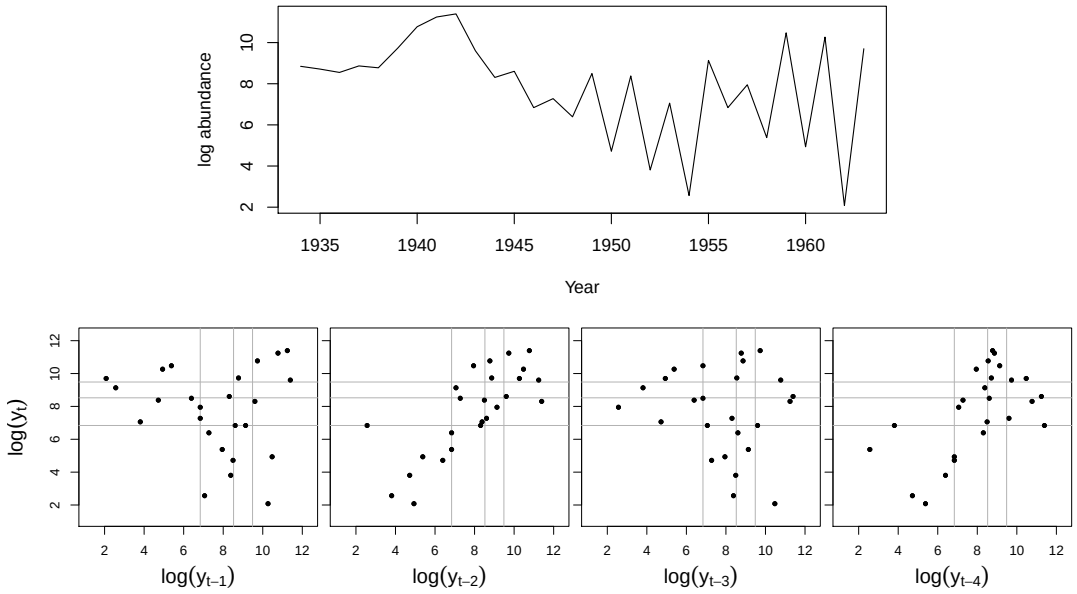


Figure 3.2: Time-series plot and lag scatter plots for the natural logarithm of pink salmon abundance from 1934 to 1963. In the lag plots, y_t denotes abundance at time t and horizontal/vertical lines separate $K = 4$ quantile-based bins used to assign $\{y_t\}$ into discrete states $\{s_t\}$.

transformations such as the natural logarithm.

The MTDg(5) model fit to the pink salmon time series clearly identifies lag 2 as the most influential (posterior means for λ_ℓ, $\ell = 0, 1, \ldots, 5$ are 0.17, 0.05, 0.67, 0.01, 0.07, and 0.03, respectively). The estimated $\boldsymbol{Q}^{(2)}$ also closely agrees with the lag-2 scatter plot in Figure 3.2. In contrast, the MMTD(5,2) model shifts considerable posterior weight toward order 2 despite the shrinkage prior on order. Uncertainty, stemming from noisy dynamics and a small sample size, again results in a multimodal posterior, as seen in the density plots for $\boldsymbol{\Lambda}$ in Figure 3.3. This uncertainty is also apparent in posterior inferences for the lag inclusion index. Credible intervals on the lag inclusion index are wide enough to warrant their omission from Figure 3.4. In these plots, we see essential agreement between the two prior settings for $\{\boldsymbol{\lambda}^{(r)}\}$, with lag 2 being most prominent. Lags 4 and 5 also appear to contribute in some of the favored configurations.

It is important to examine the MMTD estimate of $\boldsymbol{\mathcal{Q}}^{(2)}$ to verify that the model is not attempting to fit first-order dynamics with a second-order chain. If this were the case, estimates of transition probabilities in $\boldsymbol{\mathcal{Q}}^{(2)}$ would repeat across the second lag index (in this case most likely representing lag 4 and/or 5). The posterior mean estimate of $\boldsymbol{\mathcal{Q}}^{(2)}$, shown in Figure 3.5, appears not to have this problem, as consecutive 4×4 sub-matrices appear not to repeat. This is consistent under both priors. Hence, lag 2 may not be the only important lag.

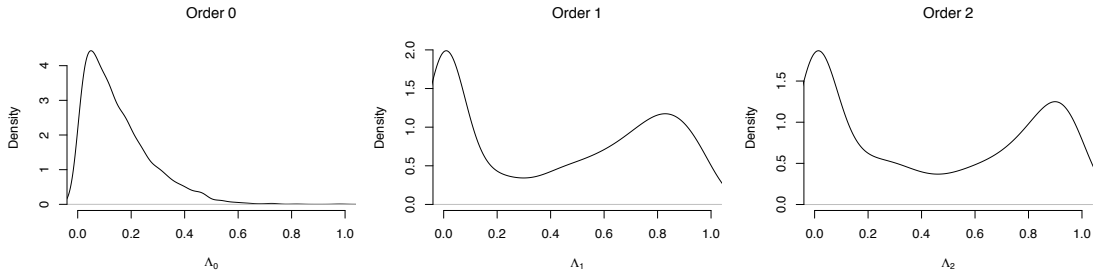


Figure 3.3: Marginal posterior density plots for $\boldsymbol{\Lambda}$ in the pink salmon analysis using a SBM prior on order and SDM priors for lag configuration weights.

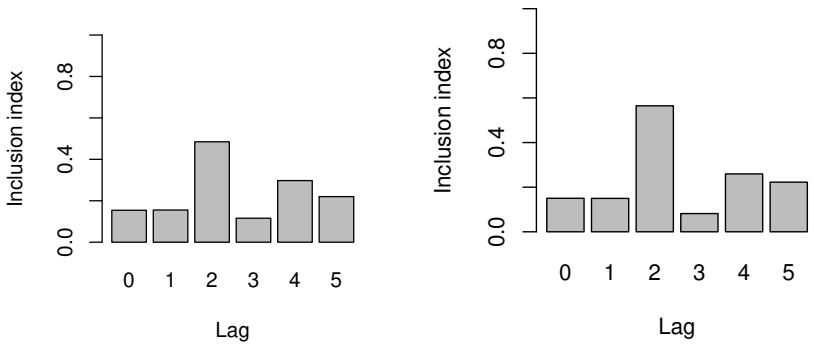


Figure 3.4: Posterior mean inclusion index for each lag in the pink salmon analysis under the MMTD(5,2) model with Dirichlet priors (left) and SDM priors (right) on lag configuration weights.

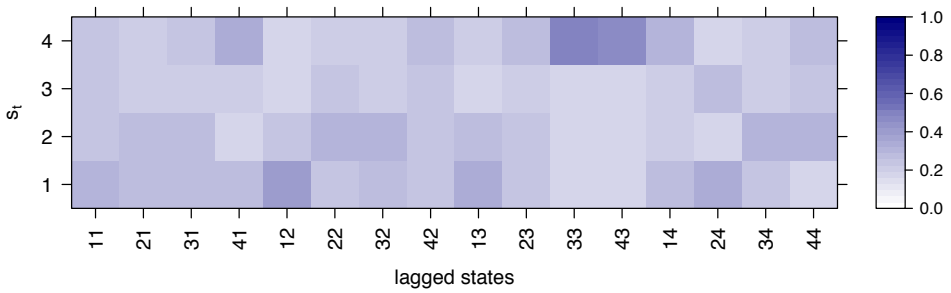


Figure 3.5: Posterior mean point estimate of the matricized $\boldsymbol{\mathcal{Q}}^{(2)}$ from the MMTD(5,2) pink salmon analysis with SDM priors on $\{\boldsymbol{\lambda}^{(r)}\}$. Rows (along the y-axis) represent states to which the transition occurs, and columns (along the x-axis) represent the states occupied by the first two selected lags, with the state corresponding to the most recent lag changing index first.

3.6 Summary

We have explored two extensions of the original mixture transition distribution model for high-order Markov chains. The first is a Bayesian approach to a recent extension capable of identifying one or more important lags. The second captures higher-order interactions and can potentially yield useful inferences for order and lag importance. To accomplish the latter, the mixture of mixture transition distributions uses an over-specified model and sparsity-inducing priors to shrink back to an interpretable and informative structure. This is accomplished in a single model without necessitating iterative selection. Furthermore, our MCMC algorithm allows us to evaluate uncertainty about the model structure and transition probabilities. We demonstrated that this model can outperform some of the standard methods in transition probability estimation, and shown its practical utility in data analysis.

The over-specified MMTD model can offer insights into order and active lags provided the modeler approaches analysis attentively. In cases of large sample size or near-determinism, the true structure will immediately manifest in inferences for the mixture weight parameters. More often, lag importance should be aggregated

and extracted in post-processing as we demonstrated in Section 3.5.1 with the lag inclusion index. If multiple lag patterns are prominent in the mixture model or different order components have non-overlapping lag patterns selected in each, the actual order of the time series may be higher than the highest selected model order. We also recommend checking the mixture transition tensors for redundancy, a sign of lower-order dependence. Absent a clearly identified lag structure through inferences on the mixture weights, we claim that this too can be informative.

Both the MTDg and MMTD models can approximate high-order dynamics by exploiting constructive additivity among lower-order transition probabilities. However, when this does not apply, a full jump to the next order in the MMTD is required. For example, in the salmon data analysis with five lags, the first-order mixture has five components associated with a 5×5 transition matrix, whereas the second-order mixture has ten components associated with a $5 \times 5 \times 5$ transition tensor. A parsimonious compromise might rely more on a factorization structure, similar to the one noted in Section 3.2.1. We do not pursue this here, but rather choose to emphasize the interpretable structure of the proposed MMTD model (3.4) which showcases a model-averaging flavor with added flexibility.

Chapter 4

Density Autoregression with the Gaussian Process Mixture Transition Distribution

4.1 Introduction

We now shift focus from discrete to continuous state spaces, a more natural setting for the population dynamics applications previously considered. While many standard models on the time domain are Markovian (linear autoregressive models being the primary example), they are typically viewed separately from traditional Markov chain models, in part because the methodology addresses distinct challenges. For example, Markov chains are traditionally unstructured, leading to an explicit conflict with the "curse of dimensionality" when incorporating additional lags. In contrast, conditional modeling in continuous spaces often relies on simplifying assumptions such as linearity and additivity when expanding the lag-embedding space. In this sense, mixture transition distribution models fit

into the mainstream with continuous spaces more than they do in their originally proposed discrete domain. We explore and pursue a flexible MTD model for continuous state spaces in this chapter.

That the MTD framework extends beyond discrete state spaces was first noted by Martin and Raftery (1987). The general MTD formulation for the conditional distribution F on time series $\{y_t\}_{t=1}^T \in \mathbb{R}^T$ is given by

$$F_t(y_t \mid y_{t-1}, \ldots, y_1) = \sum_{\ell=1}^{L} \lambda_\ell \, G_\ell(y_t \mid y_{t-\ell}) \,, \tag{4.1}$$

where each mixture component contains a univariate transition law G_ℓ associated with a specific lag. As before, each $\lambda_\ell \geq 0$, and $\sum_{\ell=1}^{L} \lambda_\ell = 1$. The most popular, and perhaps simplest model belonging to this family is the Gaussian MTD (GMTD) proposed by Le et al. (1996), wherein G_ℓ corresponds to a Gaussian distribution with linear mean $\beta_\ell \, y_{t-\ell}$ and variance σ_ℓ^2. They further include a mixture component containing a full linear autoregressive (AR) model of order L. With this simple form, the GMTD offers flexibility and better captures characteristics unavailable to standard AR models. Further modifications include a zero-mean component with large variance to accommodate outliers, and a random-walk specification to accommodate flat stretches. Le et al. (1996) derive conditions for weak stationarity and autocorrelation properties of the GMTD model before demonstrating its use with time series from financial and chemical process applications.

Despite this flexibility to model what are often termed as "nonlinear" time series, the GMTD model has a linear and additive transition mean. If we denote the coefficients for the full AR component as $\beta_{01}, \ldots, \beta_{0L}$, then the conditional transition mean for the GMTD is $\mathrm{E}(y_t \mid y_{t-1}, \ldots, y_1) = \sum_{\ell=1}^{L} (\lambda_0 \, \beta_{0\ell} + \lambda_\ell \, \beta_\ell) \, y_{t-\ell}$. While this linear structure is important for deriving stationarity conditions, we forego this restriction in favor of estimating nonlinear dependence. Thus we will

consider each $G_\ell(y_t \mid y_{t-\ell})$ to have a separate location $\mu_\ell + f_\ell(y_{t-\ell})$ consisting of a level and continuous nonlinear function $f_\ell(y_{t-\ell})$ mapping the relevant lag to $\mathbb{R}$. If the location of G_ℓ also represents the conditional mean, then we have $\mathrm{E}(y_t \mid y_{t-1}, \ldots, y_1) = \mu + \sum_{\ell=1}^{L} \lambda_\ell \, f_\ell(y_{t-\ell})$ where $\mu = \sum_{\ell=1}^{L} \lambda_\ell \, \mu_\ell$. This resembles the conditional mean obtained from the popular generalized additive model family (Hastie and Tibshirani, 1990), which has been applied to autoregressive models (Chen and Tsay, 1993; Wong and Kohn, 1996). Huang and Yang (2004) further consider order selection using BIC in this context. As Le et al. (1996) note, however, the MTD formulation is distinct from generalized additive models in that the errors arise from a mixture. That is, rather than averaging surfaces into a single composite with homogeneous error, the MTD model uses the functions f_ℓ to define the error mixture, which can vary widely across the input space. Thus the primary objective is not approximating a multi-dimensional surface, but rather flexibly modeling the transition distribution as a function of lags when several mixture components are active, and identifying low-order nonlinear dependence while inferring relevant lags when few components are active.

We propose to model the unknown functions $\{f_\ell\}$ with Gaussian process (GP) priors, which have been applied extensively for time series (see Gregorčič and Lightbody, 2009; Kocijan et al., 2003; Gutjahr et al., 2012 for examples in the nonlinear autoregressive context) and nonlinear regression generally (Rasmussen and Williams, 2006), including mixture modeling (Shi et al., 2003) and generalized additive formulations (Duvenaud et al., 2011). Retaining the sparsity-inducing prior for the mixing weights from Chapter 2, we propose a model with basic form

$$F_t(y_t \mid y_{t-1}, \ldots, y_1) = \lambda_0 \, \mathrm{N}\left(y_t \mid \mu_0, \sigma_0^2\right) + \sum_{\ell=1}^{L} \lambda_\ell \, \mathrm{N}\left(y_t \mid \mu_\ell + f_\ell(y_{t-\ell}), \, \sigma_\ell^2\right)$$

$$f_\ell \overset{\text{ind.}}{\sim} \mathrm{GP}, \quad \boldsymbol{\lambda} \sim \mathrm{SBM}\,, \tag{4.2}$$

where $\mathrm{N}(\cdot \mid \mu, \sigma^2)$ corresponds to a univariate Gaussian distribution with mean μ and variance σ^2, and $\boldsymbol{\lambda} = (\lambda_0, \lambda_1, \ldots, \lambda_L)$. To emphasize the MTD structure and utilize the model in a similar way to the approach taken in Chapter 2, we drop the full AR component and retain the zero-indexed component (intercept) with no lag dependence. Because each mixture component employs a parameterized distribution over an unbounded space, the intercept and shrinkage for $\boldsymbol{\lambda}$ do not share the same interpretations as in the MTDg and MMTD models, nor is the reduction procedure employed by Tank et al. (2017) as important for identifiability. We retain the intercept in order to accommodate what Le et al. (1996) term replacement-type outliers, to add flexibility to the mixture, and to allow the possibility of a (Gaussian) stationary distribution in the case of no serial dependence. The stick-breaking mixture prior for $\boldsymbol{\lambda}$ allows for sparsity, jumps to omit inactive lags, and stochastic ordering of active lags to reflect the belief that recent lags generally carry greater influence.

We outline a hierarchical model to implement the proposed model (4.2) and discuss prior selection and implementation in Section 4.2. In Section 4.3, we demonstrate the model with simulated and real time series. We then discuss two possible model extensions in Section 4.4, and conclude with discussion in Section 4.5.

4.2 Model

The full hierarchical specification for the Gaussian-process mixture transition distribution (GPMTD) model follows standard conventions for both Gaussian process regression and the MTD models presented earlier. As before, we break the mixture with latent component membership indicators for each time point, $z_t \in \{0, 1, \ldots, L\}$. To distinguish y_t from its lags and to emphasize that covariates

could be incorporated into the framework, we denote the time-delay vector as $\boldsymbol{x}_t = (x_{t,1}, \ldots, x_{t,L}) \equiv (y_{t-1}, \ldots, y_{t-L})$. For the Gaussian process priors, we exclusively employ Matérn covariance functions with Euclidean distance and a fixed smoothness parameter $\nu \in \{2.5, +\infty\}$, the former value ensuring a twice-differentiable regression function f and the latter corresponding to the squared exponential covariance function (Rasmussen and Williams, 2006).

Treating the first L observations of the time series as fixed, and implicitly conditioning the top level on lags in $\boldsymbol{x}_t$, the full hierarchical representation for the model in (4.2) is given by

$$
\begin{aligned}
y_t \mid z_t, \mu_0, \sigma_0^2, \{(\mu, \sigma^2, f)_\ell\}_{\ell=1}^L &\overset{\text{ind.}}{\sim} \begin{cases} \mathrm{N}\left(\mu_0, \sigma_0^2\right) & \text{if } z_t = 0, \\ \mathrm{N}\left(\mu_\ell + f_\ell(x_{t,\ell}), \sigma_\ell^2\right) & \text{if } z_t = \ell \in \{1, \ldots, L\}, \end{cases} \\
&\text{for } t = L+1, \ldots, T, \\
\Pr(z_t = \ell \mid \boldsymbol{\lambda}) = \lambda_\ell, \text{ for } \ell = 0, 1, \ldots, L, &\text{ independently for } t = L+1, \ldots, T, \\
\boldsymbol{\lambda} &\sim \mathrm{SBM}(\eta_\lambda, \pi_{1\lambda}, \pi_{3\lambda}, \boldsymbol{\gamma}_\lambda, \boldsymbol{\delta}_\lambda), \\
\mu_\ell \overset{\text{ind.}}{\sim} \mathrm{N}\left(m_0^{(\ell)}, v_0^{(\ell)}\right), \ \sigma_\ell^2 &\overset{\text{ind.}}{\sim} \mathrm{IG}\left(\nu_\sigma^{(\ell)}/2, \nu_\sigma^{(\ell)} s_0^{(\ell)}/2\right), \quad \text{for } \ell = 0, 1, \ldots, L, \\
f_\ell \mid \kappa_\ell, \sigma_\ell^2, \nu, \psi_\ell &\overset{\text{ind.}}{\sim} \mathrm{GP}\left(\mathbf{0}, \kappa_\ell\, \sigma_\ell^2\, \rho(x, x'; \nu, \psi_\ell)\right), \quad \text{for } \ell = 1, \ldots, L, \\
\kappa_\ell \mid \nu_\kappa, \kappa_0 &\overset{\text{ind.}}{\sim} \mathrm{IG}\left(\nu_\kappa/2, \nu_\kappa\, \kappa_0/2\right), \quad \text{for } \ell = 1, \ldots, L, \\
\psi_\ell \mid \nu_\psi, \psi_0 &\overset{\text{ind.}}{\sim} \mathrm{IG}\left(\nu_\psi/2, \nu_\psi\, \psi_0/2\right), \quad \text{for } \ell = 1, \ldots, L, \\
p(\nu_\kappa) \propto 1_{(\nu_\kappa \in \mathcal{V}_\kappa)}, \ \kappa_0 &\sim \mathrm{Ga}(a_\kappa, b_\kappa), \\
p(\nu_\psi) \propto 1_{(\nu_\psi \in \mathcal{V}_\psi)}, \ \psi_0 &\sim \mathrm{Ga}(a_\psi, b_\psi),
\end{aligned}
\tag{4.3}
$$

where the SBM is the stick-breaking mixture prior of Chapter 2; $\mathrm{IG}(a, b)$ denotes the inverse-gamma distribution with shape a and scale b; the Gaussian process is characterized by the zero mean function denoted with $\mathbf{0}$ and covariance function

$\kappa_\ell \, \sigma_\ell^2 \, \rho(\cdot, \cdot; \nu, \psi_\ell)$ utilizing correlation function ρ in the Matérn class with smoothness parameter ν and length-scale parameter ψ; $\mathcal{V}_\kappa$ and $\mathcal{V}_\psi$ are finite, discrete sets of positive real numbers; and $\text{Ga}(a, b)$ denotes a gamma distribution with mean a/b. Each inverse-gamma distribution is parameterized in terms of a scaled inverse Chi-squared distribution with degrees of freedom and prior harmonic mean, which aid both with interpretation and computation (potentially reduced posterior correlation among the parameters). We also parameterize the GP variance as the product $\kappa \, \sigma^2$ to aid with interpretation of κ as a signal-to-noise ratio, as well as computation, obtaining a tractable collapsed conditional distribution for each σ^2 parameter. Because $\boldsymbol{x}_t$ contains lags of the time series, it may be reasonable to assume some degree of homogeneity among $\{f_\ell\}$ across lags, for which we allow hierarchical borrowing-of-strength in the parameters governing the covariance functions. Indeed, even with ν fixed, κ and ψ are not fully identified (Zhang, 2004). Hence, we employ informative and hierarchically connected priors.

4.2.1 Prior and implementation

The GPMTD model is somewhat robust to prior choice, provided the parameters governing variances are on an appropriate scale. Experience simulating time series from the model suggests that values of κ on the order of 10^2 or 10^3 are necessary (when ψ is on the order of 1) for smooth nonlinear dynamics to be visually manifest. We typically set all $m_0^{(\ell)} = 0$ and $v_0^{(0)}$ large (orders of magnitude greater than the range of the data) to accommodate outliers, and lag-specific $v_0^{(\ell)}$ either large or commensurate with the range of the data. Absent strong beliefs about observation noise, we set all $\nu_\sigma = 5.0$ to ensure two finite moments in the inverse-gamma priors, with $s_0^{(0)}$ large (approximately one order of magnitude greater than the range of the data) and $s_0^{(\ell)} = 1.0$.

We have used $\mathcal{V}_\kappa = \mathcal{V}_\psi = \{5.0, 7.5, 10.0, 25.0, 50.0\}$ to define default discrete uniform priors on the degrees of freedom (concentration) parameters. We also use as default values $a_\kappa = 10.0$ and $b_\kappa = 0.1$, yielding a prior mean of 100.0 for κ_0 (the prior harmonic mean for each κ); and $a_\psi = 10.0$ and $b_\psi = 1.0$, yielding a prior mean of 10.0 for ψ_0 (the prior harmonic mean for each ψ). If one has strong prior beliefs regarding the strength of the dynamic signal relative to observation noise, we recommend first carefully considering an informative prior for each σ^2_ℓ for $\ell > 0$, and then setting an informative prior for the κ parameters by possibly increasing the values in $\mathcal{V}_\kappa$ and concentrating the gamma prior for κ_0.

The parameters in the SBM prior for the mixing weights should be thoughtfully considered in the context of each analysis, especially in cases with sample sizes $T < 50$. For example, priors overly concentrated on λ_0 in conjunction with a small σ^2_0 can result in an unintended bimodal transition distribution. We recommend following the procedures outlined in Sections 2.2.2 and 2.2.3 for selecting a SBM prior with a level of sparsity reflecting prior beliefs about the number of active lags in the time series. We employ default values of $\eta_\lambda = 1{,}000$, $\pi_{1\lambda} = 0.5$, $\pi_{3\lambda} = 0.25$, $\boldsymbol{\gamma}_\lambda = \mathbf{1}$, and $\boldsymbol{\delta}_\lambda = \mathbf{1}$ where $\mathbf{1}$ is a vector of ones. This results in a marginal prior density with peaks near the extremes and near-uniformity between 0 and 1 for each λ_ℓ. Note that unlike the MTDg and MMTD models in Chapter 3, the intercept weight λ_0 is subject to the small and large SBM beta mixture components.

The hierarchical model in (4.3) admits a convenient Gibbs sampling scheme for posterior inference. We make a few general remarks and highlight details unique to this model, and defer all remaining details to Appendix C. As is standard with Gaussian process regression, we work with the finite-dimensional distributions of the independent prior processes for $\{f_\ell\}$, which are multivariate Gaussian with mean 0 everywhere and covariance between all input pairs (x, x'), $x, x' \in \mathbb{R}$,

parameterized as in (4.3). Let $\boldsymbol{f}_\ell$ denote a length $T - L$ vector for which the ith element is the realization $f_{i,\ell} \equiv f_\ell(x_{i,\ell})$. To encourage mixing of the MCMC chain, we marginalize the full posterior over all $\{(\mu, \sigma, \boldsymbol{f})_\ell\}$ before updating $\{(\kappa, \psi)_\ell\}$, the only parameters for which collapsed/full conditional distributions are not tractable. For each $\ell = 1, \ldots, L$, we jointly update the pair $(\kappa, \psi)_\ell$ with a random-walk Metropolis step using bivariate Gaussian proposals on the logarithmic scale. Given these updates, conditionally conjugate updates are available for individual parameters in $\{(\mu, \sigma, \boldsymbol{f})_\ell\}$. We note that each f_ℓ must be evaluated at every $x_{t,\ell}$ (denoted as $f_{t,\ell}$) to facilitate full conditional draws for $\{z_t\}$, given as

$$\Pr(z_t = \ell \mid \cdots) = \frac{\lambda_0 \,\mathrm{N}(y_t \mid \mu_0, \sigma_0^2) 1_{(\ell=0)} + \lambda_\ell \,\mathrm{N}(y_t \mid \mu_\ell + f_{t,\ell}, \sigma_\ell^2) 1_{(\ell>0)}}{\lambda_0 \,\mathrm{N}(y_t \mid \mu_0, \sigma_0^2) + \sum_{j=1}^{L} \lambda_j \,\mathrm{N}(y_t \mid \mu_j + f_{t,\ell}, \sigma_j^2)}, \quad (4.4)$$

for $\ell = 0, 1, \ldots, L$, and $t = L+1, \ldots, T$, where in this context, $\mathrm{N}(\cdot \mid \mu, \sigma^2)$ denotes a Gaussian density function with mean μ and variance σ^2.

4.2.2 Inference and forecasting

Given posterior samples of model parameters fit through time T, it is straightforward to obtain a forecast distribution and other important quantities, including posterior uncertainty, for y_{T+1}. For each sample, one may calculate the first line of (4.2) over a grid of y_{T+1} values to estimate the one-step-ahead forecast distribution. Likewise, one may replace each distribution in (4.2) with conditional means to obtain the forecast mean. This procedure extends to transition mean and density estimates for any fixed values of inputs $(y_{t-1}, \ldots, y_{t-L})$ by evaluating (4.2) over a multidimensional grid of values for each posterior sample of model parameters.

Calculation of transition density and mean estimates requires values for each lag $(\{y_{t-\ell}\}_{\ell=1}^{L})$, regardless of inferences for $\boldsymbol{\lambda}$. However, one may be interested in these quantities conditional on a certain configuration of active lags. Suppose

that inference for $\boldsymbol{\lambda}$ in a model fit using $L = 3$ indicates that only the first two lags carry significant weight. They may specify a grid of values for the first two lags over which to evaluate (4.2), substitute dummy or default values, such as the mean, for y_{t-3}, and examine the transition density or mean as a function of y_{t-1} and y_{t-2} only. We caution that one should test the resulting inferences for sensitivity to the default values used for inactive lags before making conclusions. For example, one could replace mean values for inactive lags with random values drawn uniformly across the range of $\{y_t\}$.

Finally, one may make K-step-ahead forecasts by inductively simulating $(z, y)_{T+k}$ pairs, for $k = 1, \ldots, K$, following the first two levels of (4.3), for each posterior sample. The primary challenge here lies in the need to extend the $\{\boldsymbol{f}_\ell\}$ Gaussian process realizations to include the $f_\ell(y_{T+k-\ell})$ that do not already exist, for which a naive computation approach involves repeatedly inverting a growing covariance matrix. When repeated for each posterior simulation, this results in a computational burden commensurate with MCMC. Given a current model state (i.e., full sample of all model parameters) the procedure to draw $f_\ell(y_{T+k-\ell})$ begins by calculating $c_k = \kappa_\ell \, \sigma_\ell^2$, and $(\boldsymbol{c}_k)_i = \kappa_\ell \, \sigma_\ell^2 \, \rho(y_{T+k-\ell}, x_i; \nu, \psi_\ell)$ for all x_i associated with the entries $f_{i,\ell}$. Then using the existing $\boldsymbol{f}_\ell$, draw a realization $f_\ell(y_{T+k-\ell}) \sim \mathrm{N}\left(\boldsymbol{c}_k'(\boldsymbol{C}^{(\ell)})^{-1}\boldsymbol{f}_\ell, \; c_k - \boldsymbol{c}_k'(\boldsymbol{C}^{(\ell)})^{-1}\boldsymbol{c}_k\right)$, where $\boldsymbol{C}^{(\ell)}$ is the existing covariance matrix for $\boldsymbol{f}_\ell$. Lastly, concatenate $\boldsymbol{C}^{(\ell)}$ with c_k on the diagonal and $\boldsymbol{c}_k$ along an outer column and row, and concatenate $\boldsymbol{f}_\ell$ with the new draw from f_ℓ. One can avoid re-calculating the new $(\boldsymbol{C}^{(\ell)})^{-1}$ from scratch by storing the previous inverse and using the inversion formula for partitioned matrices (Rasmussen and Williams, 2006, p. 201).

4.3 Data illustrations

We demonstrate properties of the GPMTD model with two simulated and two real time series. The first simulation in Section 4.3.1 highlights lag selection and nonliear dynamics. The second simulation in Section 4.3.2 explores the model's fitness for approximating higher order dynamics in a time-delay embedding context. We then apply the GPMTD to a noisy time series known for non-Gaussian transitions in Section 4.3.3, and finally to a time series for which we anticipate a certain lag dependence structure in Section 4.3.4.

Each of the following analyses included at least three MCMC runs with chains initialized at default values (i.e., independent standard normal mixture components, uniform λ, and all observations allocated to the intercept). A Metropolis adaptation phase was followed by 5,000 burn-in iterations. A final run of 10,000 iterations was thinned to 2,000 inference samples (1,000 were used for some two-dimensional plots), which are reported for one chain. Unless otherwise reported, inferences for functionals of (4.2) with respect to fewer than L lags were obtained by inserting default mean values for inactive lags, which could be identified, for example, as $\{x_{t,\ell} : \mathrm{E}(\lambda_\ell \mid \{y_t\}) < c_\lambda\}$ for some small positive value c_λ (such as 0.01).

4.3.1 Simulated data: single lag

We first revisit the simulated time series introduced in Section 2.3.2, generated from

$$y_t = y_{t-2} \exp(2.6 - y_{t-2}) + \epsilon_t \,, \quad \epsilon_t \overset{\text{iid}}{\sim} \mathrm{N}(0, (0.09)^2) \,, \tag{4.5}$$

which features first-order nonlinear dynamics as a function of the second lag only. Previously, the time series was binned into discrete states from which the active

lag was inferred using the MTD model. We now fit the GPMTD model to the original real-valued time series with $L = 5$ and $T = 105$ (so that 100 observations contribute to the likelihood).

All three MCMC chains converge to the same region of the parameter space, although one earlier run showed a posterior mode with observations allocated to the fourth lag. This mode is not surprising given the cyclical behavior of even lags evident in Figure 2.5. Model inferences are decisive in favor of a single lag, with the 0.025 posterior sample quantile of λ_2 being greater than 0.99. The estimated transition mean as a function of y_{t-2} (holding other lags fixed), with 95% pointwise credible intervals, is shown in Figure 4.1 together with the data and true transition mean function. The dynamics are successfully recovered within the range of observed transitions, except on the far left, where the estimated curve tends back toward the component level μ_2 (which has posterior mean around 0.9, and standard deviation 2.4) in a smooth manner. This is likely influenced by the stationary covariance function and bias from the default prior on the component-specific

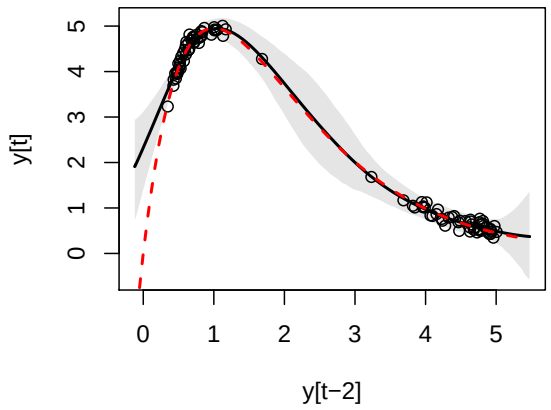


Figure 4.1: GPMTD fit to the single-lag dynamical simulation with noise. The solid black curve depicts the model estimate of the overall transition mean as a function of the second lag only, together with a 95% credible interval shaded in gray. The true transition mean function is given by the dashed red curve. All observed two-step transitions are included as points.

observation variance σ_2^2. One could argue that the stationary covariance function does not allow sufficient uncertainty in the central region ($y_{t-2} \in (2, 3)$) with no observed transitions.

4.3.2 Simulated data: time-delay embedding

Our second simulation example explores the GPMTD model's fitness for approximating higher-order dynamics. We do so with an example of statistical state-space reconstruction via time-delay embedding, which attempts to reconstruct a multidimensional attractor using lags from a single time series. The modeling objective for this example is to infer a suitable embedding dimension and estimate the corresponding transition map.

We proceed by first simulating a long time series (with sufficient burn-in) from the following two-dimensional deterministic system used to represent predator-prey dynamics with interaction (Basson and Fogarty, 1997),

$$\begin{aligned} y_t &= y_{t-1} \exp(r - a y_{t-1} - b z_{t-1}) \,, \\ z_t &= z_{t-1} \exp(r - a z_{t-1} + b y_{t-1}) \,, \end{aligned} \tag{4.6}$$

using $r = 2.75$, $a = 0.5$, and $b = 0.07$. In this case, substitution yields an analytical expression for a time-delay embedding of this system in two lags using either the $\{y_t\}$ or $\{z_t\}$ series alone. The resulting transition surface for the $\{y_t\}$ series is more regular when we consider the dyanamics on the $\log(y)$ scale, a natural transformation given that the model applies to non-negative-valued species abundance. Figure 4.2 shows the original transition surface from the first line of (4.6) with the generated time series points overlaid. A trace for 100 successive values of $\log(y_t)$ is shown in Figure 4.3. Figure 4.4 shows the time-delay embedding

transition surface for $\log(y)$, one of the inferential targets in this example. Note the outer "wall" corresponding to super-exponential growth just outside the observed data range. The continuous surface changes directions quickly at this border, with points falling on both sides of a steep and narrow trench.

It is immediately apparent that the GPMTD model is inadequate to fully capture this non-additive, intricate function of two lags. If the model admitted general functions of two inputs, or at least interactions, one could enforce near determinism with the priors on component-specific variances $\{\sigma^2_\ell\}$. This practice is discouraged with the GPMTD (unless the modeler is confident that only one lag

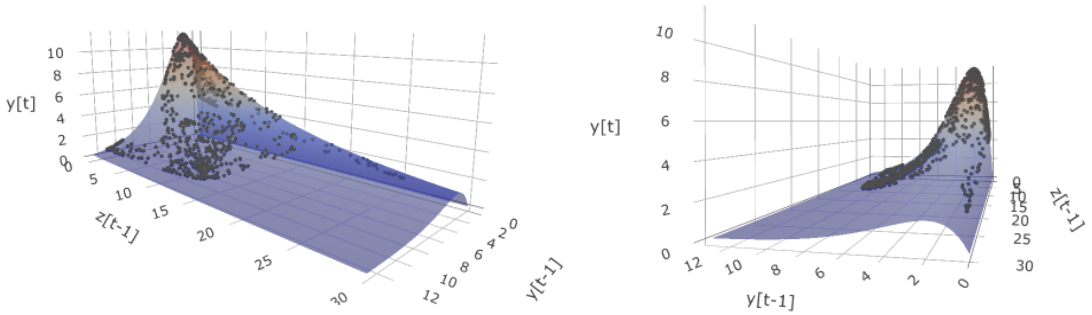


Figure 4.2: Transition surface from the deterministic nonlinear system (4.6). Simulated values are included as points on the surface. Multidimensional plots were generated with *Plotly* (Plotly Technologies Inc., 2015).

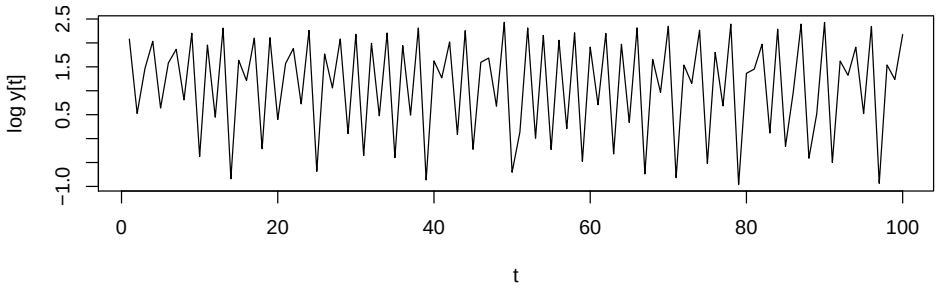


Figure 4.3: Trace of 100 steps of the log-transformed y_t series from the simulated deterministic nonlinear system.

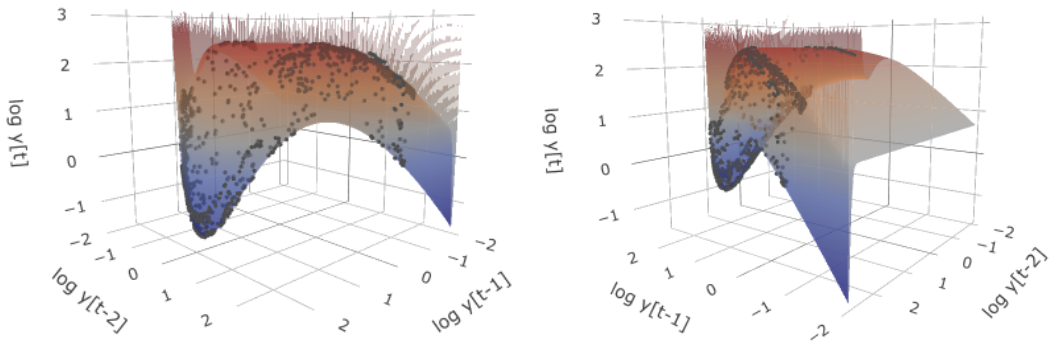


Figure 4.4: Transition surface for the time-delay embedding of $\log(y)$ from the nonlinear deterministic system (4.6). Simulated values are included as points on the surface.

is active) for two reasons. The first is that the model will attempt to interpolate apparent noise resulting from projection into one dimension, which occurs in this example. Second, the mixture of densities defining the model will produce multiple highly separated modes for most combinations of lag values. For these reasons, we forego pursuing a high-fidelity estimate of the transition surface with the GPMTD, allowing for observation noise to "smooth" over some finer features of the surface.

As before, we fit the GPMTD model with default priors and initial values to a $\{\log(y_t)\}$ series of length $T = 105$ and $T = 505$ using a lag horizon of $L = 5$. All three chains converge to similar log-likelihood values and lag configuration for the shorter time series. Two of the three chains likewise converge for the longer series, while one chain remains stuck at a mode with significantly lower log-likelihood.

The model fit to the shorter time series decisively selects lag 1 only (with a 0.025 posterior sample quantile above 0.99), which appears reasonable given the sample size and the fit depicted in Figure 4.5. The model fails to capture only a few points in the border trench along $\log(y_{t-1}) \in (-0.5, 0.5)$, $\log(y_{t-2}) \approx 2.3$. Because these observations are not allocated to another mixture component and treated as outliers, the component-specific standard deviation (effectively the global error

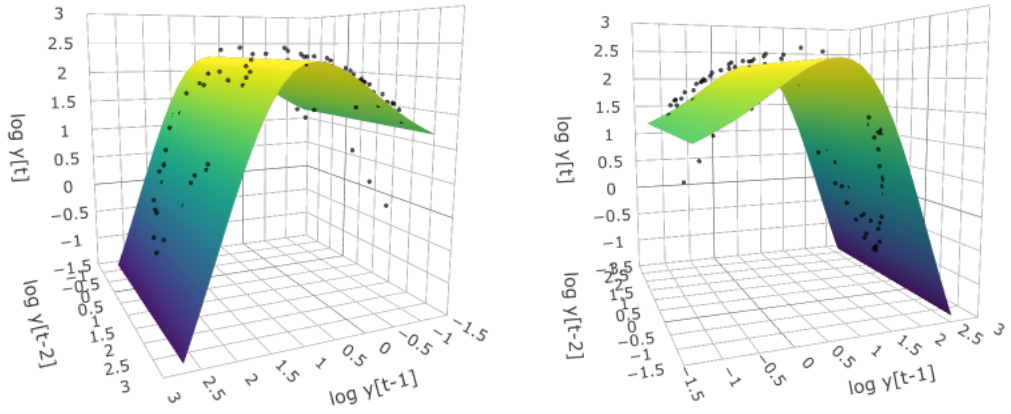


Figure 4.5: GPMTD model fit ($T = 105$ and $L = 5$) to the time-delay embedding of $\log(y)$ simulated from the nonlinear deterministic system. Plots include the posterior mean estimate for the transition surface and observed transitions as points.

standard deviation since $\lambda_1 \approx 1$) is estimated high at 0.4.

The model fit to the longer time series provides a surprisingly robust approximation, considering the level of model mis-specification. Two lags are selected, with λ_1 and λ_2 receiving a 0.78, 0.22 split in posterior mean (both 95% intervals have approximate length 0.12). The posterior mean estimate of the transition surface is given in Figure 4.6, together with marginal estimates of f_1 and f_2 and their assigned observations (classified if the observations are assigned to the corresponding lag with at least 0.5 posterior probability). The most obvious omission in the estimated surface is the outer wall or border. This is expected, as the lower trench is not clearly identified in one dimension. Assuming noisy observations, f_1 and f_2 fit the corresponding one-dimensional projections well, while the overall estimated transition surface appears attenuated, a result of the global mixture. Similarly, transition density estimates (not shown) for lag values along the two shoulders and central dip of the surface are bimodal with small variances. The second-lag component successfully captures the "outliers" near $\log(y_{t-1}) \approx -0.75$,

$\log(y_{t-2}) \approx 2.3$, producing an appropriate mixture transition density in this region.

Overall, we caution that despite its ability to produce GAM-like estimates for transition surfaces, the lag-dependent error structure of the GPMTD model is not suited to this application for high-order dynamics. The model is better poised to estimate possibly nonlinear, lag-dependent transition densities in the presence of noise. Such a scenario is presented with the two examples that follow.

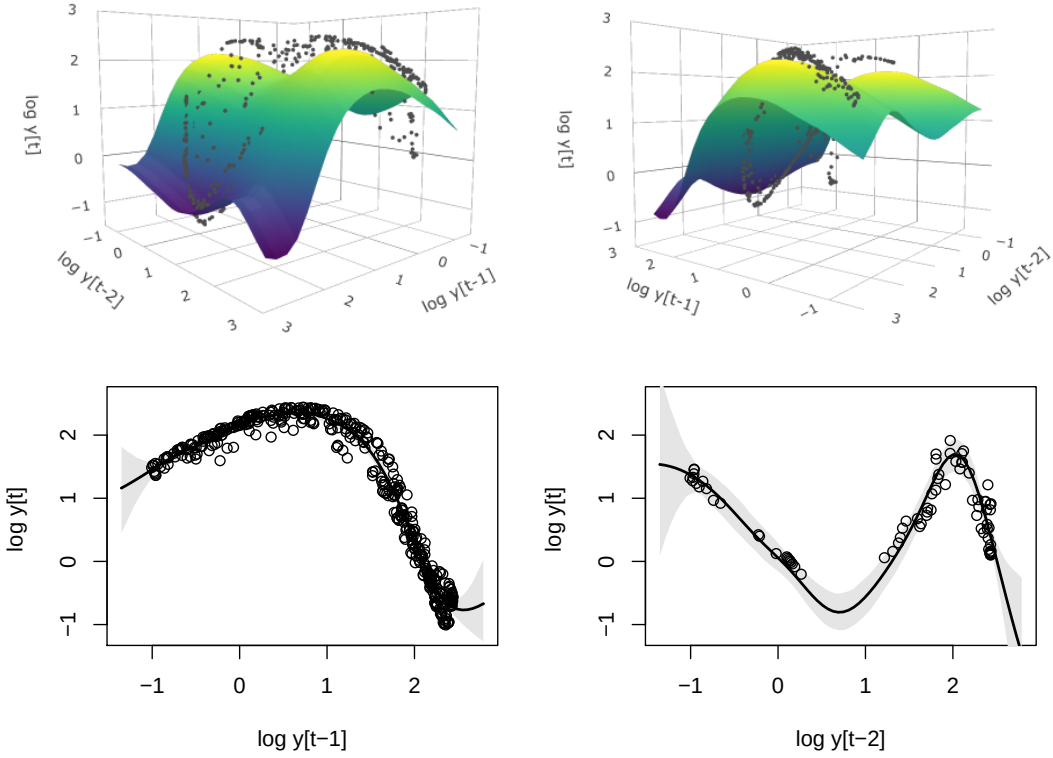


Figure 4.6: GPMTD model fit ($T = 505$ and $L = 5$) to the time-delay embedding of $\log(y)$ simulated from the nonlinear deterministic system. Posterior mean estimate for the transition surface (top) and lag-specific f_1 and f_2 functions (with pointwise 95% intervals, bottom). Data values are included as points. In the lower plots, points are included with a lag if allocated to that lag (with posterior probability greater than 0.5).

4.3.3 Old Faithful data

Our first illustration of the GPMTD with real data uses the well-known environmental time series from the Old Faithful geyser in Yellowstone National Park, U.S.A. Scientists and park authorities have recorded eruption durations and inter-eruption waiting times in order to better understand the geyser mechanism and make accurate predictions. Forecasting eruption time has proven challenging, prompting speculation that Old Faithful is a nonlinear chaotic system. Indeed, Nicholl et al. (1994) reach this conclusion. Historically, eruption durations have provided the most accurate predictions of the subsequent eruption times. However, Raye (2005) uses lags of waiting times only to make comparable predictions. Azzalini and Bowman (1990) follow a statistical approach, concluding that a second-order Markovian model appropriately captures the dominant signal.

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 4.7 shows a trace of eruption waiting times in minutes, together with a pair of scatter plots of waiting times against the first two lags. Despite high noise levels, dependence on at least one lag is clearly discernable. The relationship between consecutive waiting times appears mostly consistent across values of the second lag, but a trend may exist. The vanilla GPMTD model is unlikely to detect higher-order dynamics, which we revisit in Chapter 5. We do, however, expect the model to capture the nonlinear and non-Gaussian features apparent in Figure 4.7.

We fit the GPMTD model with $L = 5$ and $L = 10$. All chains converge to the same region in the parameter space, with exception of one run with $L = 5$ that switched to an allocation with some observations assigned to the second lag. We report results from one of the $L = 10$ runs. Model inferences are decisive in favor

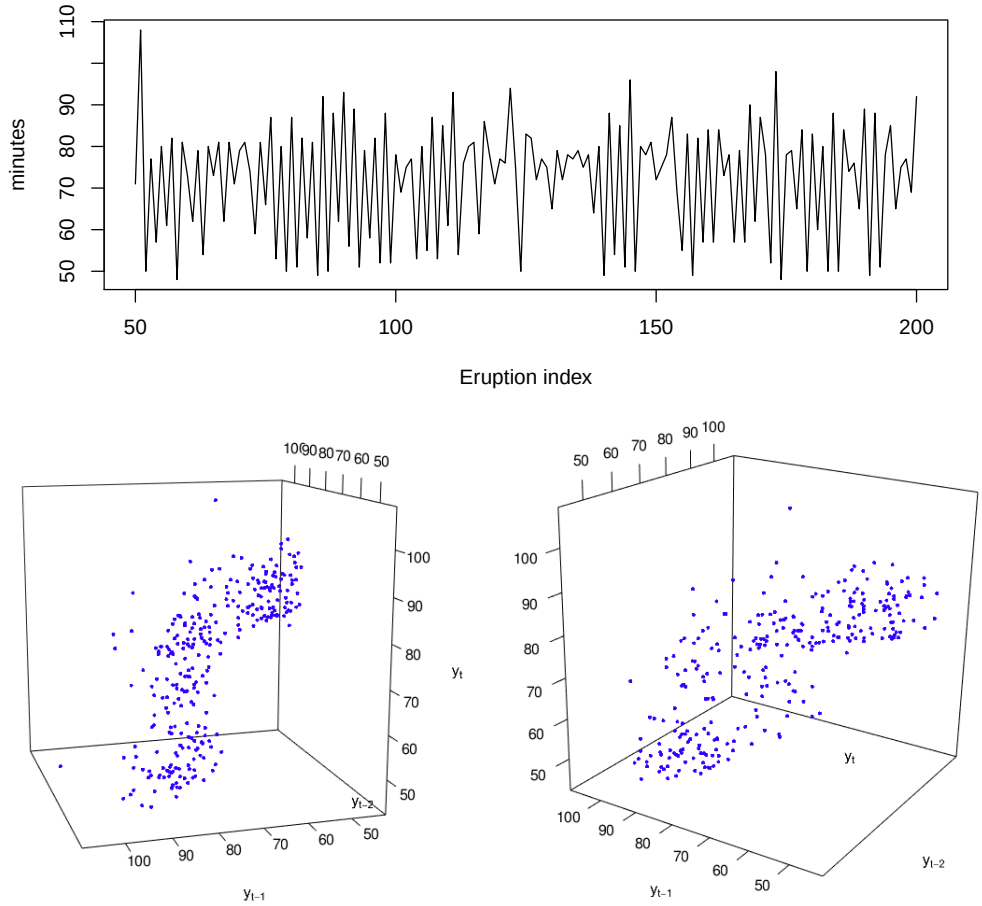


Figure 4.7: Trace of 150 consecutive Old Faithful eruption waiting times in minutes (top). This window of the middle half of the time series typifies the data, with exception of the run of long waiting times between index 120 and 140. The scatter plots (bottom) show waiting times in minutes against the first two lags for the full time series.

of a single lag, with λ_0 and λ_1 accounting for more than 99% of the allocation in the posterior mean of $\boldsymbol{\lambda}$. Point estimates and 95% credible intervals for each λ_ℓ are reported in Table 4.1. The intercept carries significant weight in order to provide bimodality in the transition distribution, while the first lag component captures nonlinear dependence. This trade-off is depicted in Figure 4.8 with a lag-1 scatter plot, where model-based lag allocation is indicated by color. Blue points are assigned to the intercept component with posterior probability greater than

ℓ	Mean	95% Interval
0	0.428	(0.332, 0.512)
1	0.571	(0.486, 0.666)
2	<0.001	(<0.001, 0.002)
3	<0.001	(<0.001, 0.001)
4-10	<0.001	(<0.001, <0.001)

Table 4.1: Posterior summary for $\lambda_\ell,\ \ell = 0, \ldots, 10$ in the GPMTD analysis of Old Faithful waiting times. Lag $\ell = 0$ refers to the intercept.

0.5, and red points are likewise assigned to the first lag. The solid red and blue curves give posterior mean inferences for the respective component means. The solid black curve depicts our pointwise estimate of the transition mean functional, together with a 95% credible interval shaded in gray. The transition mean is less useful for lagged values above 70 minutes, where it begins to straddle the bimodal transition density.

Figure 4.9 summarizes posterior inferences for transition densities for three values of the first lag $y_{t-1} \in \{50, 66, 80\}$. Because the more concentrated density associated with lag 1 is located above the mean of the wide intercept density at $y_{t-1} = 50$, the model incorrectly yields a left-skewed density for $y_{t-1} = 50$. One could argue from Figure 4.8 that the transition density at this lag should exhibit right skew. The means of mixture components $\ell = 0$ and 1 intersect near $y_{t-1} = 66$, appropriately resulting in a scale mixture of normal distributions for the transition. At $y_{t-1} = 80$, the mixture captures the obvious bimodality.

High levels of noise, together with nonlinear lag dependence, make the Old Faithful time series an interesting candidate for illustrating both strengths of the GPMTD model. When the model employs mixing for both flexible density autoregession *and* nonlinear transition surfaces simultaneously, we entreat practitioners to carefully scrutinize and validate inferences. For example, because the mixing weights are global, they may not be optimized for the transition density

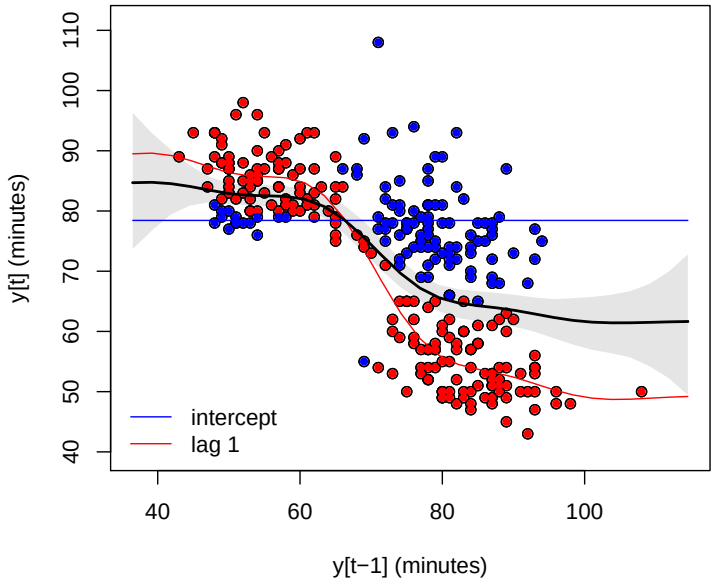


Figure 4.8: Single-step transition scatter plot with component-specific inferences from the GPMTD fit to Old Faithful waiting times. Blue points indicate membership in the intercept mixture component (with posterior probability greater than 0.5), and red points indicate the same for the first lag mixture component. The solid red and blue curves report the posterior mean for the respective component means. The solid black curve depicts the model estimate of the overall transition mean, together with a 95% credible interval shaded in gray.

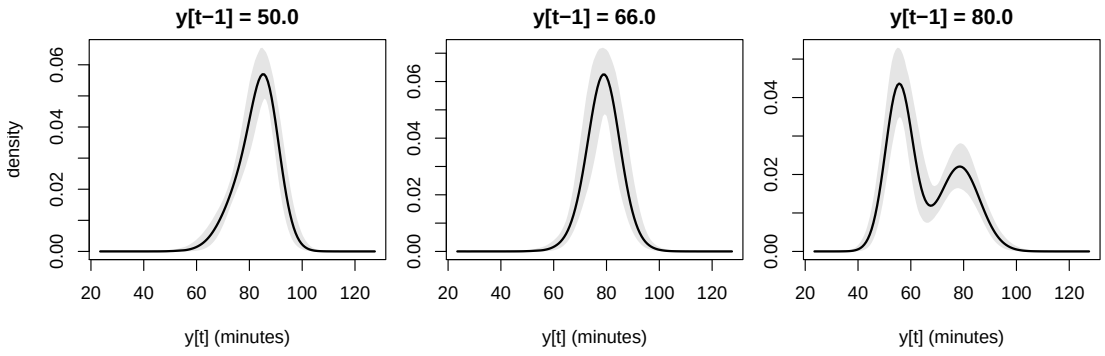


Figure 4.9: GPMTD transition density estimates for Old Faithful waiting times at three fixed values of the first lag: $y_{t-1} = 50$, $y_{t-1} = 66$, and $y_{t-1} = 80$ minutes. The solid line indicates the pointwise posterior mean and gray shading indicates 95% intervals.

specifically at $y_{t-1} = 80$ in the Old Faithful model. While the parsimonious representation (4.2) has such limitations, it is quite flexible relative to the mixtures of linear autoregressive models in the literature, efficiently capturing nonlinear and non-Gaussian lag-dependent dynamics.

4.3.4 Pink salmon data

We now apply the model to the pink salmon data introduced in Section 3.5.2, which consist of annual escapement of pink salmon in a stream in Alaska from 1934 to 1963. We expect the two-year life cycle of pink salmon to drive serial dependence, which was corroborated by MMTD model fit to a discetized version of the time series shown in Figure 3.2.

We fit the GPMTD with up to $L = 5$ lags to the logarithm of annual escapement using the same default prior, initialization, and MCMC sampling employed for other analyses. All chains converge to the same estimated posterior distributions. Not surprisingly, λ_2 has a posterior mean of 0.975 with a 95% equal-tailed interval of (0.683, 0.999). Lags 1 and 4 have the next highest upper (0.975) quantiles at 0.095 and 0.046, respectively. The estimated transition mean function with pointwise 95% intervals is given for the second lag (fixing other lags) in Figure 4.10. The diagonal dotted line has a unit slope dividing regions of population increase and decrease. Although the interval seldom leaves this line, population decline is readily apparent, particularly with the even-year population (in Figure 3.2), which experienced repeated interventions throughout the 1950s and early 1960s (Bradshaw and Heintz, 2003). We have not attempted to model additional covariates or interventions, but this series demonstrates the important feature of lag selection in the GPMTD model.

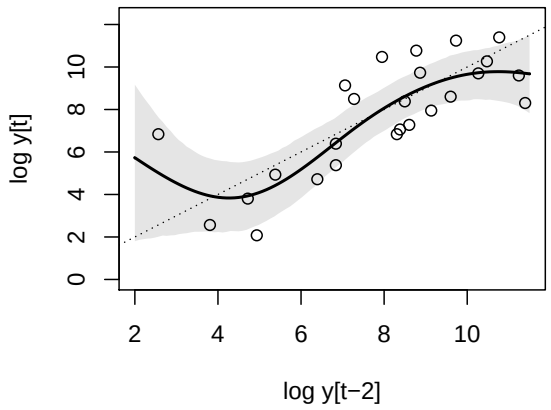


Figure 4.10: GPMTD fit to the logarithm of annual pink salmon escapement, with a scatter plot of all two-step transitions. The solid black curve gives the overall transition mean, together with a 95% credible interval shaded in gray. The reference line has unit slope and passes through the origin.

4.4 Extensions to the GPMTD

In this section, we motivate and propose two extensions of the GPMTD model. The first relaxes Gaussianity of each mixture component and the second relaxes additivity to allow higher-order interactions necessary for an important application of the GPMTD. We provide details for implementation and discussion.

4.4.1 Mixture components with long tails and skew

The intercept component in (4.2) is instrumental for the Old Faithful example in that it provides a vehicle both for bimodality (when $y_{t-1} > 70$ minutes) and a pair of outliers (at $y_{t-1} \approx 70$ minutes). As noted in Section 4.1, Le et al. (1996) also use an independent component for outliers. If, however, certain characteristics of the transition distribution systematically associate with a certain lag, it is more appropriate to accommodate them in the corresponding mixture component. Adding flexibility to the mixture component distributions further helps disentangle two model objectives: transition density estimation through mixtures and lag

selection. If the mixture is used primarily for lag selection, our method relates to Hansen (1994), who explores using parametric extensions of Gaussian transition densities. In this section, we apply two standard extensions aimed at increasing flexibility without sacrificing parsimony or computational convenience.

We first consider allowing long-tailed component distributions. Using the student t distribution's well-known representation as a scale mixture of Gaussian distributions, we introduce independent latent variables $\{\varphi_{\ell,t}\}$, associated with each component and observation, and distributed gamma with shape $\eta_\ell/2$ and rate $\eta_\ell/2$. The variance of each component in the first line of (4.3) becomes $\sigma_\ell^2/\varphi_{\ell,t}$. We complete the specification with independent gamma priors for $\eta_\ell - 2$, for $\ell = 0, \ldots, L$, ensuring two finite moments in the mixture components. This extension preserves Gaussianity of full conditional updates in the Gibbs sampler with the following minor changes. The identity matrices that appear in $\boldsymbol{W}$ and in $\boldsymbol{\Sigma}$ in Step 4 of the component-specific Gibbs scan in Appendix C.1 are replaced with $\text{diag}(\varphi_{\ell,t_{i1}}^{-1}, \ldots, \varphi_{\ell,t_{in_\ell}}^{-1})$. A fifth and sixth step are added to the Gibbs scan to update each $\varphi_{t,\ell}$ (conditionally conjugate gamma) and η_ℓ (non-conjugate, updated with Metropolis or with a discrete prior). Steps 1, 3, and 4 of the full Gibbs sampler in Appendix C.2 likewise reflect observation-specific variance scaling by $\{\varphi_{\ell,t}\}$, but retain their basic forms.

We next admit skewness in addition to long tails. We employ the scale mixture of skew-normal distributions of Cancho et al. (2011), who develop a framework for Bayesian inference in nonlinear regression with skewed and/or long-tailed errors. The skew-normal distribution derives from the construction of Azzalini (1985). Random variable Y is said to follow the skew-normal distribution if it has density

$$\phi(y \mid \mu, \sigma^2, \xi) = 2\phi\left(\frac{y-\mu}{\sigma}\right)\Phi\left(\xi\frac{y-\mu}{\sigma}\right), \tag{4.7}$$

where $\phi(\cdot)$ is the standard Gaussian density function, $\Phi(\cdot)$ is the standard Gaussian cumulative distribution function, and the parameter $\xi \in \mathbb{R}$ influences skewness. A stochastic representation for Y (Henze, 1986) facilitates modeling with the skew-normal distribution, for which significant development in the last few decades includes both nonlinear regression and mixture models (Lin et al., 2007).

Cancho et al. (2011) induce a scale mixture by including a latent positive-valued random variable φ, which for our purposes will have the same gamma distribution introduced above, and replacing all instances of σ in (4.7) with $\sigma/\sqrt{\varphi}$. Integrating (4.7) with respect to the density of φ produces the scale mixture of skew normal distributions, which in our case is a skew-t with η degrees of freedom. They report the stochastic representation as $Y = \mu + \Delta V + \sqrt{\tau/\varphi}\, V_1$, where $\Delta = \sigma\,\delta$ with $\delta = \xi/\sqrt{1+\xi^2}$, $\tau = \sigma^2(1-\delta^2)$, $V = |V_0|/\sqrt{\varphi}$, and V_0 and V_1 are independent standard Gaussian random variables. Setting $\xi = 0$ results in a scale mixture of normal distributions, while fixing $\varphi = 1$ produces skew only. This parameterization yields conditional conjugacy and thus convenient posterior sampling if we specify independent Gaussian and inverse-gamma priors for Δ and τ, respectively.

Conditional on allocation membership $z_t = \ell$, which we omit from the notation for simplicity, the modified contribution of y_t to the GPMTD is given in generative order as

$$\begin{aligned} \varphi_t \mid \eta &\sim \mathrm{Ga}\left(\eta/2, \eta/2\right) , \\ p(V_t \mid \varphi_t, \eta) &\propto \mathrm{N}\left(V_t \mid b_v, \varphi_t^{-1}\right) 1_{(V_t > b_v)} , \\ y_t \mid V_t, \varphi_t, \mu, f, \sigma^2, \eta &\sim \mathrm{N}\left(\mu + f(x_{t,\ell}) + \Delta V, \tau/\varphi\right) , \end{aligned} \tag{4.8}$$

where $b_v = -\sqrt{\eta/\pi}\,\Gamma\left([\eta-1]/2\right)/\Gamma\left(\eta/2\right)$. The shift by b_v ensures that the component has mean $\mu + f(x_{t,\ell})$ (Cancho et al., 2011). Note that the $(\varphi_t, V_t, \mu, f, \Delta, \tau, \eta)_\ell$ tuple is specific to mixture component ℓ. This setup admits a sampling scheme

similar to that given in Appendix C.1, in which matrix $\boldsymbol{W}$ again includes $\{\varphi_t\}$ and $\mathbf{1}\mu$ is replaced with a linear regression form in which $\{V_t\}$ populates the second column of a design matrix with coefficient vector (μ, Δ). The full conditional distributions for φ_t and V_t are gamma and truncated normal, respectively (Cancho et al., 2011). Because σ^2 is coupled with other parameters, the Gaussian process variance $\kappa\,\sigma^2$ is replaced with a single, unrelated variance parameter.

4.4.2 Higher-order interactions with the GPMMTD

Allowing smooth nonlinear functions to the mixture means adds significant flexibility beyond the models in Le et al. (1996). However, using single-lag dependence in each mixture component of (4.2) limits the scope of the GPMTD primarily to nonhomogeneous mixture error distributions and/or selection of a single relevant lag. Although the transition mean resulting from (4.2) is additive and can approximate functions with mild interactions of inputs, the mixture of densities (rather than means) can result in the transition mean function falling in a region of low density, as seen in the time-delay embedding example of Section 4.3.2. To accurately model high-order transition surfaces with low noise in the GPMTD framework, it becomes necessary to admit f functions of multiple lags.

The most common approach to modeling functions with several inputs in the Gaussian process regression framework involves separable covariance functions, composed through the product of single-variable correlation functions (Rasmussen and Williams, 2006). This covariance structure naturally arises when one considers the process resulting from the product of independent functions, each modeled with an independent GP. The length-scale parameter can also serve as a proxy for variable selection in automatic relevance determination (ARD, Neal, 1996). We present an alternative to ARD in the context of lag selection with an extension

of the GPMTD model that follows the mixture of mixture transition distribution construction of Chapter 3. For positive integer $R < L$, the general model is given by

$$
\begin{aligned}
F_t(y_t \mid y_{t-1}, \ldots, y_1) = \Lambda_0 \, \mathrm{N}\left(y_t \mid \mu_0, \sigma_0^2\right) + \\
\Lambda_1 \sum_{\ell=1}^{L} \lambda_\ell^{(1)} \, \mathrm{N}\left(y_t \mid \mu_\ell + f_\ell^{(1)}(y_{t-\ell}), \sigma_{1,\ell}^2\right) + \\
\Lambda_2 \sum_{1 \le \ell_1 < \ell_2 \le L} \sum \lambda_{(\ell_1,\ell_2)}^{(2)} \, \mathrm{N}\left(y_t \mid \mu_{\ell_1,\ell_2}^{(2)} + f_{\ell_1,\ell_2}^{(2)}(y_{t-\ell_1}, y_{t-\ell_2}), \sigma_{2,\ell_1,\ell_2}^2\right) + \ldots + \\
\Lambda_R \sum_{1 \le \ell_1 < \ldots < \ell_R \le L} \cdots \sum \lambda_{(\ell_1,\ldots,\ell_R)}^{(R)} \, \mathrm{N}\left(y_t \mid \mu_{\ell_1,\ldots,\ell_R}^{(R)} + f_{\ell_1,\ldots,\ell_R}^{(R)}(y_{t-\ell_1}, \ldots, y_{t-\ell_R}), \sigma_{R,\ell_1,\ldots,\ell_R}^2\right) ,
\end{aligned}
\tag{4.9}
$$

with $\lambda_{(\ell_1,\ldots,\ell_r)}^{(r)} \in \boldsymbol{\lambda}^{(r)}$, a probability vector of length $\binom{L}{r}$, for $r = 1, \ldots, R$. We propose to simplify (4.9) by specifying only one set of parameters (μ, f, σ^2) for all mixture components at a given level $r > 1$, in conjunction with a sparse Dirichlet mixture (SDM) prior for each $\{\boldsymbol{\lambda}^{(r)} : r = 2, \ldots, R\}$ to favor selection of only one component. Retaining the SBM prior on $\boldsymbol{\Lambda} = (\Lambda_0, \Lambda_1, \ldots, \Lambda_R)$ and $\boldsymbol{\lambda}^{(1)}$ retains characteristics of the GPMTD and encourages shrinkage of the over-specified model. We will refer to the simplified specification as the canonical Gaussian process mixture of mixture transition distributions (GPMMTD) model.

As with the original MMTD, the model in (4.9) can be "flattened" to a corresponding GPMTD representation with higher-order functions $\{f\}$. Increasing the number of arguments to each f represents the primary innovation and challenge of the GPMMTD model. If we select R to be modest and significantly below the lag horizon L, we can specify more general covariance functions than the standard isotropic or separable ARD kernels. One option relaxes isotropy (but retains stationarity) by replacing the Euclidean distance input of the squared exponential or Matérn correlation function with a Mahalanobis distance $d =$

$\sqrt{(\boldsymbol{x}-\boldsymbol{x}')^\top \boldsymbol{\Psi}(\boldsymbol{x}-\boldsymbol{x}')}$ for positive-definite matrix $\boldsymbol{\Psi}$ (as in Ecker and Gelfand, 1999). That the parameters in $\boldsymbol{\Psi}$ are not identifiable and grow quadratically with the dimension of $\boldsymbol{x}$ can be addressed with informative priors and maintaining a moderate value of R.

The hierarchical formulation of the GPMMTD again utilizes augmentation with latent variables $Z_t \in \{0, 1, \ldots, R\}$ and $\boldsymbol{z_t} \in \cup_r\{(\ell_1, \ldots, \ell_r) : 1 \leq \ell_1 < \ldots < \ell_r \leq L\}$, and is given as follows. For $t = L+1, \ldots, T$; $\ell = 1, \ldots, L$, $1 \leq \ell_1 < \ldots < \ell_r \leq L$; and $r = 0, 1, \ldots, R$, we have

$$
\begin{aligned}
y_t \mid Z_t, \boldsymbol{z}_t, \{(\mu, \sigma^2, f)\} &\overset{\text{ind.}}{\sim} \begin{cases} \mathrm{N}\left(\mu_0, \sigma_0^2\right) & \text{if } Z_t = 0, \\ \mathrm{N}\left(\mu_{z_t}^{(1)} + f_{z_t}^{(1)}(x_{t,z_t}), \sigma_{1,z_t}^2\right) & \text{if } Z_t = 1, \\ \mathrm{N}\left(\mu^{(Z_t)} + f^{(Z_t)}(\boldsymbol{x}_t(\boldsymbol{z}_t)), \sigma_{Z_t}^2\right) & \text{if } Z_t > 1, \end{cases} \\
&\text{for } t = L+1, \ldots, T, \\
\Pr(Z_t = r \mid \boldsymbol{\Lambda}) = \Lambda_r, \quad \Pr(\boldsymbol{z}_t = (\ell_1, \ldots, \ell_r) \mid Z_t = r, \boldsymbol{\lambda}^{(r)}) &= \lambda^{(r)}_{(\ell_1, \ldots, \ell_r)}, \\
&\text{for } r = 0, 1, \ldots, L, \text{ independently for } t = L+1, \ldots, T, \\
\boldsymbol{\Lambda} &\sim \mathrm{SBM}(\boldsymbol{\pi}_1, \boldsymbol{\pi}_3, \eta, \boldsymbol{\gamma}, \boldsymbol{\delta}), \qquad (4.10) \\
\boldsymbol{\lambda}^{(1)} \sim \mathrm{SBM}(\eta_\lambda, \pi_{1\lambda}, \pi_{3\lambda}, &\boldsymbol{\gamma}_\lambda, \boldsymbol{\delta}_\lambda), \quad \boldsymbol{\lambda}^{(r)} \overset{\text{ind.}}{\sim} \mathrm{SDM}(\boldsymbol{\alpha}_{\lambda^{(r)}}, \beta_{\lambda^{(r)}}) \quad \text{for } r = 2, \ldots, R,
\end{aligned}
$$

where $\boldsymbol{x}_t(\boldsymbol{z}_t)$ refers to the elements of $\boldsymbol{x}_t$ selected by $\boldsymbol{z}_t$, and the top level implicitly conditions on lags in $\boldsymbol{x}_t$. The remaining hierarchical structure follows with (4.3), except that $\psi^{(r)}$ for $r > 1$ is replaced by a vector of parameters used to construct $\boldsymbol{\Psi}^{(r)}$ through, for example, some factorization that preserves positive definiteness.

Most time-series analyses are conducted at the data-sampling time step without further consideration of sampling frequency, which can substantially affect inferences and forecasts. For example, sampling at high frequency and analyzing relatively few lags can cause nonlinear dynamics to appear linear. Thus, in addition

to dependence order, an optimal time delay is sometimes sought in time-delay-embedding applications (Kantz et al., 2004, p. 38). Joint inference for both is easily accommodated if we consider a trivial and simplifying adjustment to the MMTD framework. As before, the outer sum in (4.9) mixes over possible orders. However, instead of enumerating all lag combinations of order r up to horizon L, the inner sum could index spacing among the r lags so that component ℓ corresponds to the lag set $(\ell, 2\ell, \ldots, r\ell)$. For example, when $r = 2$, the first component corresponds to lags $(1, 2)$, the second component corresponds to lags $(2, 4)$, the third to $(3, 6)$, and so forth. This arrangement is less general in that it assumes all lags (at the selected sampling frequency) up to the active order are relevant. Its estimation also requires a longer time series. However, it can substantially reduce the number of components under consideration and provide a parsimonious vehicle for modeling long-range dependence.

4.5 Discussion

The models proposed in this chapter, in addition to providing a Bayesian implementation for the continuous-state GMTD, contributes three possible extensions to the original framework: 1) nonlinear transition dynamics, 2) model-based order and lag selection, and 3) higher-order interactions (through the GPMMTD). Although the original GMTD accommodates non-Gaussian transition distributions, we further allow the mixture kernels to exhibit nonlinear lag dependence and non-Gaussian features. Thus the Gaussian process mixture transition distribution model can be considered a parsimonious, semiparametric model for nonlinear transition density estimation. Additionally, it can be used to identify nonlinear dynamics with low noise in one (GPMTD) or several (GPMMTD) lags.

The GPMTD model is inherently Markovian, directly modeling a probability

distribution governing transitions. It consequently most naturally resides in the class of time-series models for dynamical (rather than measurement) error. It can nevertheless be extended, in a state-space framework or otherwise, to include other features common in time series, such as covariate dependence, trends, and periodic fluctuations. The most straightforward way to incorporate covariates is through additional mixture components dedicated to the exogenous variables. As this breaks the natural ordering of mixture components, one would need to reconsider the prior for $\boldsymbol{\lambda}$. Incorporating trends, periodicity, and covariates outside the MTD structure presents more of a challenge, as these would most naturally fit into a linear superposition with a latent GPMTD process. Estimation of GPMTD parameters would be no more complicated in such a model. However, updates for parameters governing external structures would necessitate re-evaluation of the GPMTD component-mean functions $\{f_\ell\}$ at each iteration of MCMC (or optimization), potentially creating a heavy computational burden.

The mixture autoregressive (MAR) model of Wong and Li (2000), consisting of a finite mixture of Gaussian AR models of potentially varying order, is considered to be the sequel to the GMTD in the literature. The MAR model indeed contains the GMTD as a special case if we set most AR coefficients equal to zero. It however does not generalize the linear transition mean, and perhaps more importantly, diverges from the parsimonious and interpretable representation as a mixture of low-order transition distributions. We have proposed and demonstrated a model that preserves these important and distinguishing characteristics of the MTD and GMTD models. Nevertheless, the MAR and related mixtures of autoregressive models provide an important foundation for a rich class of Markovian models explored in Chapter 5.

Chapter 5

Bayesian Nonparametric Density Autoregression

5.1 Introduction

All models in previous chapters employ mixing distributions that are based on independent processes. While convenient for computation and inference for the transition distribution, this approach limits model flexibility. In the case of continuous state spaces, we have demonstrated that Gaussian process priors in the MTD component conditional means can allow for simple lag-dependent density estimation and nonlinear dynamics. In this chapter, we simplify the mixture kernels and instead use dependence in the mixture weights to accommodate nonlinear dynamics. Importantly, serial dependence in the weights (or latent process) is driven by observed variables, as summarized in Figure 1.1. We further break from the MTD framework by allowing kernel dependence on all lags (up to the specified horizon), and move to a Bayesian nonparametric framework, admitting countable mixtures. In this section, we briefly review nonlinear time series methods utilizing

mixtures, Bayesian nonparametric (BNP) methods, including regression, BNP applications to time-series problems, and approaches to order and lag selection.

5.1.1 Nonlinear time series via mixtures

While the term nonlinearity has been used to describe various qualitative characteristics of time series, we specifically refer to nonlinear dynamics, or nonlinearity in the function mapping past observations to the present. References for nonlinear autoregressive models, including generalized additive models and Gaussian process priors, are found in Section 4.1. Dependent mixtures through hidden Markov models, also capable of capturing nonlinear dynamics, are briefly discussed in Section 1.1.1.

The class of threshold autoregressive models (Tong, 1990) provide a parsimonious approximation of nonlinear dynamics, and have remained popular for decades. Although these can be formulated as finite mixtures of linear autoregressive models, the most popular form switches among a finite set of regressions as a deterministic function of a single lag, yielding a piecewise-linear transition function. Mixtures-of-experts (MoE) models (Jordan and Jacobs, 1994; Peng et al., 1996; Carvalho and Tanner, 2005, 2006) are closely related. They have the form

$$p(y_t \mid y_{t-1}, \ldots, y_{t-L}) = \sum_{j=1}^{J} q_j(y_{t-1}, \ldots, y_{t-L}; \varphi_j)\, k_j(y_t \mid y_{t-1}, \ldots, y_{t-L}, \phi_j)\,,$$

where parameterized weight functions $q_j(\cdot; \varphi_j)$ of past values provide probabilistic thresholding, usually through a link function, to activate "expert" kernel models $k_j(\cdot \mid \phi_j)$. In the case of Gaussian experts, these kernels have linear autoregressive means, and nonlinearity is provided by the weight functions. More similar to the model we propose, Glasbey (2001) and Kalliovirta et al. (2015) use normalized

kernel functions for local weighting in place of link functions on linear combinations of lags.

5.1.2 Dirichlet process mixtures

All mixture models to this point have been finite, with the number of components fixed at the number of lags under consideration. Because we now rely on the mixture structure to provide flexibility, we use a Bayesian nonparametric framework, for which the theoretical number of mixture components is infinite and the number of active components is random.

We rely on the Dirichlet process (Ferguson, 1973), which provides a prior for the random mixing distribution, G, for some generic parameter $\theta \in \Theta$. We say G follows a Dirichlet process and write $G \sim \mathrm{DP}(\alpha, G_0)$, where α is a positive scalar concentration parameter and G_0 is a base probability measure with support on Θ. We focus particularly on the stick-breaking representation of this process (Sethuraman, 1994), wherein the random probability measure can be written as

$$G(\theta) = \sum_{h=1}^{\infty} \omega_h \, \delta_{\theta_h^*}(\theta) \, . \tag{5.1}$$

Here, $\delta_{\theta_h^*}(\cdot)$ is a dirac-delta measure, or atom, at θ_h^*, $\theta_h^* \overset{\text{iid}}{\sim} G_0$, and $\{\omega_h\}$ arise from the stick-breaking process involving latent beta random variables specified in Section 5.2.1. Convolution of (5.1) with a continuous likelihood kernel density $k(y \mid \theta)$ results in a Dirichlet process mixture (DPM) model (Antoniak, 1974) suitable for density estimation.

A common approach extending the DP to accommodate covariates $x \in \mathcal{X}$ is the dependent Dirichlet process (DDP, MacEachern, 2000). The extension utilizes the stick-breaking representation by defining stochastic processes $\{\theta_h^*(x) :$

$h = 1, \ldots, \infty,\ x \in \mathcal{X}\}$ and/or $\{\omega_h(x) : h = 1, \ldots, \infty,\ x \in \mathcal{X}\}$ that maintain G_0 as the marginal distribution for θ_h^* and marginal beta distributions for latent stick-breaking variables. Often, either the weights or atoms are modeled with a stochastic process while the other set remains common across $x \in \mathcal{X}$.

The stick-breaking representation has led to other extensions, such as the probit stick-breaking model of Rodríguez and Dunson (2011). Here, the latent beta variables are replaced with probit link functions, waiving the marginal DP in favor of a convenient and familiar model for covariate dependence.

5.1.3 Bayesian nonparametric regression

Dirichlet process mixture and dependent Dirichlet process models comprise much of the BNP regression methods proposed in the literature. Müller et al. (2015, Ch. 4) provide an overview. We briefly review BNP regression here to place our proposed model in context, and because our chosen dependence structure yields operational equivalence with regression models on general covariates.

An early approach to fully nonparametric regression arises from conditional inferences available through joint density estimation. Müller et al. (1996) propose an approach termed curve fitting using mixtures. Assuming random covariates x, they estimate the joint density of (y, x) with a DP mixture of multivariate Gaussian kernels. Conditional densities and mean functionals (regression), which through this construction are very flexible, are then easily obtained in posterior analysis. This approach has motivated several extensions as well as simplifications (see Wade et al., 2014b for a review). Wade et al. (2014a) note that covariates in this model disproportionately drive clustering behavior, and they propose a hierarchical extension to the DP to address this issue. Other approaches explicitly characterize clustering dependence on covariates (Park and Dunson, 2010).

Regression methods based on the DDP typically begin with common atoms as in De Iorio et al. (2004), which have an equivalent representation as Gaussian DPM models with linear-regression-type kernel means. Extensions targeting the weights include Chung and Dunson (2009), who employ the probit stick-breaking formulation and incorporate component-specific stochastic-search variable selection. Dunson and Park (2008) retain beta stick-breaking latent variables, but weight them directly with covariate-dependent density kernels. Reich et al. (2012) utilize a prior reminiscent of automatic relevance detection on the kernels to select regressors. Fuentes-García et al. (2009) propose a geometric weights formulation that depends on covariates through a link-transformed Gaussian process. Recently, Barrientos et al. (2017) proposed a DDP-based framework for nonparametric regression for bounded response, considering both covariate-dependent atoms and weights, and general link functions for stick breaking.

5.1.4 Bayesian nonparametric methods for time series

Bayesian nonparametric modeling applications to problems in time series have seen rapid growth over the past two decades. We restrict attention to a few methods involving DPM or closely related models.

Although the model we propose does not incorporate Markov dependence among the latent variables defining the mixture, BNP applications to state-space and hidden-Markov models (HMM) bear mentioning here. Taddy and Kottas (2009) propose a HMM with a fixed number of hidden states and independent DP priors mixing on each emission distribution. In a slightly different formulation, Yau et al. (2011) mix emission distributions with *both* a latent hidden process and a DP prior. Rodríguez and Ter Horst (2008) work in a state-space context, mixing the distribution of observables with a common-weights DDP in which the

atoms evolve according to a linear, Gaussian random walk, resulting in a countable mixture of dynamic linear models (DLM, West and Harrison, 1997). Caron et al. (2007) use DPM priors for the observation and state noise distributions in DLMs. Alternately, Fox et al. (2011) propose Markov-switching DLMs in which the space of hidden states is countable, building on the infinite HMM of Beal et al. (2002).

DP mixtures of linear autoregressive models, which are closer to our formulation, can be viewed as a nonparametric extension of the mixture autoregressive model of Wong and Li (2000). Lau and So (2008) proposed a model of this type, specifying a DP prior for the mixing distribution of the autoregressive coefficients, variance, and autoregressive order in each kernel. Di Lucca et al. (2013) frame a similar model as a DDP, ultimately demonstrating first-order mixtures of autoregressive models with common weights for continuous and binary time series. Tang and Ghosal (2007a) propose and establish posterior consistency for a class of nonlinear autoregressive models, intended for ergodic time series, which mix over the parameters of a specific link function in the means of Gaussian kernels. Tang and Ghosal (2007b) explores posterior consistency for BNP estimation of transition densities more generally.

DP mixtures of linear autoregressive models with lag-dependent weights can likewise be viewed as a nonparametric extension of mixture-of-experts models. Müller et al. (1997) proposed a model of this type using a finite mixture construction, but placing a DP prior on the mixing distribution for the coefficients and variance of the autoregressive kernel, as well as for location parameters of normalized Gaussian weight kernels on lags.

Stationarity, or time invariance of the marginal probability distribution $p(y_t)$, is not a common feature of the flexible models described thus far. Mena and Walker (2005) and Martınez-Ovando and Walker (2011) do 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 flexible 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 weights. Kalli and Griffin (2018) extend this framework to a stationary multivariate autoregressive model of multiple lags, although it is demonstrated with a single lag.

DeYoreo and Kottas (2017) use construction similar to Antoniano-Villalobos and Walker (2016), but do not assume stationarity. They empirically demonstrate superior flexibility over the stationary model. Our proposed model is a multiple-lag analogue, and the construction procedure is outlined in Section 5.2.1.

5.1.5 Order and lag selection

We have included references to work in the BNP literature that address the problem of lag selection (and variable selection, in the case of regression). 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 autoregressive characteristic polynomial together with a reversible-jump algorithm to infer order. Wood et al. (2011) likewise use reversible jump as part of a two-stage MCMC sampler to infer component-specific order and perform Bayesian model averaging

in their time-weighted mixture of autoregressive models. Techniques for Bayesian variables selection, including BNP, are further reviewed in Section 5.4.

This chapter proceeds as follows. In Section 5.2, we propose a Bayesian nonparametric time-series model for density autoregression and present details for implementation and inference. In Section 5.3, we illustrate the model fit to synthetic and real data. In Section 5.4, we extend the model to incorporate inferences about relevant lags and demonstrate its use on data. Section 5.5 compares density estimation performance for models proposed in this chapter and Chapter 4 with simulated nonlinear time series featuring skewness, heteroscedasticity, and different lag structures. We conclude with discussion in Section 5.6.

5.2 Model

Our modeling objective is to develop a general-purpose and fully nonparametric time-homogeneous Markovian model for continuous-state time series that is sufficiently flexible to: 1) estimate possibly 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 the values of up to L lags. If y_t and $\boldsymbol{y}_{t-1} \equiv (y_{t-1}, \ldots, y_{t-L})$ denote the observation at time t and first L lags, respectively, the general model formulation for the transition density can be written as

$$f(y_t \mid \boldsymbol{y}_{t-1}) = \sum_{h=1}^{\infty} \underbrace{q_h(\boldsymbol{y}_{t-1})}_{\text{local weights}} \underbrace{\mathrm{N}(y_t \mid \mu_h(\boldsymbol{y}_{t-1}), \sigma_h^2)}_{\text{mixture kernels}}, \tag{5.2}$$

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 $\boldsymbol{y}_{t-1} \in \mathbb{R}^L$. We further use kernel mean functions $\mu_h(\boldsymbol{y}_{t-1})$ that are linear in the lags, resulting in a local mixture of linear transition densities. The third objective of order and lag selection is accomplished through a stochastic-search prior structure.

The model is fully nonparametric in the sense that its form derives from a prior for joint density estimation that enjoys full support in the Kullback Leibler sense (Wu et al., 2008). Time homogeneity in the model 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, and dependence on, lagged observations. Apparently time-dependent structural changes can sometimes be attributed to differences in dynamics among disjoint regions of the phase space, which can be captured by the proposed model. 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 can and does occur. We therefore encourage responsible investigation before drawing conclusions from this or any statistical model for which observations are time-indexed.

We proceed with a complete model derivation and specification in Section 5.2.1, which also describes a parameterization that is useful for interpretation and implementation, and address truncation of the countable mixture. Section 5.2.2 discusses the roles of model parameters and gives recommended prior settings. Section 5.2.3 briefly outlines the Markov chain Monte Carlo algorithm used for posterior inferences and addresses implementation. Finally, Section 5.2.4 discusses

model inferences, including transition density estimation.

5.2.1 Model specification

One avenue to arrive at the conditional density form in (5.2) begins with a prior for joint density estimation. To highlight operational equivalence with nonparametric regression, we use $y \in \mathbb{R}$ to represent a generic continuous response and $\boldsymbol{x} \in \mathbb{R}^L$ to denote a vector of continuous covariates. In the context of the final model, however, we have $\boldsymbol{x}_t \equiv \boldsymbol{y}_{t-1}$.

We begin as in Müller et al. (1996) by considering y and $\boldsymbol{x}$ to arise jointly from a Gaussian DPM,

$$\begin{aligned} f_{YX}(y, \boldsymbol{x} \mid G) &= \int \mathrm{N}\left((y, \boldsymbol{x}) \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \, \mathrm{d}G(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \,, \\ G \mid \alpha, G_0 &\sim \mathrm{DP}(\alpha, G_0) \,, \end{aligned}$$

or equivalently, under the stick-breaking representation (Sethuraman, 1994),

$$f_{YX}(y, \boldsymbol{x} \mid G) = \sum_{h=1}^{\infty} \omega_h \, \mathrm{N}\left((y, \boldsymbol{x}) \mid \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h\right) \,, \tag{5.3}$$

where $\{\omega_h\}$ are constructed as

$$\omega_1 = v_1, \;\; \omega_h = v_h \prod_{j=1}^{h-1} (1 - v_j), \;\; \text{for } h > 2, \;\; \text{and } v_h \overset{\text{iid}}{\sim} \mathrm{Beta}(1, \alpha) \,, \tag{5.4}$$

and $(\boldsymbol{\mu}, \boldsymbol{\Sigma})_h \overset{\text{iid}}{\sim} G_0$. Conditioning on $\boldsymbol{x}$, we obtain

$$\begin{aligned} f_{Y|X}(y \mid \boldsymbol{x}, G) = \frac{f_{YX}(y, \boldsymbol{x} \mid G)}{\int f_{YX}(y, \boldsymbol{x} \mid G) \, \mathrm{d}y} &= \frac{\sum_{h=1}^{\infty} \omega_h \, \mathrm{N}_{(h)}(\boldsymbol{x}) \, \mathrm{N}_{(h)}(y \mid \boldsymbol{x})}{\sum_{j=1}^{\infty} \omega_j \, \mathrm{N}_{(j)}(\boldsymbol{x})} \\ &= \sum_{h=1}^{\infty} \underbrace{q_h(\boldsymbol{x})}_{\text{local weights}} \underbrace{\mathrm{N}(y_t \mid \mu_h(\boldsymbol{x}), \sigma_h^2)}_{\text{mixture kernels}} , \end{aligned} \tag{5.5}$$

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 (5.5) 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 covariance $\sigma^2 \equiv (\sigma^y)^2 - \boldsymbol{\Sigma}^{yx}(\boldsymbol{\Sigma}^x)^{-1}\boldsymbol{\Sigma}^{xy}$). The second line of (5.5) reveals the local linear structure of the model with lag-dependent weights and mixture kernels with means depending linearly on $\boldsymbol{x}$. Local weighting allows the model to capture nonlinearity while the mixture structure accommodates non-Gaussianity.

This procedure yields a model structure for a conditional density satisfying the requirements of the proposed model (5.2). Specifically, we have $\sum_{h=1}^{\infty} \omega_h = 1$ almost surely (Ishwaran and James, 2001). Then, 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 $\boldsymbol{x} \in \mathbb{R}^L$, producing a valid weight function.

Although $\boldsymbol{x}$ (representing $\boldsymbol{y}_{t-1}$) can legitimately be considered random in the time-series context, the Markovian likelihood $p(\{y_t\}) = \prod_t p(y_t \mid \boldsymbol{y}_{t-1})$ requires that the conditional transition density (5.5) form the basis of the model. Furthermore, unless we assume stationarity of the process $\{y_t\}$ and impose corresponding restrictions on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, the joint mixture density in (5.3) 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 *not* assume stationarity of the time series (DeYoreo and Kottas, 2017). This decision carries

two important consequences. First, the proposed model does not estimate a joint density for consecutive observations, so that the density implied by reversing (5.5) is not interpretable. Second, the densities $\{\mathrm{N}_{(h)}(\boldsymbol{x})\}$ serve only to support the function of the mixture weights $\{q_h(\cdot)\}$ and are not related to a joint probability distribution for lags.

The model likelihood for time series data, conditional on the first L observations, is $\prod_{t=L+1}^{T} f_{Y|X}(y_t \mid \boldsymbol{y}_{t-1}, G)$. This likelihood, based on (5.5), 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. All three implement their respective models only for a single lag. This weight structure has also been explored for finite mixtures with the stationarity assumption (Glasbey, 2001; Kalliovirta et al., 2015). The general regression formulation in (5.2) with local weights $q_h(\boldsymbol{x})$ resembles mixture-of-experts constructions. The re-weighting of $\{\omega_h\}$ with probability density kernels on $\boldsymbol{x}$ distinguishes our model from nonparametric extensions of MoE, such as dependent Dirichlet process models and the kernel stick-breaking model class introduced by Dunson and Park (2008), both of which introduce covariate dependence through the $\{v_h\}$ variables in (5.4).

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 kernel covariance matrix according to the factorization $\boldsymbol{\Sigma} = \boldsymbol{B}^{-1}\boldsymbol{\Delta}(\boldsymbol{B}^{-1})'$ where

$\boldsymbol{\Delta} = \text{diag}(\sigma^2, \delta_1^x, \ldots, \delta_{L,}^x)$ and $\boldsymbol{B}$ is the upper unit-triangular matrix

$$\boldsymbol{B} = \begin{pmatrix} 1 & \beta_1^y & \beta_2^y & \cdots & \beta_{L-1}^y & \beta_L^y \\ 0 & 1 & \beta_{1,2}^x & \cdots & \beta_{1,L-1}^x & \beta_{1,L}^x \\ 0 & 0 & 1 & \cdots & \beta_{2,L-1}^x & \beta_{2,L}^x \\ \vdots & \vdots & & \ddots & & \vdots \\ & & & & 1 & \beta_{L-1,L}^x \\ 0 & & \ldots & & 0 & 1 \end{pmatrix}. \tag{5.6}$$

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 5.4.

The primary advantage of this parameterization stems from the sequential decomposition of the joint Gaussian density for y and $\boldsymbol{x}$ into $L+1$ univariate Gaussian densities. Specifically,

$$\begin{aligned} \text{N}\left((y, \boldsymbol{x}) \mid \boldsymbol{\mu}, \boldsymbol{B}^{-1}\boldsymbol{\Delta}(\boldsymbol{B}^{-1})'\right) = &\,\text{N}(x_L \mid \mu_L^x,\, \delta_L^x) \times \\ &\text{N}(x_{L-1} \mid \mu_{L-1}^x - \beta_{L-1,L}^x(x_{L-1} - \mu_{L-1}^x),\, \delta_{L-1}^x) \times \\ \times \cdots \times &\,\text{N}\left(x_1 \mid \mu_1^x - \sum_{\ell=2}^{L} \beta_{1,\ell}^x(x_\ell - \mu_\ell^x),\, \delta_1^x\right) \times \\ &\,\text{N}\left(y \mid \mu^y - \sum_{\ell=1}^{L} \beta_\ell^y(x_\ell - \mu_\ell^x),\, \sigma^2\right). \end{aligned} \tag{5.7}$$

The typical application of this parameterization constructs the vector sequentially from front to back, resulting in a lower unit-triangular $\boldsymbol{B}$ that conforms to the standard definition of Cholesky factorization. Instead, we construct from back

(most distant lag) to front (y) so that the response density depends on the entire $\boldsymbol{x}$ vector while maintaining order convention for time-delay embedding vectors. This fully parameterized representation of the covariance matrix is flexible, as each β parameter is unrestricted and δ parameters need only be positive. Furthermore the popular inverse-Wishart prior can be constructed as a special case (Daniels and Pourahmadi, 2002). This representation also allows substantial control over the marginal weight kernel 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 of $\boldsymbol{B}$, and $\boldsymbol{\Delta}^x = \text{diag}(\delta_1^x, \ldots, \delta_L^x)$.

The final term in (5.7) 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 (5.5), 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.

DP truncation

A primary challenge in implementing the model in (5.5) is the infinite summation in the denominator of the weights. While Antoniano-Villalobos and Walker (2016) address this problem by introducing multiple sets of auxiliary variables and consider slice sampling in the style of Kalli et al. (2011), all previous implementations of this model class ultimately rely on truncation of the infinite mixture. We also truncate, following the blocked Gibbs strategy of Ishwaran and James (2001).

There are both theoretical and practical considerations when selecting the truncation level, H. Given a value of the DP concentration parameter α, we can calculate the prior expected truncation error in the weights, $\text{E}(\omega_H) = \text{E}(\prod_{h=1}^{H-1}(1 -$

$v_h)) = [\alpha/(1+\alpha)]^{H-1}$. We can also monitor this final weight ω_H throughout MCMC sampling to ensure it remains small. Because the mixture provides flexibility in the model, it may be necessary to increase the truncation level to estimate transition densities/functions that exhibit complex local behavior. We therefore also advocate monitoring the number of occupied clusters throughout MCMC to ensure that it does not approach H.

Hierarchical formulation

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 cluster h. We denote all cluster-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 $\boldsymbol{\beta}_r^x$ (for $r = 1, \ldots, L-2$), and $\beta_{L-1}^x \equiv \beta_{L-1,L}^x$ taken from the corresponding rows of $\boldsymbol{B}$, and $\boldsymbol{\delta}^x = (\delta_1^x, \ldots, \delta_L^x)$. We again simplify notation by using $\mathrm{N}_{(h)}(\cdot)$ to indicate that all parameters used to specify that the mean and covariance are indexed by h. The hierarchical formulation of our model is given by

$$
\begin{aligned}
y_t \mid \boldsymbol{x}_t, s_t = h, \{\boldsymbol{\eta}\} &\overset{\text{ind.}}{\sim} \mathrm{N}_{(h)}\left(y_t \mid \mu^y - \sum_{\ell=1}^{L} \beta_\ell^y (x_{t,\ell} - \mu_\ell^x),\, \sigma^2\right), \\
&\qquad \text{for } t = L+1, \ldots, T, \text{ and } h = 1, \ldots, H, \\
\Pr(s_t = h \mid \boldsymbol{x}_t, \{\boldsymbol{\eta}\}, \boldsymbol{\omega}) &= \frac{\omega_h \, \mathrm{N}_{(h)}(\boldsymbol{x}_t \mid \boldsymbol{\mu}^x, \boldsymbol{\Sigma}^x)}{\sum_{j=1}^H \omega_j \, \mathrm{N}_{(j)}(\boldsymbol{x}_t \mid \boldsymbol{\mu}^x, \boldsymbol{\Sigma}^x)}, \qquad (5.8) \\
\omega_1 = v_1,\ \omega_h = v_h \prod_{j=1}^{h-1} (1 - v_j),\ &\text{for } j = 2, \ldots, H-1, \text{ and } \omega_H = \prod_{j=1}^{H-1} (1 - v_j), \\
v_j \mid \alpha &\overset{\text{iid}}{\sim} \mathrm{Beta}(1, \alpha), \text{ for } j = 1, \ldots, H-1, \\
\boldsymbol{\eta}_h \mid G_0 &\overset{\text{iid}}{\sim} G_0(\boldsymbol{\eta}_h), \text{ for } h = 1, \ldots, H, \\
\alpha &\sim \mathrm{Ga}(a_\alpha, b_\alpha),
\end{aligned}
$$

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 5.2.3).

We complete the model 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 $\ell = 1, \ldots, L$.

5.2.2 Prior settings

The priors for the hierarchical model in Section 5.2.1 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, if capturing nonlinear dynamics and transition density estimation are of primary interest, one may consider centering and scaling the (possibly de-trended) time series to unit marginal variance, and basing hyperparameter settings on default values. In this section, we recommend default values derived from the marginal center and range of the time-series data.

Before we recommend default hyperparameter values, it is useful to discuss the function and interpretation of model parameters. The first consideration is that the model (5.5) 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 function or surface spanning $\mathbb{R}^L$ for each mixture component h that follows the contours of a L-variate Gaussian density down-weighted by ω_h. The cluster-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, $\boldsymbol{\beta}^y$ provide the cluster-conditional mean as a first-order linear combination of $\boldsymbol{x}$, and σ^2 provides observation error variance around the cluster's mean.

One primary functional of interest derived from the transition density in (5.5) is the conditional expectation $\mathrm{E}(y \mid \boldsymbol{x}) = \int y\, f(y \mid \boldsymbol{x}, G)\ \mathrm{d}y = \sum_h q_h(\boldsymbol{x})\, \mu_h(\boldsymbol{x})$, to which we refer as the transition mean functional. 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 clusters, 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 clusters to concentrate on small regions, analogous to using many knots in spline models.

We can visualize the effects of prior settings through prior simulation in low-dimensional models. As an example, Figure 5.1 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 clusters with low values of α results in transition mean functions with few changepoints and long stretches of near linearity, whereas allowing more clusters increases variability in the curve. Low values for δ^x likewise encourage rigid transition mean curves with abrupt changepoints. Increasing the variance in the weight kernels has a smoothing effect, as expected. Note that for some regions of the lag space, the transition density is multimodal, and so the transition mean does not follow any of the lines corresponding to mixture components in that

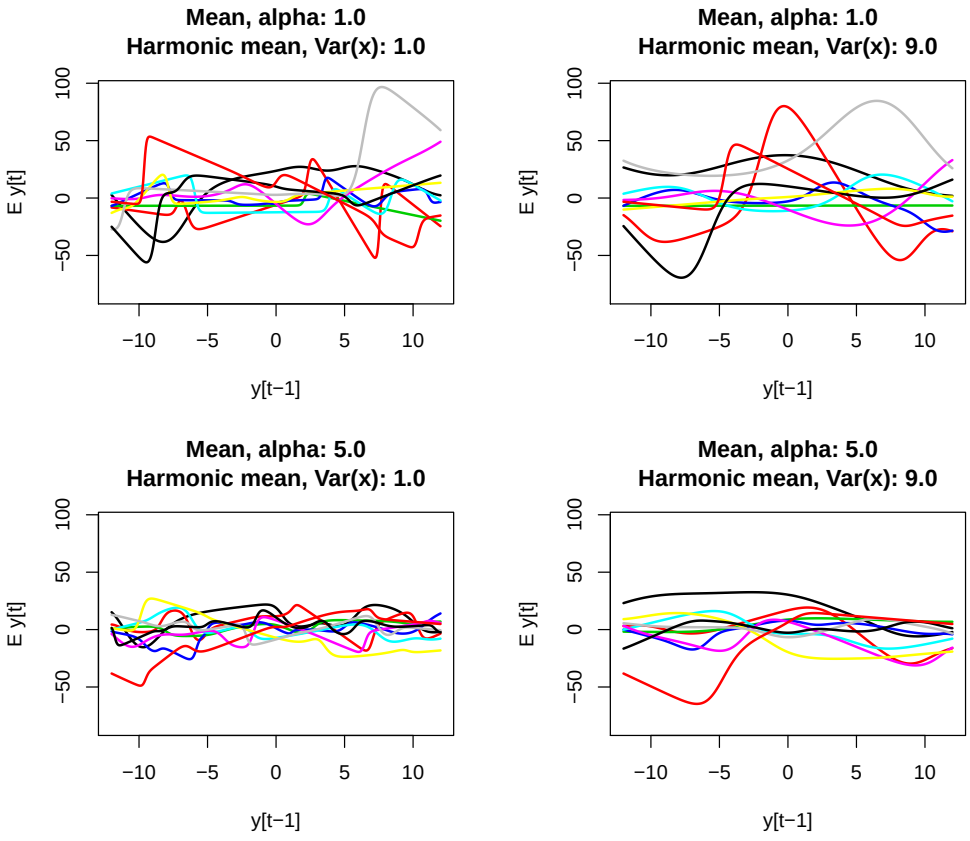


Figure 5.1: Ten prior realizations of the transition mean for the proposed nonparametric model with a single lag, under combinations of prior settings for α and δ^x.

region. As with the transition mean, we can use prior simulation to elucidate the effects of prior settings for transition densities to aid practitioners in specifying desired characteristics and performing sensitivity analysis.

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 $\alpha \log\left((\alpha + T - L)/\alpha\right)$ as a rough prior estimate for the number of clusters. While this applies in the prior joint model, the number of clusters in our conditional model (5.5) 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 $\boldsymbol{S}_0^* = \text{diag}([\text{range}(y)/6.0]^2, 1.0, \ldots, 1.0)$, with range$(y)$ representing the range of the time series, empirical or user-defined. This specification 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. We set $\nu^* = 50\,(L + 1 + 2)$ and $\boldsymbol{\Psi}_0^* = s_{00}^{-1}\,\text{diag}([\text{range}(y)/2.0]^2, 16.0, \ldots, 16.0)$ to avoid extreme values in $(\boldsymbol{\Lambda}_0^*)^{-1}$ occasionally encountered during MCMC, thus promoting stability and identifiability. We scale (divide) $\boldsymbol{\Psi}_0^*$ by s_{00}, the prior estimate of σ^2, to partially compensate and control 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 $\mathcal{R} \in [5.0, 25.0]$). We use $\boldsymbol{m}_0^x = \bar{y}\,\mathbf{1}$ 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 $\boldsymbol{\Psi}_0^{\mu_x} = [\text{range}(y)/1.0]^2\, \boldsymbol{I}_L$. 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.

5.2.3 Computation

We outline the Markov chain Monte Carlo algorithm used to obtain posterior samples from the proposed model and discuss details for our approach to implementation challenges. The algorithm consists of a Gibbs sampler containing a variety of update methods for parameter blocks. If we condition on the first L observations, the hierarchical model (5.8) yields the full joint posterior distribution over all model parameters up to proportionality,

$$
\begin{aligned}
p(\cdots \mid \{y_t\}_{t=1}^T) \propto \\
&\prod_{t=L+1}^{T} \left[\frac{\omega_{s_t}\, \mathrm{N}_{(s_t)}(\boldsymbol{x}_t \mid \boldsymbol{\mu}^x, \boldsymbol{\Sigma}^x)}{\sum_{j=1}^{H} \omega_j\, \mathrm{N}_{(j)}(\boldsymbol{x}_t \mid \boldsymbol{\mu}^x, \boldsymbol{\Sigma}^x)}\, \mathrm{N}_{(s_t)}\left(y_t \mid \mu^y - \sum_{\ell=1}^{L} \beta_\ell^y (x_{t,\ell} - \mu_\ell^x),\, \sigma^2 \right) \right] \times \\
&\prod_{h=1}^{H} \left[\mathrm{Beta}(v_h \mid 1, \alpha)^{1_{(h<H)}}\; p(\boldsymbol{\eta}_h \mid G_0) \right]\; \mathrm{Ga}(\alpha \mid a_\alpha, b_\alpha) \times \\
&\mathrm{N}(\boldsymbol{\beta}_0^*)\; \mathrm{IWish}\left((\boldsymbol{\Lambda}_0^*)^{-1} \right)\; \mathrm{Ga}(s_0)\; \mathrm{N}(\boldsymbol{\mu}_0^x)\; \mathrm{IWish}((\boldsymbol{\Lambda}^{\mu_x})^{-1}) \times \\
&\prod_{r=1}^{L-1} \left[\mathrm{N}(\boldsymbol{\beta}_{0,r}^{\beta_x})\; \mathrm{IWish}((\boldsymbol{\Lambda}_{0,r}^{\beta_x})^{-1}) \right]\; \prod_{\ell=1}^{L} \mathrm{Ga}(s_{0,\ell}^x)\,, \qquad (5.9)
\end{aligned}
$$

where

$$
\begin{aligned}
p(\boldsymbol{\eta}_h \mid G_0) = & \, \mathrm{N}\left((\mu^y_{(h)}, \boldsymbol{\beta}^y_{(h)}) \mid \boldsymbol{\beta}^*_0, \sigma^2_{(h)}(\boldsymbol{\Lambda}^*_0)^{-1}\right) \, \mathrm{IG}\left(\sigma^2_{(h)} \mid \frac{\nu_{\sigma^2}}{2}, \frac{\nu_{\sigma^2} s_0}{2}\right) \times \\
& \mathrm{N}\left(\boldsymbol{\mu}^x_{(h)} \mid \boldsymbol{\mu}^x_0, (\boldsymbol{\Lambda}^{\mu_x})^{-1}\right) \prod_{r=1}^{L-1} \mathrm{N}\left(\boldsymbol{\beta}^x_{r,(h)} \mid \boldsymbol{\beta}^{\beta_x}_{0,r}, (\boldsymbol{\Lambda}^{\beta_x}_{0,r})^{-1}\right) \times \\
& \prod_{\ell=1}^{L} \mathrm{IG}\left(\delta^x_{\ell,(h)} \mid \frac{\nu^{\delta^x}_\ell}{2}, \frac{\nu^{\delta^x}_\ell s^x_{0,\ell}}{2}\right)
\end{aligned} \tag{5.10}
$$

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

Latent states

The latent states identifying cluster membership for each observation y_t are updated individually, for $t = L+1, \ldots, T$, with their discrete full conditional distributions $\Pr(s_t = h \mid \cdots) \propto \omega_h \, \mathrm{N}_{(h)}(\boldsymbol{x}_t \mid \boldsymbol{\mu}^x, \boldsymbol{\Sigma}^x) \, \mathrm{N}_{(h)}\left(y_t \mid \mu^y - \sum_{\ell=1}^{L} \beta^y_\ell (x_{t,\ell} - \mu^x_\ell), \sigma^2\right)$, for $h = 1, \ldots, H$.

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 (5.9), admit $H-1$ independent beta full conditional distributions (Ishwaran and James, 2001). In our model, the full conditional distributions are given as

$$
\begin{aligned}
p(\{v_h\} \mid \cdots) & \propto \prod_{t=L+1}^{T} \left[\frac{\omega_{s_t}}{\sum_{j=1}^{H} \omega_j \, \mathrm{N}_{(j)}(\boldsymbol{x}_t \mid \boldsymbol{\mu}^x, \boldsymbol{\Sigma}^x)}\right] \prod_{h=1}^{H-1} \mathrm{Beta}(v_h \mid 1, \alpha) \\
& \propto \frac{\prod_{h=1}^{H-1} \mathrm{Beta}(v_h \mid 1 + n^*_h, \alpha + \sum_{k=h+1}^{H} n^*_k)}{\prod_{t=L+1}^{T} \sum_{j=1}^{H} v_j \prod_{i=1}^{j-1}(1 - v_i) \;\; \mathrm{N}_{(j)}(\boldsymbol{x}_t \mid \boldsymbol{\mu}^x, \boldsymbol{\Sigma}^x)} \,,
\end{aligned} \tag{5.11}
$$

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 clusters. We define $v_H = 1$ for convenience in notation. This full conditional distribution is unchanged from the distribution reported in DeYoreo and Kottas (2017), with exception that the Gaussian kernels appearing in the denominator are now multivariate Gaussian for the vector $\boldsymbol{x}_t$. This small 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 hyperrectangle slice sampler of Neal (2003) (summarized in Figure 8 of that article) to update all v_h, $h = 1, \ldots, H-1$, simultaneously. Details are given in Appendix D.

Cluster-specific parameters

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

$$
\begin{aligned}
p(\boldsymbol{\eta}_h \mid \cdots) \propto \\
&\prod_{t:s_t=h} \left[\mathrm{N}_{(h)}(\boldsymbol{x}_t \mid \boldsymbol{\mu}^x, \boldsymbol{\Sigma}^x)\, \mathrm{N}_{(h)}\left(y_t \;\middle|\; \mu^y - \sum_{\ell=1}^{L} \beta_\ell^y (x_{t,\ell} - \mu_\ell^x),\, \sigma^2 \right) \right] \times \\
&\prod_{t=L+1}^{T} \left[\sum_{j=1}^{H} \omega_j\, \mathrm{N}_{(j)}(\boldsymbol{x}_t \mid \boldsymbol{\mu}^x, \boldsymbol{\Sigma}^x) \right]^{-1} \mathrm{N}\left((\mu_{(h)}^y, \boldsymbol{\beta}_{(h)}^y) \mid \boldsymbol{\beta}_0^*, \sigma_{(h)}^2 (\boldsymbol{\Lambda}_0^*)^{-1} \right) \times \\
&\mathrm{IG}\left(\sigma_{(h)}^2 \mid \frac{\nu_{\sigma^2}}{2}, \frac{\nu_{\sigma^2} s_0}{2} \right) \mathrm{N}\left(\boldsymbol{\mu}_{(h)}^x \mid \boldsymbol{\mu}_0^x, (\boldsymbol{\Lambda}^{\mu_x})^{-1} \right) \times \\
&\prod_{r=1}^{L-1} \mathrm{N}\left(\boldsymbol{\beta}_{r,(h)}^x \mid \boldsymbol{\beta}_{0,r}^{\beta_x}, (\boldsymbol{\Lambda}_{0,r}^{\beta_x})^{-1} \right) \prod_{\ell=1}^{L} \mathrm{IG}\left(\delta_{\ell,(h)}^x \mid \frac{\nu_\ell^{\delta^x}}{2}, \frac{\nu_\ell^{\delta^x} s_{0,\ell}^x}{2} \right),
\end{aligned}
\tag{5.12}
$$

for $h = 1, \ldots, H$. To improve mixing of the y-indexed, cluster-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 $\prod_t \sum_j \omega_j \, \mathrm{N}_{(j)}(\boldsymbol{x_t})$, precluding simple conjugate updates.

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

$$
\begin{aligned}
p(\boldsymbol{\eta}_h^x \mid \cdots, -\boldsymbol{\eta}_h^y) \propto & \prod_{t:s_t=h} \mathrm{N}_{(h)}(\boldsymbol{x}_t \mid \boldsymbol{\mu}^x, \boldsymbol{\Sigma}^x) \prod_{t=L+1}^{T} \left[\sum_{j=1}^{H} \omega_j \, \mathrm{N}_{(j)}(\boldsymbol{x}_t \mid \boldsymbol{\mu}^x, \boldsymbol{\Sigma}^x) \right]^{-1} \times \\
& \mathrm{N}\left(\boldsymbol{\mu}_{(h)}^x \mid \boldsymbol{\mu}_0^x, (\boldsymbol{\Lambda}^{\mu_x})^{-1}\right) \prod_{r=1}^{L-1} \mathrm{N}\left(\boldsymbol{\beta}_{r,(h)}^x \mid \boldsymbol{\beta}_{0,r}^{\beta_x}, (\boldsymbol{\Lambda}_{0,r}^{\beta_x})^{-1}\right) \times \\
& \prod_{\ell=1}^{L} \mathrm{IG}\left(\delta_{\ell,(h)}^x \mid \frac{\nu_\ell^{\delta x}}{2}, \frac{\nu_\ell^{\delta x} s_{0,\ell}^x}{2}\right) \det(\boldsymbol{\Lambda}_{1,h}^*)^{-1/2} \times \\
& \left[\nu_{\sigma^2} s_0 + \boldsymbol{y}_{(h)}' \, \boldsymbol{y}_{(h)} + (\boldsymbol{\beta}_0^*)' \, \boldsymbol{\Lambda}_0^* \, \boldsymbol{\beta}_0^* - (\boldsymbol{\beta}_{1,h}^*)' \, \boldsymbol{\Lambda}_{1,h}^* \, \boldsymbol{\beta}_{1,h}^*\right]^{-(\nu_{\sigma^2}+n_h^*)/2},
\end{aligned}
\tag{5.13}
$$

where $\boldsymbol{\Lambda}_{1,h}^* = \boldsymbol{D}_h' \, \boldsymbol{D}_h + \boldsymbol{\Lambda}_0^*$; $\boldsymbol{\beta}_{1,h}^* = (\boldsymbol{\Lambda}_{1,h}^*)^{-1}(\boldsymbol{\Lambda}_0^* \, \boldsymbol{\beta}_0^* + \boldsymbol{D}_h' \, \boldsymbol{y}_{(h)})$; $\boldsymbol{y}_{(h)}$ 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 $(1, \mu_{1,(h)}^x - x_{t,1}, \ldots, \mu_{L,(h)}^x - x_{t,L})$ for each t such that $s_t = h$. Note that proportionality in (5.13) is preserved with respect to the $\{\mu_{\ell,(h)}^x\}$, which appear in the regression means for y_t. Aside from the mixture denominator factor, 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 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 (5.13). 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

$$\sigma_h^2 \mid \cdots, -\boldsymbol{\beta}_h^* \sim \mathrm{IG}\left(\frac{\nu_{\sigma^2} + n_h^*}{2}, \frac{\nu_{\sigma^2}\, s_0 + \boldsymbol{y}'_{(h)}\, \boldsymbol{y}_{(h)} + (\boldsymbol{\beta}_0^*)'\, \boldsymbol{\Lambda}_0^*\, \boldsymbol{\beta}_0^* - (\boldsymbol{\beta}_{1,h}^*)'\, \boldsymbol{\Lambda}_{1,h}^*\, \boldsymbol{\beta}_{1,h}^*}{2}\right),$$

$$\boldsymbol{\beta}_h^* \mid \sigma_h^2, \cdots \sim \mathrm{N}\left(\boldsymbol{\beta}_{1,h}^*, \sigma_h^2\, (\boldsymbol{\Lambda}_{1,h}^*)^{-1}\right).$$

Parameters in the base measure

Let $n^* = \sum_{h=1}^{H} 1_{(n_h^*>0)}$ count the total number of occupied clusters. The posterior conditional density for $\boldsymbol{\beta}_0^*$ is proportional to $\prod_{\{h:n_h^*>0\}} [\mathrm{N}(\boldsymbol{\beta}_h^* \mid \boldsymbol{\beta}_0^*, \sigma_h^2\, (\boldsymbol{\Lambda}_0^*)^{-1})]\, \mathrm{N}(\boldsymbol{\beta}_0^* \mid \boldsymbol{b}_0^*, \boldsymbol{S}_0^*)$, 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 $\prod_{\{h:n_h^*>0\}} [\mathrm{N}(\boldsymbol{\beta}_h^* \mid \boldsymbol{\beta}_0^*, \sigma_h^2\, (\boldsymbol{\Lambda}_0^*)^{-1})]\, \mathrm{IWish}((\boldsymbol{\Lambda}_0^*)^{-1} \mid \nu^*, \nu^*\, \boldsymbol{\Psi}_0^*)$, 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 $\prod_{\{h:n_h^*>0\}} [\mathrm{IG}(\nu_{\sigma^2}/2,\, \nu_{\sigma^2}\, s_0/2)]\, \mathrm{Ga}(s_0 \mid a_{s_0}, b_{s_0})$, 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 clusters.

DP concentration parameter

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

Implementation

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 clusters $\{s_t\}$, for which we use output from a clustering algorithm applied to $(y_t, x_{1,t}, \ldots, x_{L,t})$, for all $t = L+1, \ldots, T$. For example, we use hierarchical clustering with complete linkage and assign the observations into H clusters.

MCMC begins with an adaptation phase used to tune the covariance of the random-walk proposal distributions for $\{\boldsymbol{\eta}_h^x\}_{h=1}^H$. 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., $[0.02, 0.25]$, to promote exploration across the multimodal posterior) 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.

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. This further indicates a potential area to improve parameter economy in the weight functions. Indeed, weight kernels with local independence between elements of $\boldsymbol{x}$, as proposed by Shahbaba and Neal (2009), can approximate any shape if we allow for additional mixture components. Finally, such a parsimonious replacement becomes necessary if we include many lags, as the number of covariance parameters for *each* cluster grows quadratically with L. We focus here on low-order dependence $L \leq 5$, and aid mixing by iterating between adaptation and pre-burn-in runs before beginning an official burn-in run. We do note that despite the mixing challenges, MCMC chains for parameters and functionals of interest are typically stable.

5.2.4 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 5.4. The transition mean functional and estimates of the transition density function are straightforward to compute, as the stick-breaking representation and blocked Gibbs sampler yield a complete approximation of the random mixing distribution G at each iteration of MCMC. 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 (5.5), given as

$$\tilde{f}_{Y|X}(y \mid \boldsymbol{x}) = \sum_{h=1}^{H} \tilde{q}_h(\boldsymbol{x}) \; \mathrm{N}_{(h)}(y \mid \mu(\boldsymbol{x}), \sigma^2) \,, \tag{5.14}$$

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 create pointwise estimates and intervals for $\tilde{f}_{Y|X}$. 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

$$\tilde{Q}_u(y \mid \boldsymbol{x}) = u - \sum_{h=1}^{H} \tilde{q}_h(\boldsymbol{x}) \, \Phi\left([y - \mu_{(h)}(\boldsymbol{x})] / \sigma_{(h)} \right) \,, \tag{5.15}$$

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 (5.8) for each posterior sample. Such samples propagate both forecast and inferential uncertainty, and can be useful for assessing model performance with validation data.

5.3 Data illustrations

We illustrate the proposed model with three examples. The first two synthetic data examples highlight 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 $\mathcal{R} = 10.0$ (instead of the default value of 5.0), and higher degrees of freedom ν_{σ^2} and n_{s_0} (in the interval $[7.0, 10.0]$ instead of the default 5.0). For each example, multiple MCMC chains were randomly initialized using the strategy described in Section 5.2.3, with four adaptation phases followed by 400,000 burn-in samples. The next 600,000 iterations were then thinned to 1,000 for inference. Unless otherwise noted, samples were found appropriate for inferences and reported for one of the chains. Occasionally, chains abort due to numerical instabilities stemming from $\log(\omega_H)$ causing failure of the gamma update for α, and failure of Cholesky factorization.

5.3.1 Simulated data: single lag

We begin with the simulated time series introduced in Section 2.3.2, and previously analyzed in Section 4.3.1. The series was generated from

$$y_t = y_{t-2} \exp(2.6 - y_{t-2}) + \epsilon_t \,, \quad \epsilon_t \overset{\text{iid}}{\sim} \mathrm{N}(0, (0.09)^2) \,, \tag{5.16}$$

featuring first-order nonlinear dynamics as a function of the second lag only. We fit the proposed model to the original real-valued time series with $L = 2$, $T = 102$ (so that 100 observations contribute to the likelihood), and $H = 40$. The two different prior signal-to-noise ratio specifications produce similar results. Multiple chains produce similar traces of the log-likelihood and number of occupied clusters (not shown), which ranges between three and five. All traces of σ^2 for the most occupied cluster (not shown) converge to approximately 1.5 times the true value of 0.0081, due in part to the prior estimates $s_{00} \in \{0.060, 0.121\}$.

The dynamics are successfully recovered in data-rich regions of the phase space despite using an over-specified model with two lags. The upper panel of Figure 5.2 shows a posterior mean estimate of the surface, in which most variation occurs

along the second lag. We can informally assess the influence of the first lag with the second-order model by checking for sensitivity in inferences for the transition mean along values of the first lag. For example, the lower-left panel of Figure 5.2 plots the pointwise posterior mean and 95% credible intervals for the transition mean over a grid of values for the second lag, in which all values for the first lag have been fixed at their mean. The lower-right panel replicates this plot with grid values for the first lag drawn uniformly over the range of the data. This perturbation

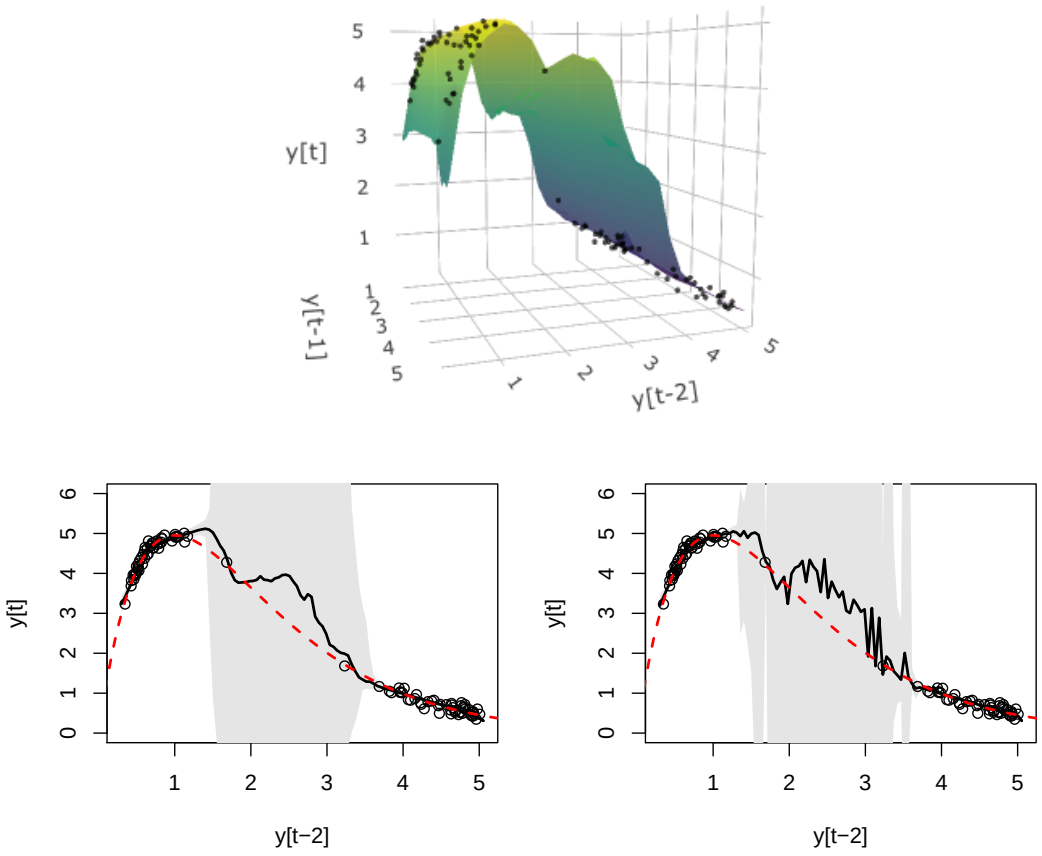


Figure 5.2: Nonparametric model fit to the single-lag dynamical simulation with noise ($T = 102$, $L = 2$). The upper panel shows the pointwise posterior mean of the transition mean surface. The lower panel shows posterior mean and 95% intervals for the transition functions over a grid of values for lag 2 with first lag values fixed at a mean value (left), and drawn uniformly (right). Data points are included, as well as the true transition map (dashed red).

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 range from approximately -10 to 17 in the fixed-grid case and -20 to 22 in the random-grid case. This appears to stem from the weight functions concentrating locally around the data, leaving other regions to revert to the prior.

5.3.2 Simulated data: time-delay embedding

Our second simulation example explores second-order dynamics with the example of statistical state-space reconstruction via time-delay embedding introduced in Section 4.3.2. We fit the model to lags of only one variable in a bivariate series generated from a two-dimensional deterministic system. Although the embedding dimension is ultimately of inferential interest, here we assume knowledge of a two-dimensional embedding and demonstrate the proposed model's ability to fit the surface.

As a mixture, the proposed model (5.5) is overtly stochastic and intended for transition density estimation. However, one can encourage more deterministic behavior through the prior, informing component-specific variances $\{\sigma_h^2\}$. Note that the hierarchical structure proposed in (5.8) uses σ^2 in the prior for $(\mu^y, \boldsymbol{\beta}^y)$. Thus, forcing σ^2 to be small requires compensating with larger values along the diagonal of the prior harmonic mean of the covariance matrix for coefficients, $\boldsymbol{\Psi}_0^*$, in order to maintain flexibility. This is default behavior recommended in Section 5.2.2, and is partially specified through the prior signal-to-noise ratio $\mathcal{R}$. We also recommend increasing the prior expectation of α to accommodate local structure in the transition map. Encouraging small weight variances δ^x may also be necessary to capture local behavior and (attempt to) avoid multimodal transition densities.

We elect to encourage rather than enforce this behavior with the mild prior settings described at the beginning of this section.

We fit the proposed model to the $\{\log(y_t)\}$ series of length $T = 102$ using $L = 2$ lags and truncation $H = 40$. Likelihood traces among six runs jump between three primary values, in close correspondence with the number of clusters, which range from five to ten. All chains settle on small σ^2 values for the most occupied cluster. Results appear insensitive to the two prior specifications. Furthermore, transition mean surface estimates for two different prior specifications are nearly visually indistinguishable. The estimated surface for the fit with higher prior signal-to-noise ratio is shown in Figure 5.3. The dynamics are captured well and the fit is superior to the first-order additive approximation of the GPMTD in Section 4.3.2.

The fit to the same simulation of length $T = 502$ (not shown) appears successful at capturing the transition surface (in Figure 4.4), including both sides of the steep and narrow "trench" near the lowest region. The low variances in the weight kernels

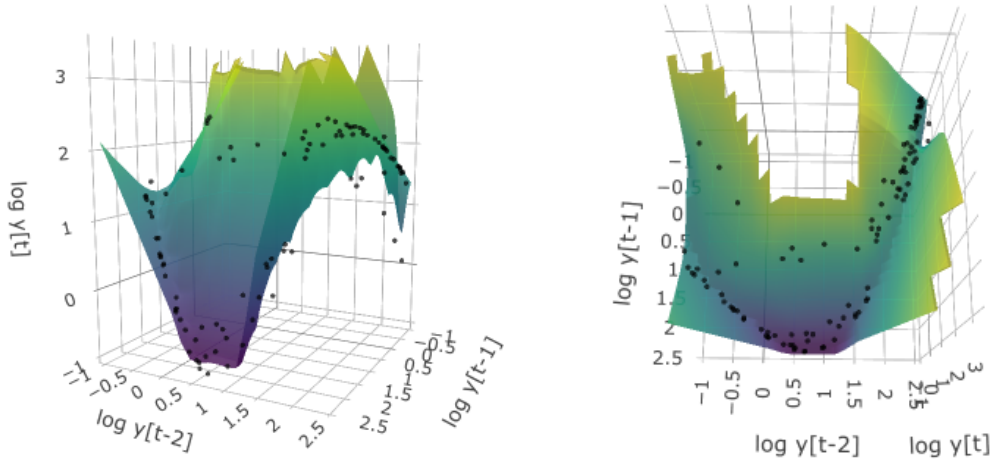


Figure 5.3: Nonparametric model fit ($T = 105$, $L = 2$, and $\mathcal{R} = 10.0$) to the time-delay embedding of $\log(y)$ simulated from the nonlinear deterministic system. Plots show the posterior mean estimate for the transition surface as a function of the first two lags. Data values are included as points.

required to capture the complex local behavior can lead to erratic interpolations and ineffective surface estimation in data-sparse regions.

5.3.3 Old Faithful data

We return to the Old Faithful waiting time series introduced at length in Section 4.3.3. There we noted that most dependence appears in the first lag, although a trend along the second lag may exist. We expect the nonparametric model to capture both the nonlinear and non-Gaussian features apparent in Figure 4.7. We consider two lags, consistent with the conclusions of Azzalini and Bowman (1990).

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 between values corresponding to the number of occupied mixture components, which ranges from two to five. Estimated transition mean surfaces, one of which is shown in Figure 5.4, are primarily driven by the first lag, with minor tilt along the second. As before, the transition mean functional is less informative for values of the first lag above 70 minutes, when the transition distribution becomes bimodal. In this region, estimates of transiton quantiles may be more appropriate than the transition mean. Inferences for quantiles over a grid of fixed lag values are easily obtained from posterior samples by following the procedure described in Section 5.2.4. Figure 5.5 shows pointwise posterior mean estimates of the 0.2 and 0.8 quantile surfaces as functions of the two lags. Credible intervals for all three surfaces (excluded for simplicity in the plots) are reasonable, falling within the range of the data.

Figure 5.6 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

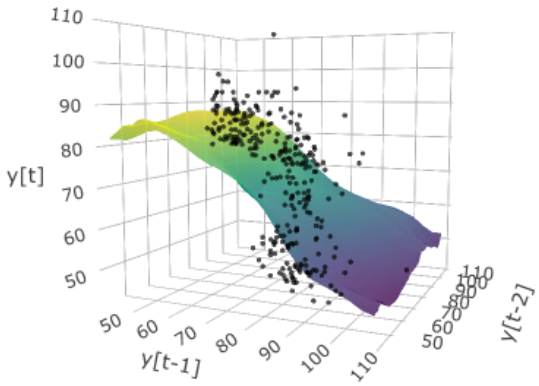


Figure 5.4: Nonparametric model fit to Old Faithful waiting times in minutes ($T = 291$, $L = 2$, and $\mathcal{R} = 5.0$), with a pointwise posterior mean estimate of the transition mean surface. Observed transitions are included as points.

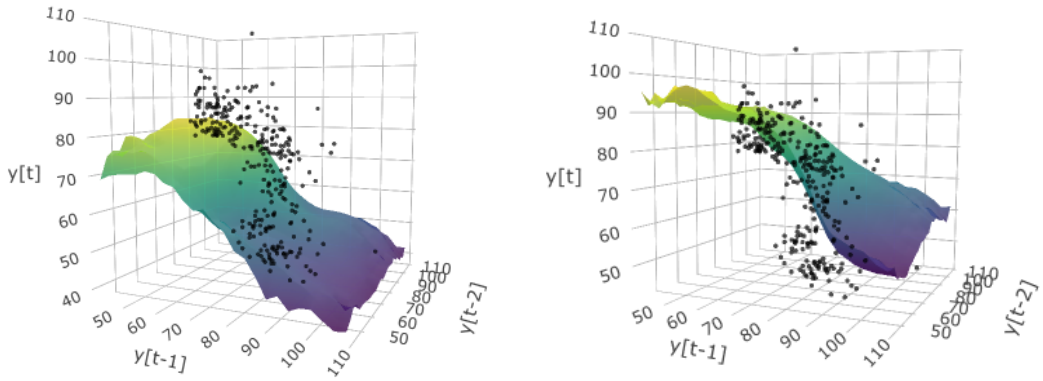


Figure 5.5: Pointwise posterior mean estimates for the 0.2 (left) and 0.8 (right) quantiles of the transition distribution of Old Faithful waiting times in minutes using the nonparametric model fit ($T = 291$, $L = 2$, and $\mathcal{R} = 5.0$). Observed transitions are included as points.

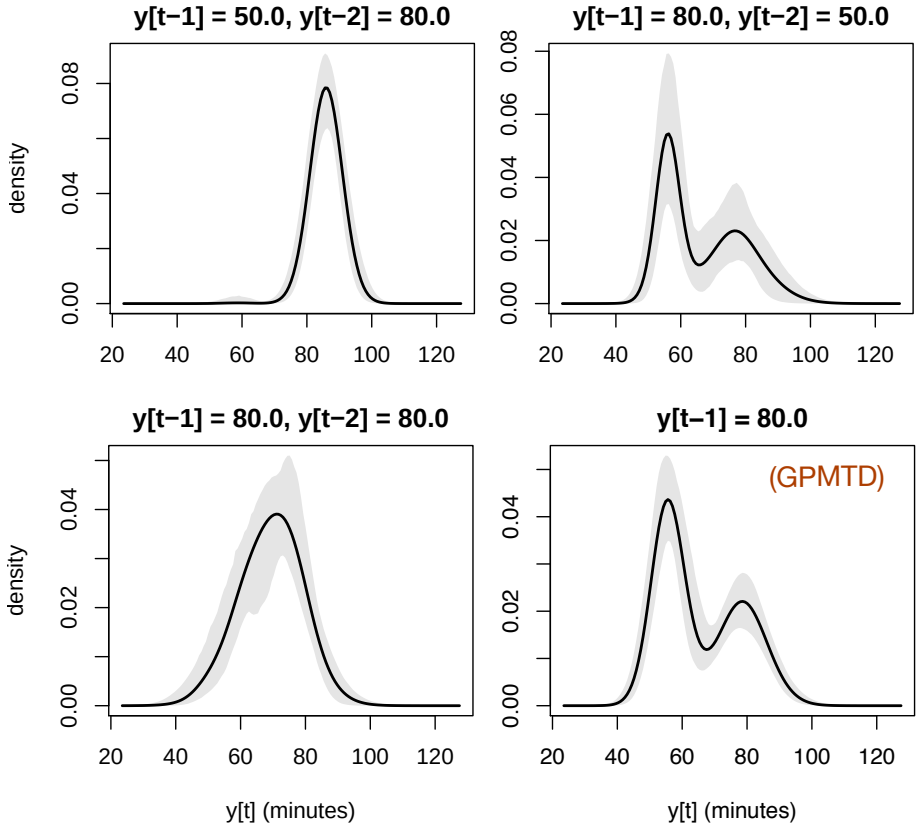


Figure 5.6: Posterior mean and 95% interval estimates for the transition density of Old Faithful waiting times at three pairs (all but bottom-right panel) of fixed values of the first two lags using the nonparametric model fit ($T = 291$, $L = 2$, and $\mathcal{R} = 5.0$). For comparison, the bottom-right panel replicates one plot from Figure 4.9 corresponding to a GPMTD model fit.

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 evidence for dependence on the second lag, and hence, second (or higher)-order dependence. Other runs show similar structure, including separation of a lower left mode for $y_{t-1} = y_{t-2} = 80$. The density estimate for $y_{t-1} = 80$ in the GPMTD model in Figure 4.9 is included in Figure 5.6 for comparison. Density dependence on the second lag is missed by the GPMTD model, in which mixture weights are constant across all values of lags.

One simple way to verify dependence on the second lag is to dichotomize by whether durations are at least 68.0 minutes, a natural cutoff in the scatter plots, and count transitions. Fixing the first lag at the high value (≥ 68.0) and the second lag at the low value, the transition counts are 68 to low and 31 to high. If the second lag is at the high value (still fixing the first lag at the high value), the transition counts are 30 to low and 61 to high. Using independent Beta$(0.5, 0.5)$ priors on these two transition probabilities yields a posterior probability greater than 0.999 that their absolute difference is at least 0.1, providing strong evidence for dependence on the second lag. Furthermore, multiple runs of the MMTD model on the binary time series with with $L = 5$ and $R = 3$ (highest mixing order) primarily favor a second-order chain depending on lags 1 and 2.

5.4 Lag selection

We now discuss extending the model (5.5) to include inferences for relevant lags (or for variable-selection generally). 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 capture the system dimensionality, 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.

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. Other approaches include adaptive shrinkage through scale mixtures

of normal priors for regression coefficients (Park and Casella, 2008; Armagan et al., 2013), g-priors (Zellner, 1986; Liang et al., 2008), and samplers that explore model spaces of differing dimensions (Green, 1995).

There is also a growing literature for variable selection in BNP regression modeling. Barcella et al. (2017) and references therein provide a recent review that discusses approaches for covariate-dependent Dirichlet process mixture, dependent Dirichlet process, and product partition models. Most approaches involve binary indicator variables associated with each covariate that either turn a contributing probability density "on" (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 lag selection, along with two variants, in Section 5.4.1. Section 5.4.2 discusses inference, including the Gibbs sampling update for lag inclusion, other modifications to MCMC, and sampling for functionals. In Section 5.4.3, we revisit the data illustrations from Section 5.3 and include two additional data sets.

5.4.1 Model extension

In the model (5.5), both mixture kernels and weights depend on $\boldsymbol{x}$, necessitating coordination across multiple parameters for model-based variable selection. To this end, we employ binary variables $\{\gamma_\ell\}$, for $\ell = 1, \ldots, L$, to indicate dependence of y on x_ℓ if γ_ℓ=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. We consider three approaches to modifying the weight kernels $\mathrm{N}_{(h)}(\boldsymbol{x}_t \mid \boldsymbol{\mu}^x, \boldsymbol{\Sigma}^x)$ to control

dependence in the mixture weights.

The first approach is analogous to the changes in the mixture kernels, wherein we replace $\beta^x_{r,\ell}$ with $\gamma_r\,\gamma_\ell\,\beta^x_{r,\ell}$ for each cluster, and replace δ^x_ℓ with $\delta^x_\ell + c\,(1-\gamma_\ell)$ with $c \gg \text{range}(\{y_t\})$. The effect of these changes is most clearly understood in the context of the sequential construction of weight kernels given in (5.7). If $\gamma_\ell = 0$, then $x_{t,\ell}$ does not contribute to any conditional mean in the sequence. Furthermore, the univariate Gaussian density for x_ℓ in (5.7) will have mean μ^x_ℓ and inflated variance $\delta^x_\ell + c$. These modifications will *effectively* (but not absolutely) nullify the contribution of x_ℓ to the overall joint kernel as long as c is very large relative to the marginal variability in the time series, $\{\boldsymbol{x}_t\}$ contains no extreme outliers, and μ^x_ℓ (which is regulated by $\boldsymbol{\mu}^x_0$ and $\boldsymbol{\Lambda}^{\mu_x}_0$) remains close to the range of the data for all $h = 1, \ldots, H$. If these conditions are met, the univariate density in (5.7) associated with $x_{t,\ell}$ becomes approximately constant with respect to $x_{t,\ell}$ and μ^x_ℓ across all $h = 1, \ldots, H$, allowing it to approximately factor and cancel out of both numerator and denominator of the weight function $\tilde{q}_h(\boldsymbol{x})$. This produces a "soft" lag-selection procedure.

The second proposed modification to the weight kernels is analogous to the first, but totally removes the effect of deselected lags. As before, we replace $\beta^x_{r,\ell}$ with $\gamma_r\,\gamma_\ell\,\beta^x_{r,\ell}$. Instead of inflating the variance of the univariate Gaussian distribution associated with $x_{t,\ell}$, we remove the univariate density altogether, replacing it with 1. This is equivalent to appropriately subsetting $\{\boldsymbol{\beta}^x_r\}$ and $\boldsymbol{\delta}^x$ prior to constructing the covariance matrix $\boldsymbol{\Sigma}^x$, reducing the dimensionality of $\text{N}_{(h)}(\boldsymbol{x}_t \mid \boldsymbol{\mu}^x, \boldsymbol{\Sigma}^x)$ 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 DP mixture model for y only. This approach reduces computational burdens and offers a clean, complete lag selection, conditional on $\boldsymbol{\gamma}$. We refer to this as the subsetting method for lag selection.

The third proposed modification provides clean lag selection by selecting active elements of $\boldsymbol{\mu}^x$ and $\boldsymbol{\Sigma}^x$ *after* construction with the full, unmodified $\{\boldsymbol{\beta}_r^x\}$ and $\boldsymbol{\delta}^x$. This is equivalent to marginalizing the weight kernels as $\int \mathrm{N}_{(h)}(\boldsymbol{x}_t)\ \mathrm{d}\bar{\boldsymbol{x}}_t$, where $\bar{\boldsymbol{x}}_t$ contains the $x_{t,\ell}$ for which $\gamma_\ell = 0$. Again if $n_\gamma = 0$, the model reduces to a DP mixture for y. While more computationally demanding, this approach helps avoid large jumps in $\{\boldsymbol{\beta}_r^x\}$ and $\boldsymbol{\delta}^x$ that result from re-purposing these parameters when $\boldsymbol{\gamma}$ changes during MCMC. This is similar in spirit to saturation MCMC schemes (Robert and Casella, 2004, pp. 444-445), as conditional on $n_\gamma < L$, the weight function is over-parameterized. However, we are less concerned with interpreting the parameters in the weight function, and thus dispense with transformations between sub-models. We refer to this as the marginalization method for lag selection.

We have elected for a single set of global $\{\gamma_\ell\}$ indicators. However, 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 cluster h, in which case the indicators become part of $\boldsymbol{\eta}_h$. This approach is adopted by Chung and Dunson (2009), who develop formal hypothesis testing for variable inclusion. For the remainder of this chapter, we assume $\boldsymbol{\gamma}$ is global.

Our proposed modifications for lag selection affect the hierarchical model in (5.8) only through the regression means in the mixture kernel distribution for y_t, which becomes $\mu^y - \sum_{\ell=1}^L \gamma_\ell\, \beta_\ell^y\, (x_{t,\ell} - \mu_\ell^x)$, through the construction of $\mathrm{N}_{(h)}(\boldsymbol{x}_t)$ in the discrete distribution for s_t, and through addition of a prior for $\{\gamma_\ell\}$. We again favor simplicity and assign independent Bernoulli(π_ℓ^γ) priors to each γ_ℓ. We use as default $\pi_\ell^\gamma = 0.25$ for $\ell = 1, \ldots, L$, promoting sparsity and dimension reduction.

5.4.2 Posterior inference

The proposed setup is minimally disruptive to the MCMC algorithm outlined in Section 5.2.3. 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, in 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 (5.13) with each $\mathrm{N}_{(h)}(\boldsymbol{x}_t)$ 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 $\mathrm{N}_{(h)}(\boldsymbol{x}_t)$ and kernel means.

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

$$\Pr(\gamma_\ell = 1 \mid \cdots) = \frac{\pi_\ell^\gamma a_\ell^\gamma}{\pi_\ell^\gamma a_\ell^\gamma + (1 - \pi)_\ell^\gamma b_\ell^\gamma}\,, \tag{5.17}$$

where

$a_\ell^\gamma = \prod_{t=L+1}^{T} \left[\mathrm{N}_{(s_t)}(\boldsymbol{x}_t \mid \gamma_\ell = 1)\ \mathrm{N}_{(s_t)}(y_t \mid \gamma_\ell = 1) \left(\sum_{j=1}^{H} \omega_j\, \mathrm{N}_{(j)}(\boldsymbol{x}_t \mid \gamma_\ell = 1) \right)^{-1} \right]$

and

$b_\ell^\gamma = \prod_{t=L+1}^{T} \left[\mathrm{N}_{(s_t)}(\boldsymbol{x}_t \mid \gamma_\ell = 0)\ \mathrm{N}_{(s_t)}(y_t \mid \gamma_\ell = 0) \left(\sum_{j=1}^{H} \omega_j\, \mathrm{N}_{(j)}(\boldsymbol{x}_t \mid \gamma_\ell = 0) \right)^{-1} \right].$

The Gaussian densities in these expressions are from (5.9), modified to reflect either $\gamma_\ell = 1$ or 0, and appropriately reflecting all other $\{\gamma_k\}_{k \neq \ell}$. The full update for $\boldsymbol{\gamma}$ proceeds by sampling from (5.17) to update γ_ℓ, for $\ell = 1, \ldots, L$.

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 variable 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, as demonstrated in Section 5.3. We recommend running multiple MCMC chains, initializing each with $\gamma_\ell = 1$, for all $\ell = 1, \ldots, L$. We begin with an adaptation phase in which $\boldsymbol{\gamma}$ is not updated, followed by a iterated adaptation and burn-in phases with the full sampler. We then run a final burn-in, followed by samples used for inference.

Before exploring inferences for transition mean and density functionals, it is helpful to assess lag dependence. 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. Alternatively, one can monitor the full conditional probabilities of inclusion in (5.17), which if averaged across posterior samples, produces a Rao-Blackwellized estimate for the posterior probability that $\gamma_\ell = 1$. Due to the mixing difficulties relating to $\boldsymbol{\gamma}$, we have seen little practical difference between the two estimates, as reliable Monte Carlo estimates of these posterior probabilities can necessitate unrealistically long MCMC runs. Shorter runs can nevertheless be informative.

Conditional on lag selection, posterior inference for functionals proceeds as in Section 5.2.4, with appropriate modifications to include $\boldsymbol{\gamma}$. For any value of y and $\boldsymbol{x}$, or over a multidimensional grid of values, samples for $f_{Y|X}$ are calculated from

$$\tilde{f}_{Y|X}(y \mid \boldsymbol{x}, \boldsymbol{\gamma}) = \sum_{h=1}^{H} \tilde{q}_h(\boldsymbol{x} \mid \boldsymbol{\gamma}) \, \mathrm{N}_{(h)}(y \mid \mu(\boldsymbol{x} \mid \boldsymbol{\gamma}), \sigma^2) \,, \tag{5.18}$$

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 $\tilde{\mathrm{E}}_{Y|X}(y \mid \boldsymbol{x}, \boldsymbol{\gamma}) = \sum_{h=1}^{H} \tilde{q}_h(\boldsymbol{x} \mid \boldsymbol{\gamma}) \, \mu_{(h)}(\boldsymbol{x} \mid \boldsymbol{\gamma})$. Likewise, analogous expressions include $\boldsymbol{\gamma}$ when using (5.15) to estimate quantiles. 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 (5.8), adjusted for $\boldsymbol{\gamma}$, for each posterior sample. While dependence on other parameters in (5.8) is implicit in the preceding expressions, we add explicit dependence on $\boldsymbol{\gamma}$ in order to emphasize the modifications necessary to include lag dependence.

The full expression for the transition density, marginalizing over all 2^L possible lag configurations, is

$$\tilde{f}_{Y|X}(y \mid \boldsymbol{x}) = \sum_{\boldsymbol{\gamma} \in \{0,1\}^L} \sum_{h=1}^{H} \tilde{q}_h(\boldsymbol{x} \mid \boldsymbol{\gamma}) \; \mathrm{N}_{(h)}(y \mid \mu(\boldsymbol{x} \mid \boldsymbol{\gamma}), \sigma^2) \; \Pr(\boldsymbol{\gamma}) \,, \tag{5.19}$$

where $\Pr(\boldsymbol{\gamma})$ can refer to either the prior or marginal posterior of $\boldsymbol{\gamma}$. The expression for the joint prior mass function is $\prod_{\ell=1}^{L} (\pi_\ell^\gamma)^{\gamma_\ell} \, (1 - \pi_\ell^\gamma)^{1-\gamma_\ell}$. In practice, we bypass the burdensome outer summation in (5.19) and instead calculate (5.18) across MCMC samples, which yields the desired posterior inferences marginalized with respect to the posterior of all parameters, including $\boldsymbol{\gamma}$.

Calculation of transition density and mean estimates requires the full $\boldsymbol{x} \in \mathbb{R}$, 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 $\boldsymbol{\gamma} = (1, 1, 0)$. They may use only posterior samples for which this lag configuration was active (taking into account the order of full-conditional sampling), provided a sufficiently long MCMC chain. They may then calculate

(5.18) 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 most or all posterior samples coincide with a particular configuration, one may proceed in the same way, substituting default (or average) values in for the inactive elements of $\boldsymbol{x}$ and examining inferences 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 5.3.

5.4.3 Data illustrations incorporating lag selection

We now revisit the analyses from Section 5.3 with the full model including lag selection, and include two additional examples. For each example, multiple MCMC chains were randomly initialized using the strategy described in Section 5.2.3, with four iterated burn-in and adaptation stages followed by 400,000 burn-in samples. The next 600,000 iterations were then thinned to 2,000 for inference (1,000 samples were used in the first illustration and for multidimensional surface plots). Each of the subsetting (second) and marginalization (third) methods for lag inferences were employed and compared. MCMC runs for this section fix G_0, as updating the parameters in the base measure sometimes hindered exploration in $\boldsymbol{\gamma}$.

Simulated data: linear autoregression

We begin by demonstrating 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, correctly identifying the lag structure and recovering the parameters. 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 a series of length $T = 105$ with a lag horizon of $L = 5$ (so that 100 observations contribute to the likelihood), DP truncation at $H = 35$, and default prior settings. Both methods for lag selection behave similarly and recover the true structure, with all chains decisively selecting the first two lags only.

Figure 5.7 provides trace plots for key quantities from one model fit (using the marginalization method of lag selection) including the log-likelihood, number of occupied clusters, selection indicators for the first four lags, the observation (innovation) variance for the most populated cluster, the first three β^y coefficients for the most populated cluster, the center μ^y for the most populated cluster, and the intercept $(\mu^y + \sum_{\ell=1}^{L} \gamma_\ell \, \beta_\ell^y \, \mu_\ell^y)$ for the most populated cluster, thinned to 1,000 samples. The trace for the log-likelihood indicates that the chain is no longer traversing across substantially different cluster configurations. Most observations belong to one cluster throughout MCMC. Lag indicator parameters γ_1 and γ_2 are in the "on" position for all inference samples (and the full conditional probabilities are likewise numerically 1), while trace plots for the γ_ℓ, $\ell = 3, 4$ are in the "off" position for all inference samples. The lag 5 indicator (not shown) briefly switches to the "on" position, but remains "off" for the overwhelming majority of iterations. Trace plots for the kernel parameters faithfully track the true values (note that the sign is switched for the coefficients in the model formulation). The only exception in this chain is μ^y, which in the model is replaced by μ_ℓ^x parameters in the lag

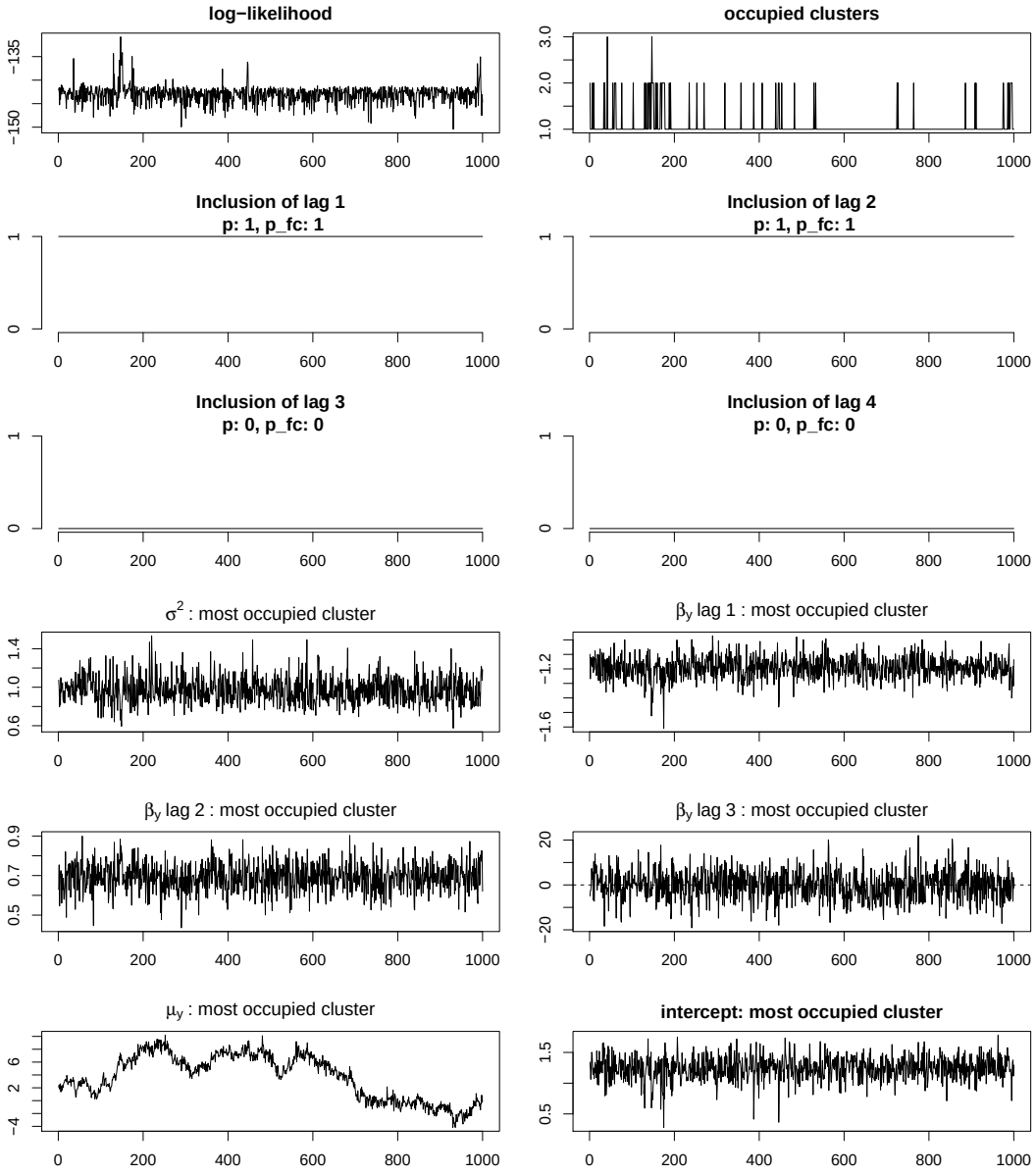


Figure 5.7: MCMC trace plots for the nonparametric model fit to the simulated second-order autoregression (marginalization method fo lag selection). In the lag-inclusion plots (second and third row), p refers to the Monte Carlo estimate of the posterior probability that $\gamma_\ell = 1$, and *p_fc* refers to the Rao-Blackwellized estimate.

summands, and thus over-parameterized for this stationary time series. Note, however, that the intercept (with true value 1.25) is stable and correctly estimated. 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 true 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, although the marginalization method for lag selection produces erratic credible intervals while the subset method produces tight and accurate intervals. Furthermore, marginal posterior standard deviations are only slightly higher (5% for one of the coefficients, 11% for the intercept) than 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.

Simulated data: single lag

Model runs ($T = 105$, $L = 5$, $H = 40$) fit to the nonlinear simulation with one active lag have mixed results. No run identifies that only lag 2 is active, but several decisively select both lags 2 and 4, with no visits to other configurations in the inference samples. Selecting these two lags is reasonable given that the data reside in two diagonal quadrants of the (y_{t-2}, y_{t-4}) lag-embedding space, as seen in Figure 5.8. All three runs employing the subset lag selection method with lower prior signal-to-noise ($\mathcal{R}$, and weaker variance) find the lag (2,4) configuration, as did one of each of the other prior specification/selection method combinations. The problem with most other runs was failure to deselect other lags. As expected,

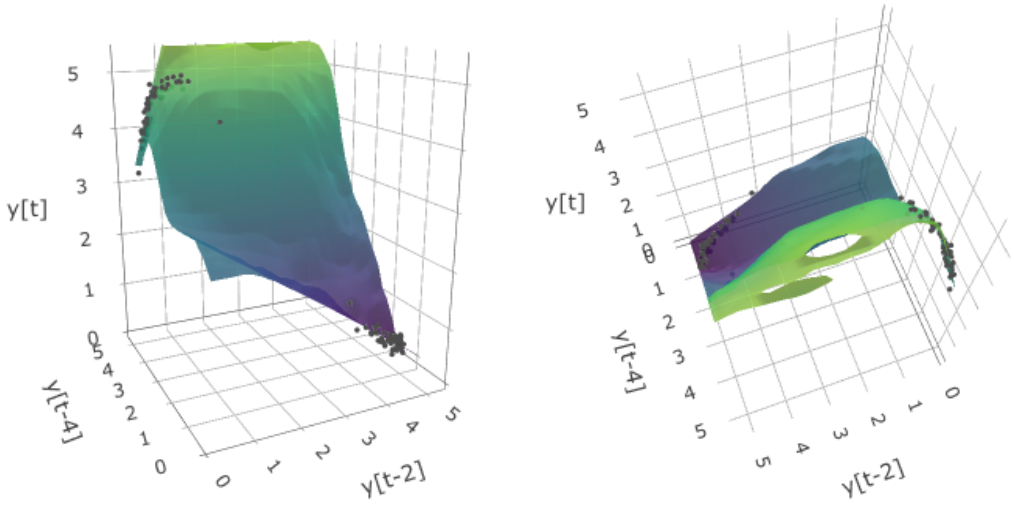


Figure 5.8: Nonparametric model fit to the single-lag dynamical simulation with noise ($T = 102$, $L = 5$, lags 2 and 4 selected). Both plots show the pointwise posterior mean estimate of the transition surface.

runs with higher $\mathcal{R}$ estimate σ^2 (between 0.010 and 0.015) to be lower and closer to the truth (0.0081). The estimated transition mean surface in lags 2 and 4, for one of the three runs in the successful group, is reported in Figure 5.8.

Simulated data: time-delay embedding

All model runs ($T = 105$, $L = 5$, $H = 40$) fit to the two-dimensional nonlinear simulation maintain all lags active throughout MCMC. All chains within each prior specification appear to stabilize at similar log-likelihood traces, and posterior mean residuals plotted against posterior mean expected values (using the marginalization method for lag selection, and fixing lags 3-5 at mean values) for one of the runs do not indicate any pattern (not shown).

Results for lag selection are identical in the $T = 505$ case. Furthermore, different chains settle on distinct log-likelihood traces, indicating a highly separated, multimodal posterior distribution.

Old Faithful data

Model runs ($T = 294$, $L = 5$, $H = 40$) fit to the Old Faithful time series also have mixed results. Only one run using the marginalization method deselects any lags, retaining lags 1, 4, and 5. The subset method with higher $\mathcal{R}$ selects the first lag, the first two lags, and all lags in its respective runs. In the lower $\mathcal{R}$ case, all lags, all but lag 4, and the first two lags are are selected. In all cases, there is no change in $\boldsymbol{\gamma}$ within inference samples of a chain. Considering the noise level in this series, we would tend to trust the model with lower prior signal-to-noise ratio, but only one of three runs yields the results we expect. In that one case, transition mean and density inferences are similar to those given in Section 5.3.3.

Pink salmon data

Model runs ($T = 30$, $L = 5$, $H = 40$) fit to the pink salmon data, analyzed in Sections 3.5.2 and 4.3.4, demonstrate sensitivity to prior specification. Recall that the data consist of log-transformed annual escapement of pink salmon in a stream in Alaska, and we anticipate dependence in even lags. All runs with the weaker prior signal-to-noise ratio ($\mathcal{R} = 5.0$) setting end with no lags selected, although one run has lag 2 active for many inference samples. This appears to be a consequence of inappropriately setting $a_\alpha = 10.0$, too high for such a small sample size. In contrast, the runs with higher $\mathcal{R} = 25.0$ and $a_\alpha = 15.0$ generally keep most lags active, probably due to over-fitting.

Another set of runs comparing ($\mathcal{R} = 5.0,\ a_\alpha = 5.0$) to ($\mathcal{R} = 7.0,\ a_\alpha = 7.0$) produces more reasonable and consistent results. The marginalization method for lag selection still struggles, deselecting all lags in several runs. All completed runs of the subset method decisively select lag 2 only, and do exhibit some mixing in lag configuration. Figure 5.9 reports posterior inferences for the transition mean

as a function of lag 2 with other lags fixed at mean values, from a run using the subset method and $\mathcal{R} = 7.0$. Comparison with Figure 4.10 reveals differences between this and the GPMTD model. The nonparametric model yields a more linear transition through the bulk of points, which may be restrictive, although the estimated transition density at $\log(y_{t-2}) = 8.0$ exhibits slight right skew (not shown). Also, credible intervals more appropriately vary with data abundance.

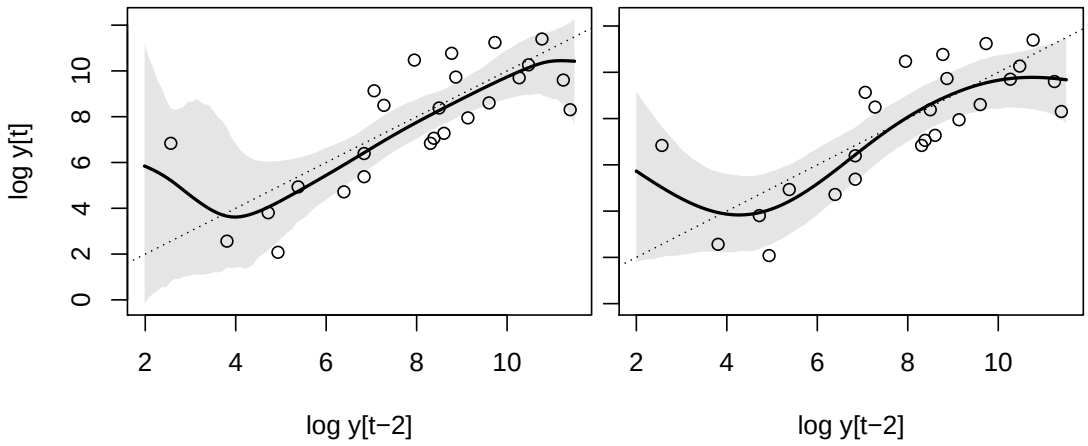


Figure 5.9: Nonparametric model fit (left) to the logarithm of annual pink salmon escapement ($T = 30$, $L = 5$, lag 2 selected). Figure 4.10, from the analogous GPMTD model fit, is replicated on the right for comparison. The plots include pointwise posterior mean estimates and 95% credible intervals 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.

5.5 Transition density estimation performance

Transition density estimation is a primary objective of the models proposed in this chapter and in Chapter 4. The Gaussian process mixture transition distribution (GPMTD) model is simpler in specification and implementation, but more limited in density flexibility and in lag dependence. To compare density estimation between the models, we fit each 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 single-lag nonlinear system in (5.16). The first modification replaces the additive Gaussian error with multiplicative log-normal error. Specifically, transitions were generated from

$$y_t = y_{t-2} \exp(2.6 - y_{t-2} + \epsilon_t)\,, \qquad \epsilon_t \overset{\text{iid}}{\sim} \mathrm{N}(0, (0.09)^2)\,, \tag{5.20}$$

corresponding to a log-normal transition density with log-mean equal to $\log(y_{t-2}) + 2.6 - y_{t-2}$ and log-scale equal to 0.09. This produces right skew and heteroscedastcity in the transition distribution, which continues to depend exclusively on the second lag. The lag scatter plot in Figure 5.10 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 log-normal, with each parameter depending on a separate lag. The lag scatter plot in Figure 5.11 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, the first 1,000 (post burn-in) observations were reserved for model fitting, and a validation set of size 1,000 was randomly sampled from the subsequent 9,000 observations.

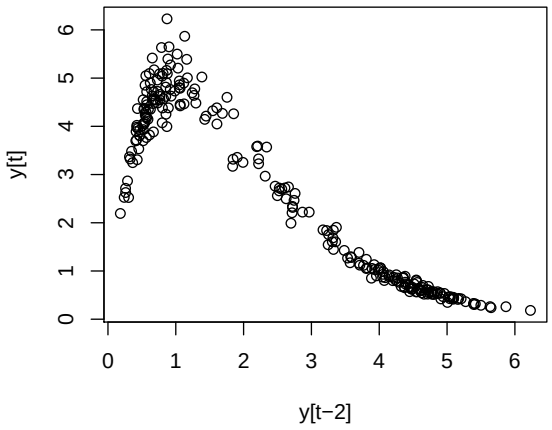


Figure 5.10: Lag scatter plot from the modified single-lag nonlinear simulation with log-normal transition density.

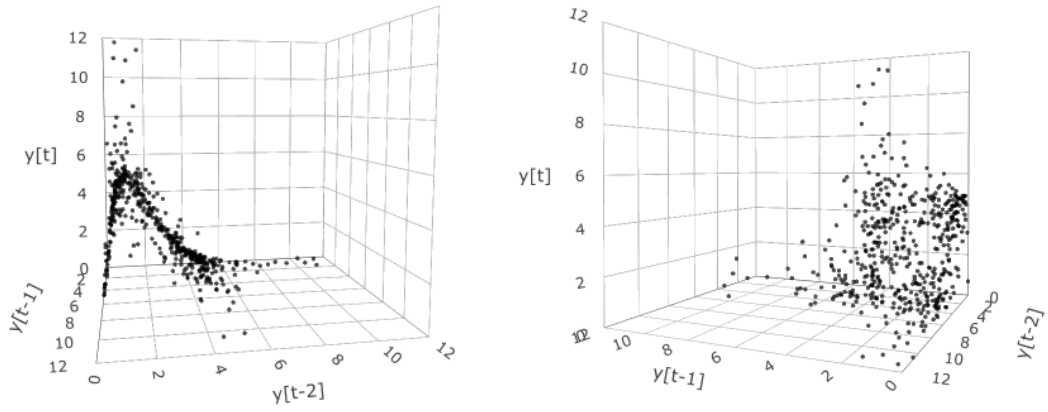


Figure 5.11: Lag scatter plot from the modified two-lag nonlinear simulation with log-normal transition density.

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 heteroscedastcity, subtle departures from Gaussianity, and subtle variation in lag dependence.

The following models were fit using default settings to simulated time series from the original system (5.16) and both modifications: The GPMTD with $L = 5$,

the proposed nonparametric model (which we denote as the BNP-WMAR model, for Bayesian nonparametric, weighted mixture of autoregressive models) with $L = 2$ and no lag selection (and G_0 fixed), and the BNP-WMAR model with $L = 5$ and lag selection (subseting method, and G_0 fixed). Two GPMTD model runs included 1,000 thinned samples each, following adaptation and 50,000 burn-in iterations. BNP-WMAR model runs included three chains of 1,000 thinned posterior samples each, following repeated rounds of adaptation and burn-in. Variability in reported loss metrics can primarily be attributed to chains exploring distinct posterior modes.

Each posterior sample was used to create density ordinates, denoted $\hat{p}(y_t \mid \boldsymbol{y}_{t-1})$ and calculated from the density analogue of the first line of (4.2), from (5.14), and from (5.18). With 2,000 replicate simulation draws $\{y_j^{(i)}\}_{i=1}^{2,000}$ from the true data-generating distribution (with density $p_{\text{true}}(y_t \mid \boldsymbol{y}_{t-1})$) for each validation pair $\{(y_j, \boldsymbol{x}_j)\}_{j=1}^{1,000}$, we approximated the Kullback-Leibler divergence using

$$\begin{aligned} \mathrm{D}_{\mathrm{KL}}(p_{\text{true}} \parallel \hat{p}) &\equiv \int p_{\text{true}}(y \mid \boldsymbol{x}) \log\left(\frac{p_{\text{true}}(y \mid \boldsymbol{x})}{\hat{p}(y \mid \boldsymbol{x})}\right) \mathrm{d}y \\ &\approx \sum_{i=1}^{2,000} \log\left(p_{\text{true}}(y^{(i)} \mid \boldsymbol{x})\right) - \log\left(\hat{p}(y^{(i)} \mid \boldsymbol{x})\right), \end{aligned} \tag{5.21}$$

averaged over validation observations and posterior simulations. Let $\hat{\mathrm{D}}_{\mathrm{KL}}$ denote the result. This loss metric is reported in Table 5.1 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 some cases, the only two completed runs).

In the single-lag, normal case, the GPMTD fit to 300 observations sets the standard with a correctly specified error distribution and nearly all weight on lag 2.

		$\hat{D}_{KL}$			
Simulation	Model	70 obs.		300 obs.	
Single-lag, normal	GPMTD(5)	0.908	1.160	0.286	0.287
	BNP-WMAR(2*)	1.827	1.992	0.272	0.393
	BNP-WMAR(5)	0.762	0.791	0.586	0.723
Single-lag, log-normal	GPMTD(5)	1.121	1.128	0.600	0.601
	BNP-WMAR(2*)	1.423	1.676	0.345	0.373
	BNP-WMAR(5)	0.696	1.110	0.424	0.483
Two-lag, log-normal	GPMTD(5)	1.401	1.405	1.101	1.103
	BNP-WMAR(2*)	1.536	1.817	0.936	1.001
	BNP-WMAR(5)	1.385	1.826	1.232	1.351

Table 5.1: Comparison of single-step transition density estimation performance, measured by K-L divergence, for the GPMTD and nonparametric models for three simulations and two sample sizes. The numbers in parentheses are L, the number of lags considered in each fit, and $*$ indicates no lag selection. Within each set, the minimum (left) and maximum (right) losses across runs are reported.

One GPMTD run with 70 observations performs poorly due to selecting lag 4. The nonparametric model without lag selection performs poorly in the small sample, presumably a consequence of fitting the first lag. This burden is alleviated in the large sample. Lag selection helps the nonparametric model in the small sample, which surprisingly outperforms the GPMTD. However, no lags are deselected in the large sample, resulting in deteriorated model efficiency.

The GPMTD model is misspecified in the other two scenarios, which is evident for large samples. Results for the nonparametric model fits to the single-lag, log-normal simulation are similar to before in that lag selection dramatically helps for small samples and detracts in large samples from failure to deselect. In the two-lag, log-normal scenario, the nonparametric model with $L = 2$ and no lag selection is the most appropriate of the three specifications. The other models, however, perform well on the small sample by selecting primarily lag 2 and apparently leveraging their mixture structures to accommodate the small number of "outliers."

Overall, this simulation study elucidates some strengths and weaknesses of the proposed models. Despite its limitations, the GPMTD model is surprisingly resourceful, using mixture components for flexibility in addition to effective selection of a primary active lag. Lag selection in the nonparametric model, as currently implemented, can be effective for short time series, even outperforming the GPMTD model. Long time series tend to sharpen and separate posterior modes, hindering lag selection in the nonparametric model.

5.6 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 global 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. Continuous covariates are also readily accommodated by extending the $\boldsymbol{x}_t$ vectors.

Results from the base model are promising, faithfully capturing known or anticipated features of the data examples. However, inference for relevant lags remains challenging. Indeed, this problem has proven challenging for linear models, and extending to globally nonlinear responses amplifies its complexity. One important consideration 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 time-delay embedding. Another issue is that binary lag-inclusion parameters do not quantify relative contributions of lags to the transition function, which can be

measured through decompositions (e.g., as in Sobol, 2001), or to the transition density itself. Ideally, weak dependence would manifest in the posterior probability of inclusion. Another challenge arises from the fact that lags, are assumed to be correlated. Consequently, different lag configurations of the same order can be (nearly) equally effective in forecasting (e.g., Figure 2.5).

Although the marginalization method for lag selection is conceptually more appealing and appears to promote better mixing in some cases, the subsetting method has consistently produced superior results in terms of lag selection. With both approaches, it has proven difficult to get lags to turn off, and even harder to get them back on. Our experience has been that lag selection within this model suffers from acute prior sensitivity, which is not surprising. The three 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 decisively identified with data. Consequently, assumptions must be made and encoded through the prior settings. We strongly recommend completing a thorough preliminary and exploratory analysis of data, with visualization if possible. Even so, we advise that practitioners run several models with a variety of prior signal-to-noise ratio and flexibility (through α and possibly δ^x) settings, each with multiple MCMC chains.

Notwithstanding these theoretical and practical challenges, lag selection is critical for dimension reduction and is an integral part of this work. One avenue for improving performance in this model would be to employ a more intricate prior, or strategies suggested by O'Hara et al. (2009). Local lag selection is an option that was explored by Chung and Dunson (2009). While this approach again raises questions about the tradeoff between signal and noise, it can improve modeling efficiency in cases of truly local dependence. Local selection could also potentially

improve mixing in MCMC, a topic that is addressed in Chapter 6.

We have found that fixing (or nearly fixing) the parameters in G_0 improves numerical stability and lag selection. Our experience is that updating G_0 makes it difficult to deselect lags. Another source of numerical instability in MCMC can be avoided by fixing α.

The proposed model is, in principle, sufficiently flexible to approximate intricate transition densities with nonlinear dependence on multiple lags. 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. Unfortunately, updates for each component-specific parameter vector are not easily distributable, due to their appearance in the denominator of the likelihood through the normalized weights. Simplifications to the model, particularly in the weight functions, can partially alleviate the computational burden at the cost of some model flexibility.

Chapter 6

Conclusion

As stated in the introduction, the primary objectives of this work are to contribute Bayesian statistical methodology and modeling strategies to estimate transition distributions, particularly when these distributions exhibit non-Gaussianity and/or nonlinear dependence on multiple lags, with emphasis in detecting and exploiting low-order dependence. To this end, we have proposed and explored flexible time-series models under a common theme of mixture modeling methods. The models, appropriate for time series of moderate length, can be useful for recovering nonlinear dynamics in systems for which it is difficult to justify use of any particular parametric model or to observe all relevant variables. Bayesian inference for these statistical models provides a coherent framework for learning the dynamics when faced with noisy data and structural uncertainty.

The models for discrete-state time series in Chapters 2 and 3 utilize finite mixtures, together with novel priors, to provide a flexible means of softly selecting relevant lags. The mixture transition distribution model supplies a parsimonious and interpretable foundation for these models. The semiparametric extension for continuous state spaces in Chapter 4 is useful for detecting dependence on a single unknown lag, and unlike similar models in the literature, accommodates

nonlinear dynamics. The fully nonparametric model in Chapter 5 forgoes the simple mixture transition distribution structure to capture general lag dependence, both in the transition mean and density. We further accommodate multiple lags while encouraging low-dimensional dependence.

We have focused throughout this thesis on introducing methodology and exploring its practical application to time series with various characteristics. While computation plays a prominent role and care has been taken with respect to algorithms and implementation, opportunities to improve computational efficiency remain. For example, the double-mixture structure in Chapter 3 necessitates a combinatorial search over lag configurations for each latent indicator variable. However, these indicators rarely change in posterior sampling, as the goal is to concentrate the majority of allocations to a single configuration. Seemingly redundant calculations also exist for the nonparametric model in Chapter 5, wherein kernel weights must be evaluated for each mixture component at each time point. While these computationally intensive steps do not preclude the analyses in this thesis, they do present bottlenecks limiting the potential for these models on longer time series with more complex dynamics. It may be worthwhile to explore modifications or approximations that reduce these computational burdens.

Another challenge consistently encountered in this work has been designing MCMC algorithms that adequately explore the complex posterior distributions inherent in mixture modeling. In Chapters 2 and 3, we employed occasional "jumpstart" Metropolis steps, and in Chapter 5, adaptation targeting low acceptance rates encouraged exploration. It was, nevertheless, common for parallel chains to settle in distinct modes. Beyond the strategies noted in Section 5.6, algorithms incorporating gradient-assisted or tempered MCMC may help in this regard.

The methods presented in this thesis contribute additional tools for Bayesian

inference of transition distributions and their associated lag dependence structures. The methods, designed for different settings, and with varying complexity and generality, aim to consolidate order selection and estimation into a single model-based framework that appropriately quantifies uncertainty. We hope that the various models for probability vectors, discrete-state Markov chains, and continuous-state time series prove useful and find application beyond the examples given.

Reflections

My interest in pursuing study of Bayesian nonparametrics quickly expanded during my first year of coursework at UC, Santa Cruz, as appreciation for both the challenge and utility of time-series methods grew with intrigue for their elegance. The following year, the opportunity to blend these interests with the fascinating (and likewise elegant) field of dynamical systems provided an exciting start to my research. Not surprisingly, the subsequent path of my work follows a nonlinear correspondence with its presentation. I began with the nonparametric model of Chapter 5 and variable selection in clustering, motivated by time-delay embedding. Challenges in creating a framework for lag selection led to another branch and a shift in focus toward symbolic dynamics and Markov chains. The priors for sparse probability vectors of Chapter 2 grew out of Bayesian implementation for the mixture transition distribution (MTD) model, wherein the standard Dirichlet priors left the models wanting. Accommodation for interactions of multiple lags prompted exploration of the extension in Chapter 3. The Gaussian process MTD model provides another (simpler) alternative to the fully nonparametric model and rounds out my work with MTD formulations.

I, along with Professor Adrian Raftery (the originator of the MTD model), maintain optimism for continuing potential of the MTD framework. Concurrent

with the work presented herein, I pursued hidden Markov/switching mixture approaches, which ultimately proved less compatible with my objectives (as discussed in Section 1.1.1). I still see potential in this area, and intend to explore computationally feasible approaches to Markov-switching and other models incorporating the new priors with MTD (or similar) structures. I also remain enthusiastic about the relatively small intersection of statistical time series methods with dynamical systems. As my skill and understanding in this area mature, I aspire to further leverage my training toward meaningful contributions to a field currently dominated by algorithmic tools.

My development as a researcher might be accurately characterized as a shedding of naiveté. It is humbling that in the course of research, growing awareness of one's own ignorance invariably outpaces accumulation of knowledge. Nevertheless, as I have come to appreciate the difficulty of the problems pursued, I also appreciate the growth that accompanies struggle. It is truly exhilarating to become conversant with scientific literature, to contribute in meetings with advisors and colleagues, and to discover common core principles in diverse disciplines, despite wide variety in their presentation and application. Through much practice, I have enjoyed refining fundamental skills such as prioritizing, organization, planning, and revision. Each step of the process for a statistician, of gathering and understanding data, learning methods, developing methods, implementing and testing, communicating, and iterating as necessary, has unique appeal and contributes to a cycle that is both enjoyable and rewarding.

Appendix A

Implementation Details for MTD Models with Sparse Probability Vectors

A.1 SBM correction for δ_k

Here we derive the correction factor (2.7) proposed to account for sparsity when using the SBM prior as an extension of a Dirichlet prior. We define a probability θ_j to be *negligible* if the corresponding Z_j is drawn from the first Beta$(1, \eta)$ component, or if Z_h was drawn from the third Beta$(\eta, 1)$ distribution for some $h < j$. Let W be the total number of non-negligible probabilities in $\boldsymbol{\theta}$. We can write $W = W_1 - W_2$ where W_1 is the minimum j such that $\xi_j = 3$ (Z_j comes from the large component) or J, whichever is smaller, and W_2 counts the number of times $\xi_j = 1$ (Z_j comes from the small component) among $j = 1, \ldots, W_1 - 1$. W_1 follows a truncated geometric distribution with success probability π_3 and has expectation $\mathrm{E}(W_1) = (1 - (1 - \pi_3)^J)/\pi_3$. Conditional on W_1, W_2 is binomial with

$W_1 - 1$ trials and success probability π_1, and thus has conditional expectation $\mathrm{E}(W_2 \mid W_1) = (W_1 - 1)\pi_1$. Putting these together, we find

$$\begin{aligned}\mathrm{E}(W) = \mathrm{E}(W_1) - \mathrm{E}_{W_1}[\mathrm{E}(W_2 \mid W_1)] &= \mathrm{E}(W_1) - \mathrm{E}_{W_1}[(W_1 - 1)\pi_1] \\ &= \mathrm{E}(W_1) - \pi_1 \,\mathrm{E}(W_1) + \pi_1 \,.\end{aligned}$$

Substituting $\mathrm{E}(W_1)$ and dividing by J yields (2.7). In the $\pi_3 = 0$ case, we have $W_1 = J$ with probability 1, so that $\mathrm{E}(W_1) = J$.

A.2 Marginal distributions

We report the marginal distributions of observations associated with the Dirichlet, SDM, and SBM models for probability vectors. These distributions can be useful for computing Bayes factors in addition to facilitating the MCMC algorithm described in Appendix A.3.2.

Consider a sequence of independent random variables $\{s_t\} \in \{1, \ldots, J\}^N$ with common distribution $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)$. Given $\boldsymbol{\theta}$, the probability of the sequence is $\prod_t \theta_{s_t} = \theta_1^{n_1} \cdots \theta_J^{n_J}$ where the sufficient statistics in $\boldsymbol{n} = (n_1, \ldots, n_J)$ count the occurrences of each category. If the ordering t is not important, the probability is multiplied by the multinomial coefficient $N!/(n_1! \cdots n_J!)$.

If $\boldsymbol{\theta}$ follows a Dirichlet distribution with shape parameter vector $\boldsymbol{\alpha}$, then the marginal (prior predictive) distribution of $\{s_t\}$ is given by

$$\begin{aligned}p(\{s_t\}) = \int p(\{s_t\} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}) \,\mathrm{d}\boldsymbol{\theta} &= \frac{\Gamma(\sum_j \alpha_j)}{\prod_j \Gamma(\alpha_j)} \int \theta_1^{\alpha_1 + n_1} \cdots \theta_J^{\alpha_J + n_J} \,\mathrm{d}\boldsymbol{\theta} \\ = \frac{\Gamma(\sum_j \alpha_j)}{\prod_j \Gamma(\alpha_j)} \frac{\prod_j \Gamma(\alpha_j + n_j)}{\Gamma(\sum_j \alpha_j + n_j)} &= \frac{\mathrm{MVB}(\boldsymbol{\alpha} + \boldsymbol{n})}{\mathrm{MVB}(\boldsymbol{\alpha})} \,, \qquad \text{(A.1)}\end{aligned}$$

where $\mathrm{MVB}(\cdot)$ denotes the multivariate beta function.

Now suppose $\boldsymbol{\theta}$ follows the SDM distribution with parameters $\boldsymbol{\alpha}$ and β, and $w_j = \prod_{h=1}^{J} \Gamma(\alpha_h + \beta 1_{(h=j)})$, then the marginal distribution of $\{s_t\}$ is given by

$$\begin{aligned} p(\{s_t\}) &= \int \theta_1^{n_1} \cdots \theta_J^{n_J} \sum_{j=1}^{J} \frac{w_j}{\sum_{h=1}^{J} w_h} \operatorname{Dir}(\boldsymbol{\theta} \mid \boldsymbol{\alpha} + \beta \boldsymbol{e}_j) \, \mathrm{d}\boldsymbol{\theta} \\ &= \sum_{j=1}^{J} \frac{w_j}{\sum_{h=1}^{J} w_h} \frac{\operatorname{MVB}(\boldsymbol{\alpha} + \beta \boldsymbol{e}_j + \boldsymbol{n})}{\operatorname{MVB}(\boldsymbol{\alpha} + \beta \boldsymbol{e}_j)} , \end{aligned} \tag{A.2}$$

where $\boldsymbol{e}_j$ denotes a vector of 0s with a 1 in the jth entry.

Before considering the SBM model, we first obtain the marginal distribution under the generalized Dirichlet distribution (Connor and Mosimann, 1969). Working directly with the stick-breaking weights, we have $p(\boldsymbol{Z}) = \prod_{j=1}^{J-1} \operatorname{Beta}(Z_j \mid a_j, b_j)$ and $p(\{s_t\} \mid \boldsymbol{Z}) = Z_1^{n_1} (1 - Z_1)^{\sum_{j=2}^{J} n_j} \times \cdots \times Z_{J-1}^{n_{J-1}} (1 - Z_{J-1})^{n_J}$. Putting these together and integrating over $\boldsymbol{Z}$ results in $p(\{s_t\}) = \prod_{j=1}^{J-1} g_j(a_j, b_j, \boldsymbol{n})$ where

$$g_j(a_j, b_j, \boldsymbol{n}) = \frac{\Gamma(a_j + b_j)}{\Gamma(a_j^* + b_j^*)} \frac{\Gamma(a_j^*)}{\Gamma(a_j)} \frac{\Gamma(b_j^*)}{\Gamma(b_j)} ,$$

with $a_j^* = a_j + n_j$, and $b_j^* = b_j + \sum_{h=j+1}^{J} n_h$. Using a similar approach, it can be shown that under the SBM model with parameters $\boldsymbol{\pi}$, η, $\boldsymbol{\gamma}$, and $\boldsymbol{\delta}$, we have

$$p(\{s_t\}) = \prod_{j=1}^{J-1} \left[\pi_1 \, g_j(1, \eta, \boldsymbol{n}) + \pi_2 \, g_j(\gamma_j, \delta_j, \boldsymbol{n}) + \pi_3 \, g_j(\eta, 1, \boldsymbol{n}) \right] . \tag{A.3}$$

A.3 MCMC algorithm details

Following the hierarchical MTD model outlined in (2.10), the joint posterior distribution of all unknown parameters is given up to proportionality:

$$p(\{z_t\}, \boldsymbol{\lambda}, \boldsymbol{Q} \mid \{s_t\}) \propto p(\boldsymbol{\lambda}) \prod_k [p\left((\boldsymbol{Q})_{\cdot,k}\right)] \prod_t [p(z_t \mid \boldsymbol{\lambda})\, p(s_t \mid z_t, \boldsymbol{Q}, \{s_{t-\ell}\}_{\ell=1}^{L})]\,, \tag{A.4}$$

where $(\boldsymbol{Q})_{\cdot,k}$ denotes the kth column of $\boldsymbol{Q}$.

A.3.1 Original algorithm

MCMC sampling for the original hierarchical MTD structure (Insua et al., 2012) can be achieved entirely with Gibbs updates. This is also the case when substituting in the SDM and SBM priors. A Gibbs sampler cycles through the parameters, drawing updates from the conditional distributions given below.

- $\Pr(z_t = \ell \mid \cdots) \propto \Pr(z_t = \ell \mid \boldsymbol{\lambda})\, p(s_t \mid z_t = \ell, s_{t-\ell}, \boldsymbol{Q}) = \lambda_\ell \left(\boldsymbol{Q}\right)_{s_t, s_{t-\ell}}$, independently for each $t = L+1, \ldots, T$.

- $p(\boldsymbol{\lambda} \mid \cdots) \propto p(\boldsymbol{\lambda}) \prod_t p(z_t \mid \boldsymbol{\lambda}) = \mathrm{Dir}(\boldsymbol{\lambda} \mid \boldsymbol{\alpha}_\lambda) \prod_t \lambda_{z_t}$, a standard Dirichlet-multinomial update using the counts of z_t in each of $\{1, \ldots, L\}$. In the case of a SDM prior for $\boldsymbol{\lambda}$, this becomes the standard SDM update given in Section 2.2.1. In the case of a SBM prior for $\boldsymbol{\lambda}$, this becomes the standard SBM update given in Section 2.2.2, which draws updated latent stick-breaking weights and constructs $\boldsymbol{\lambda}$.

- $p((\boldsymbol{Q})_{\cdot,k} \mid \cdots) \propto p((\boldsymbol{Q})_{\cdot,k}) \prod_{\{t: s_{t-z_t} = k\}} p(s_t \mid z_t, (\boldsymbol{Q})_{\cdot,k}, s_{t-z_t})$
  $= \mathrm{Dir}((\boldsymbol{Q})_{\cdot,k} \mid \boldsymbol{\alpha}_Q) \prod_{\{t: s_{t-z_t} = k\}} (\boldsymbol{Q})_{s_t,k}$, which is again a standard Dirichlet-multinomial update using transition counts. In the case of a SBM

prior for the kth column of $\boldsymbol{Q}$, this becomes the standard SBM update given in Section 2.2.2, which draws updated latent stick-breaking weights and constructs the column of $\boldsymbol{Q}$.

This Gibbs sampler can be modified to include an update for $\boldsymbol{\pi}_k$ together with the update for the kth column of $\boldsymbol{Q}$ since the only component of the posterior in (A.4) dependent on $\boldsymbol{\pi}_k$ is $p\left((\boldsymbol{Q})_{k,\cdot}\right)$. A Dirichlet prior for $\boldsymbol{\pi}_k$ results in a Dirichlet full conditional, using the counts of latent ξ variables drawn using (2.6). Updates for any other hyperparameters would require an alternate sampling scheme.

A.3.2 Modified algorithm

Sampling $\{z_t\}$ conditional on $\boldsymbol{Q}$, followed by $\boldsymbol{Q}$ conditional on $\{z_t\}$ results in a chain that tends to get stuck. To improve mixing, we instead integrate $\boldsymbol{Q}$ out of the joint posterior (A.4) and conduct Gibbs sampling between $\{z_t\}$ and $\boldsymbol{\lambda}$. At each iteration, it is then straightforward to draw $\boldsymbol{Q}$ from the conditional distribution given in Appendix A.3.1.

Let $\boldsymbol{N}$ be a $K \times K$ matrix of transition counts for which the (k_1, k_2) entry is the cardinality of $\{t : s_t = k_1 \text{ and } s_{t-z_t} = k_2\}$. Integrating $\boldsymbol{Q}$ from (A.4) yields $p(\{z_t\}, \boldsymbol{\lambda} \mid \{s_t\}) \propto p(\boldsymbol{\lambda}) \prod_t [p(z_t \mid \boldsymbol{\lambda})\, p(s_t \mid z_t, s_{t-z_t})]$, which differs from the original *only* in that

$$\prod_{t=L+1}^{T} \left[p(s_t \mid z_t, s_{t-z_t})\right] = \prod_{k=1}^{K} \psi\left((\boldsymbol{N})_{\cdot,k}, \phi_k\right) , \tag{A.5}$$

where the $\psi(\cdot, \phi)$ takes the form of (A.1) if the columns of $\boldsymbol{Q}$ are independent Dirichlet, (A.2) if the columns of $\boldsymbol{Q}$ are independent SDM, and (A.3) if the columns of $\boldsymbol{Q}$ are independent SBM; and ϕ refers to generic hyperparameters appropriate for the choice of prior.

The modified algorithm then proceeds with the standard update for $\boldsymbol{\lambda}$ given in Appendix A.3.1. Each z_t is then updated individually using its full conditional involving $\boldsymbol{\lambda}$ and (A.5) by modifying $\boldsymbol{N}$ to reflect each possible value of $z_t \in \{1, \ldots, L\}$. Unnecessary computation can be avoided by noting redundancies in the denominators of (A.1) and (A.2), and computing (A.5) only over values of k such that $s_{t-\ell} = k$ for $\ell = 1, \ldots, L$.

Allowing inference for $\{\boldsymbol{\pi}_k\}$ when utilizing SBM priors for $\boldsymbol{Q}$ is no more complicated than in the original algorithm. Simply draw $\boldsymbol{\pi}_k$ when sampling from the conditional for the kth column of $\boldsymbol{Q}$. Note, however, that integrating over $\boldsymbol{Q}$ leaves the update for $\boldsymbol{\lambda}$ conditional on $\{\boldsymbol{\pi}_k\}$. Nevertheless, the result is still a valid Gibbs sampler which in practice produces acceptable mixing.

Appendix B

Implementation Details for MMTD

B.1 Marginal distributions

We report the marginal distribution of observations associated with the SBM priors for probability vectors given originally in Appendix A.1, but modified for its use in Chapter 3. These distributions can be useful for computing Bayes factors in addition to facilitating the MCMC algorithms described in Appendices B.2 and B.3.

Consider a length-N sequence of independent random variables $\{s_t\} \in \{1, \ldots, J\}^N$ with common distribution $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)$. Given $\boldsymbol{\theta}$, the probability of the sequence is $\prod_t \theta_{s_t} = \theta_1^{n_1} \cdots \theta_K^{n_K}$ where the sufficient statistics in $\boldsymbol{n} = (n_1, \ldots, n_K)$ count the occurrences of each category. Suppose $\boldsymbol{\theta}$ follows the SBM distribution with parameters $\{\pi_{1,j}\}$, $\{\pi_{3,j}\}$, η, $\{\gamma_j\}$, and $\{\delta_j\}$. Let

$$g_j(a_j, b_j, \boldsymbol{n}) \equiv \frac{\Gamma(a_j + b_j)}{\Gamma(a_j^* + b_j^*)} \frac{\Gamma(a_j^*)}{\Gamma(a_j)} \frac{\Gamma(b_j^*)}{\Gamma(b_j)},$$

with $a_j^* \equiv a_j + n_j$, and $b_j^* \equiv b_j + \sum_{h=j+1}^{K} n_h$. Then the marginal distribution of $\{s_t\}$ (integrating over $\boldsymbol{\theta}$) has probability mass function

$$p(\{s_t\}) = \prod_{j=1}^{J-1} [\pi_{1,j}\, g_j(1, \eta, \boldsymbol{n}) + \pi_{2,j}\, g_j(\gamma_j, \delta_j, \boldsymbol{n}) + \pi_{3,j}\, g_j(\eta, 1, \boldsymbol{n})]\,. \tag{B.1}$$

B.2 MCMC algorithm details: MTDg

Following the hierarchical MTDg model outlined in (3.6), the joint posterior distribution of all unknown parameters is given up to proportionality:

$$p\Big(\{z_t\}_{t=L+1}^{T}, \boldsymbol{\lambda}, \{\boldsymbol{Q}^{(\ell)}\}_{\ell=0}^{L} \mid \{s_t\}_{t=1}^{T}\Big) \propto p(\boldsymbol{\lambda})\, p(\boldsymbol{Q}^{(0)}) \prod_{\ell=1}^{L} \prod_{k=1}^{K} \Big[p\Big((\boldsymbol{Q}^{(\ell)})_{\cdot,k}\Big)\Big] \times \prod_{t=L+1}^{T} \Big[p(z_t \mid \boldsymbol{\lambda})\, p\Big(s_t \mid z_t, \{\boldsymbol{Q}^{(\ell)}\}_{\ell=0}^{L}, \{s_{t-\ell}\}_{\ell=1}^{L}\Big)\Big]\,, \tag{B.2}$$

where $(\boldsymbol{Q}^{(\ell)})_{\cdot,k}$ denotes column k from $\boldsymbol{Q}^{(\ell)}$.

B.2.1 Full Gibbs sampler

MCMC sampling for the full augmented model (3.6) can be achieved entirely with Gibbs updates. A Gibbs sampler cycles through the parameters, drawing updates from the conditional distributions given below.

- $\Pr(z_t = \ell \mid \cdots) \propto \Pr(z_t = \ell \mid \boldsymbol{\lambda})\, p\Big(s_t \mid z_t, \{\boldsymbol{Q}^{(j)}\}_{j=0}^{L}, \{s_{t-j}\}_{j=1}^{L}\Big)$
  $= \lambda_0\, (\boldsymbol{Q}^{(0)})_{s_t}\, 1_{(\ell=0)} + \lambda_\ell\, (\boldsymbol{Q}^{(\ell)})_{s_t, s_{t-\ell}}\, 1_{(\ell \in \{1,\ldots,L\})}$, independently for each $t = L+1, \ldots, T$.

- $p(\boldsymbol{\lambda} \mid \cdots) \propto p(\boldsymbol{\lambda}) \prod_t p(z_t \mid \boldsymbol{\lambda}) = \mathrm{SBM}(\boldsymbol{\lambda} \mid \boldsymbol{\pi}_1, \boldsymbol{\pi}_3, \eta, \boldsymbol{\gamma}, \boldsymbol{\delta}) \prod_t \lambda_{z_t}$, a conjugate SBM-multinomial update using the counts of z_t in each of $\{0, 1, \ldots, L\}$. A draw from the full conditional distribution begins by drawing the latent

stick-breaking weights X_ℓ for $\ell = 0, \ldots, L-1$, each from a mixture of three beta distributions. The mixture weights for X_ℓ are the three summands in the corresponding product terms of (B.1) (with indexes shifted to begin at $\ell = 0$), where n_ℓ is the cardinality of $\{t : z_t = \ell\}$. The three beta distributions have the corresponding a^*_ℓ and b^*_ℓ shape parameters taken from the SBM prior parameters and counts. The draw for $\boldsymbol{\lambda}$ is then constructed from the sampled $\{X_\ell\}$ using (2.4).

- $p(\boldsymbol{Q}^{(0)} \mid \cdots) \propto p(\boldsymbol{Q}^{(0)}) \prod_{t:z_t=0} p\left(s_t \mid z_t, \{\boldsymbol{Q}^{(\ell)}\}_{\ell=0}^{L}, \{s_{t-\ell}\}_{\ell=1}^{L}\right)$
  $= \mathrm{Dir}(\boldsymbol{Q}^{(0)} \mid \boldsymbol{\alpha}^{(0)}) \prod_{t:z_t=0} (\boldsymbol{Q}^{(0)})_{s_t}$, a standard conjugate Dirichlet-multinomial update using the counts of s_t in each of $\{1, \ldots, K\}$ collected in $\boldsymbol{N}^{(0)}$. The full conditional is then $\mathrm{Dir}(\boldsymbol{\alpha}^{(0)} + \boldsymbol{N}^{(0)})$.

- $p\left((\boldsymbol{Q}^{(\ell)})_{\cdot,k} \mid \cdots\right) \propto p\left((\boldsymbol{Q}^{(\ell)})_{\cdot,k}\right) \prod_{\{t:z_t=\ell \text{ and } s_{t-\ell}=k\}} \times$
  $p\left(s_t \mid z_t, \{\boldsymbol{Q}^{(\ell)}\}_{\ell=0}^{L}, \{s_{t-\ell}\}_{\ell=1}^{L}\right)$
  $= \mathrm{Dir}\left((\boldsymbol{Q}^{(\ell)})_{\cdot,k} \mid \boldsymbol{\alpha}_k^{(\ell)}\right) \prod_{\{t:z_t=\ell \text{ and } s_{t-\ell}=k\}} (\boldsymbol{Q}^{(\ell)})_{s_t,k}$, a standard conjugate Dirichlet-multinomial update using the counts of s_t in each of $\{1, \ldots, K\}$ collected in the kth column of $\boldsymbol{N}^{(\ell)}$. The full conditional is then $\mathrm{Dir}\left(\boldsymbol{\alpha}_k^{(\ell)} + (\boldsymbol{N}^{(\ell)})_{\cdot,k}\right)$, independent for each $\ell = 1, \ldots, L$, and $k = 1, \ldots, K$.

B.2.2 Collapsed Gibbs sampler

Iterated full-conditional sampling of both $\{z_t\}$ and $\{\boldsymbol{Q}^{(r)}\}$ slows exploration of the joint posterior. To improve mixing, we instead integrate all Q parameters out of (B.2) and conduct Gibbs sampling between $\{z_t\}$ and $\boldsymbol{\lambda}$. At each iteration, it is then straightforward to draw each $\boldsymbol{Q}^{(r)}$ from the conditional distributions given in Appendix B.2.1.

Again we will summarize transition count information in $\{(s_t, z_t)\}$ by aggregating into sufficient statistics $\boldsymbol{N}^{(0)}$ and $\{\boldsymbol{N}^{(\ell)}\}_{\ell=1}^{L}$, a set of $K \times K$ matrices

described in Section 3.3.1 and Appendix B.2.1. Integrating $\boldsymbol{Q}^{(0)}$ from (B.2) yields $p(\{z_t\}, \boldsymbol{\lambda} \mid \{s_t\}) \propto p(\boldsymbol{\lambda}) \prod_t \left[p(z_t \mid \boldsymbol{\lambda})\, p\left(s_t \mid z_t, \{s_{t-\ell}\}_{\ell=1}^{L}\right)\right]$ which differs from the original *only* in that

$$\prod_{t=L+1}^{T} \left[p\left(s_t \mid z_t, \{s_{t-\ell}\}_{\ell=1}^{L}\right)\right] = \psi(\boldsymbol{N}^{(0)}, \phi^{(0)}) \prod_{\ell=1}^{L} \prod_{k=1}^{K} \psi\left((\boldsymbol{N}^{(\ell)})_{\cdot,k}, \phi_k^{(\ell)}\right), \tag{B.3}$$

where the $\psi(\cdot, \phi)$ takes the form of (A.1) if the columns of $\boldsymbol{Q}$ have independent Dirichlet priors, and (B.1) if the columns of $\boldsymbol{Q}$ have independent SBM priors; and ϕ refers to generic hyperparameters appropriate for the choice of prior.

The modified algorithm then proceeds with the standard update for $\boldsymbol{\lambda}$ given in Appendix B.2.1. Each z_t is then updated individually with

$$\Pr(z_t = \ell \mid \boldsymbol{\lambda}, \{s_t\}, \{z_{t'}\}_{t' \neq t}) \propto \lambda_\ell\, \psi(\boldsymbol{N}^{(0)}, \phi^{(0)}) \prod_{j=1}^{L} \psi\left((\boldsymbol{N}^{(j)})_{\cdot,s_{t-\ell}}, \phi_{s_{t-\ell}}^{(j)}\right), \tag{B.4}$$

where the $\{\boldsymbol{N}^{(j)}\}$ are modified to reflect the possible values of $z_t \in \{0, 1, \ldots, L\}$.

B.3 MCMC algorithm details: MMTD

Following the hierarchical MMTD model outlined in (3.7), the joint posterior distribution of all unknown parameters is given up to proportionality:

$$\begin{aligned} p\Big(\{\zeta_t\}_{t=L+1}^{T}, \boldsymbol{\Lambda}, \{\boldsymbol{\lambda}^{(r)}\}_{r=1}^{R}, \{\boldsymbol{\mathcal{Q}}^{(r)}\}_{r=0}^{R}, \mid \{s_t\}_{t=1}^{T}\Big) \propto \\ p(\boldsymbol{\Lambda})\, p(\boldsymbol{\mathcal{Q}}^{(0)}) \prod_{r=1}^{R} \left[p(\boldsymbol{\lambda}^{(r)}) \prod_{j=1}^{K^r} p\left((\boldsymbol{\mathcal{Q}}^{(r)})_{\cdot,j}\right)\right] \times \\ \prod_{t=L+1}^{T} \left[p\left(\zeta_t \mid \boldsymbol{\Lambda}, \{\boldsymbol{\lambda}^{(r)}\}_{r=1}^{R}\right)\, p\left(s_t \mid \zeta_t, \{\boldsymbol{\mathcal{Q}}^{(r)}\}_{r=0}^{R}, \{s_{t-\ell}\}_{\ell=1}^{L}\right)\right], \end{aligned} \tag{B.5}$$

where $(\boldsymbol{\mathcal{Q}}^{(r)})_{\cdot,j}$ denotes column j from a matricized version of $\boldsymbol{\mathcal{Q}}^{(r)}$.

B.3.1 Full Gibbs sampler

MCMC sampling for the full hierarchical model (3.7) can be achieved entirely with Gibbs updates. A Gibbs sampler cycles through the parameters, drawing updates from the conditional distributions given below. In what follows, let $Z(\zeta)$ and $\boldsymbol{z}(\zeta)$ map ζ to its corresponding Z and $\boldsymbol{z}$ respectively. Also, let $\varrho_r(\boldsymbol{s})$ be a unique map from each possible length-r vector of lagged states $\boldsymbol{s} \in \{1, \ldots, K\}^r$ to the corresponding column index of the flattened (matricized) $\boldsymbol{\mathcal{Q}}^{(r)}$. Further, let $\boldsymbol{s}_{t-1}(\boldsymbol{z})$ be a function accepting a lag configuration $\boldsymbol{z}$ and returning the values of the states at those selected lags from the vector $(s_{t-1}, s_{t-2}, \ldots, s_{t-L})$. For example, if $\boldsymbol{z}_t = (2, 5)$, then $\boldsymbol{s}_{t-1}(\boldsymbol{z}_t)$ will return the vector (s_{t-2}, s_{t-5}).

- $\Pr(\zeta_t = i \mid \cdots) \propto p\left(\zeta_t \mid \boldsymbol{\Lambda}, \{\boldsymbol{\lambda}^{(r)}\}_{r=1}^R\right) p\left(s_t \mid \zeta_t, \{\boldsymbol{\mathcal{Q}}^{(r)}\}_{r=0}^R, \{s_{t-\ell}\}_{\ell=1}^L\right)$
  $\propto \Lambda_{Z(i)} \, \lambda^{(Z(i))}_{(\boldsymbol{z}(i))} \, (\boldsymbol{\mathcal{Q}}^{(Z(i))})_{s_t, \varrho_{Z(i)}(\boldsymbol{s}_{t-1}(\boldsymbol{z}(i)))}$ independently for each $t = L+1, \ldots, T$, with $i \in \left\{0, 1, \ldots, \left[\binom{L}{1} + \binom{L}{2} + \ldots + \binom{L}{R}\right]\right\}$. Note that we define $\lambda^{(0)} \equiv 1$.

- $p(\boldsymbol{\Lambda} \mid \cdots) \propto p(\boldsymbol{\Lambda}) \prod_t p(\zeta_t \mid \boldsymbol{\Lambda}, \{\boldsymbol{\lambda}^{(r)}\}) \propto \mathrm{SBM}(\boldsymbol{\Lambda} \mid \boldsymbol{\pi}_1, \boldsymbol{\pi}_3, \eta, \boldsymbol{\gamma}, \boldsymbol{\delta}) \prod_t \Lambda_{Z(\zeta_t)}$, a conjugate SBM-multinomial update using the counts of $Z(\zeta_t)$ in each of $\{0, 1, \ldots, R\}$.

- $p(\boldsymbol{\lambda}^{(r)} \mid \cdots) \propto p(\boldsymbol{\lambda}^{(r)}) \prod_t p(\zeta_t \mid \boldsymbol{\Lambda}, \{\boldsymbol{\lambda}^{(r)}\}) \propto \mathrm{SDM}(\boldsymbol{\lambda}^{(r)} \mid \boldsymbol{\alpha}^{(r)}_\lambda, \beta^{(r)}_\lambda) \times$
  $\prod_{t:Z(\zeta_t)=r} \lambda^{(r)}_{(\boldsymbol{z}(\zeta_t))}$ independently for $r = 1, \ldots, R$. Here, $\boldsymbol{\lambda}^{(r)}$ is indexed by the $\binom{L}{r}$ possible sets of lags. This is a conjugate SDM-multinomial update using the counts of the $\binom{L}{r}$ unique lag configurations $\boldsymbol{z}_t$ within order r. The full conditional is a SDM distribution with $\beta^{(r)}_\lambda$ and with the multinomial counts added to $\boldsymbol{\alpha}^{(r)}_\lambda$, analogous to Dirichlet full conditionals.

- $p(\boldsymbol{\mathcal{Q}}^{(0)} \mid \cdots) \propto p(\boldsymbol{\mathcal{Q}}^{(0)}) \prod_{t:Z(\zeta_t)=0} p\left(s_t \mid \zeta_t, \{\boldsymbol{\mathcal{Q}}^{(r)}\}_{r=0}^L, \{s_{t-\ell}\}_{\ell=1}^L\right)$
  $= \mathrm{Dir}(\boldsymbol{\mathcal{Q}}^{(0)} \mid \boldsymbol{\alpha}_{Q^{(0)}}) \prod_{t:Z(\zeta_t)=0} (\boldsymbol{\mathcal{Q}}^{(0)})_{s_t}$, a standard conjugate

Dirichlet-multinomial update using the counts of s_t in each of $\{1, \ldots, K\}$ collected in $\boldsymbol{N}^{(0)}$. The full conditional is then $\mathrm{Dir}(\boldsymbol{\alpha}_{Q^{(0)}} + \boldsymbol{N}^{(0)})$.

- $p\left((\boldsymbol{\mathcal{Q}}^{(r)})_{\cdot,j} \mid \cdots\right) \propto p\left((\boldsymbol{\mathcal{Q}}^{(r)})_{\cdot,j}\right) \prod_{\{t:Z(\zeta_t)=r \text{ and } \varrho_r(\boldsymbol{s}_{t-1}(\boldsymbol{z}(\zeta_t)))=j\}} \times$
  $p\left(s_t \mid \zeta_t, \{\boldsymbol{\mathcal{Q}}^{(r)}\}_{r=0}^{R}, \{s_{t-\ell}\}_{\ell=1}^{L}\right)$
  $\propto \mathrm{Dir}\left((\boldsymbol{\mathcal{Q}}^{(r)})_{\cdot,j} \mid \boldsymbol{\alpha}_{Q^{(r)}}\right) \prod_{\{t:Z(\zeta_t)=r \text{ and } \varrho_r(\boldsymbol{s}_{t-1}(\boldsymbol{z}(\zeta_t)))=j\}} (\boldsymbol{\mathcal{Q}}^{(r)})_{s_t,j}$,
  independently for $r = 1, \ldots, R$, and $j = 1, \ldots, K^r$. Again, this is a standard conjugate Dirichlet-multinomial update using the transition counts collected in $(\boldsymbol{\mathcal{N}}^{(r)})_{\cdot,j}$, where $\boldsymbol{\mathcal{N}}^{(r)}$ is a matrix corresponding to the matricized version of $\boldsymbol{\mathcal{Q}}^{(r)}$. The full conditional distribution is then $\mathrm{Dir}(\boldsymbol{\alpha}_{Q^{(r)}} + (\boldsymbol{\mathcal{N}}^{(r)})_{\cdot,j})$.

B.3.2 Collapsed Gibbs sampler

Iterated full-conditional sampling of both $\{\zeta_t\}$ and $\{\boldsymbol{\mathcal{Q}}^{(r)}\}$ slows exploration of the joint posterior. To improve mixing, we instead integrate each $\boldsymbol{\mathcal{Q}}^{(r)}$ out of the joint posterior (B.5) and conduct Gibbs sampling between $\{\zeta_t\}$, $\boldsymbol{\Lambda}$ and each $\boldsymbol{\lambda}^{(r)}$. At each iteration, it is then straightforward to draw each $\boldsymbol{\mathcal{Q}}^{(r)}$ from the conditional distributions given in Appendix B.3.1.

For each $r = 1, \ldots, R$, again let $\boldsymbol{\mathcal{N}}^{(r)}$ be a matrix containing transition counts for which the (k, j) entry is the cardinality of $\{t : Z(\zeta_t) = r \text{ and } \varrho_r(\boldsymbol{s}_{t-1}(\boldsymbol{z}(\zeta_t))) = j \text{ and } s_t = k\}$. Also let the kth entry of vector $\boldsymbol{N}^{(0)}$ be the cardinality of $\{t : Z(\zeta_t) = 0 \text{ and } s_t = k\}$. Integrating all $\boldsymbol{\mathcal{Q}}^{(r)}$ from the full joint posterior proportional to (B.5) yields

$$p\Big(\{\zeta_t\}, \boldsymbol{\Lambda}, \{\boldsymbol{\lambda}^{(r)}\} \mid \{s_t\}\Big) \propto \mathrm{SBM}(\boldsymbol{\Lambda}) \prod_r \Big[\mathrm{SDM}(\boldsymbol{\lambda}^{(r)})\Big] \prod_t \Big[\Lambda_{Z(\zeta_t)}\, \lambda_{(\boldsymbol{z}(\zeta_t))}^{(Z(\zeta_t))}\Big] \times$$
$$p\left(\boldsymbol{N}^{(0)} \mid \{\zeta_t\}, \{s_t\}\right) \prod_{r=1}^{R} \left[\prod_{j=1}^{K^r} p\left((\boldsymbol{\mathcal{N}}^{(r)})_{\cdot,j} \mid \{\zeta_t\}, \{s_t\}\right)\right], \quad \text{(B.6)}$$

where $p\left((\boldsymbol{\mathcal{N}}^{(r)})_{\cdot,j} \mid \{\zeta_t\}, \{s_t\}\right)$ takes the form of (A.1) if the columns of matricized $\boldsymbol{\mathcal{Q}}^{(r)}$ follow independent Dirichlet priors, and (B.1) if they follow independent SBM priors. The same marginal distribution forms apply for $\boldsymbol{N}^{(0)}$.

The modified algorithm then proceeds with the standard updates for $\boldsymbol{\Lambda}$ and each $\boldsymbol{\lambda}^{(r)}$ given in Appendix B.3.1. Each ζ_t is then updated individually using its discrete collapsed conditional

$$p\left(\zeta_t \mid \cdots, -\{\boldsymbol{\mathcal{Q}}^{(r)}\}\right) \propto \Lambda_{Z(\zeta_t)} \lambda_{(\boldsymbol{z}(\zeta_t))}^{(Z(\zeta_t))} p\left(\boldsymbol{N}^{(0)} \mid \{\zeta_t\}, \{s_t\}\right) \times \prod_{r=1}^{R} \left[\prod_{j=1}^{K^r} p\left((\boldsymbol{\mathcal{N}}^{(r)})_{\cdot,j} \mid \{\zeta_t\}, \{s_t\}\right)\right], \tag{B.7}$$

where we modify $\{\boldsymbol{\mathcal{N}}^{(r)}\}$ to reflect each possible value of $\zeta_t \in \left\{0, 1, \ldots, \left[\binom{L}{1} + \binom{L}{2} + \ldots + \binom{L}{R}\right]\right\}$.

Appendix C

Implementation Details for GPMTD

C.1 Setup for mixture component updates

Conditional on the configuration variables $\{z_t\}$ and covariance hyperparameters ν_κ, κ_0, ν_ψ, ψ_0, we have L independent block-conditional updates for parameters in the (non-intercept) mixture components. To simplify notation, assume without loss of generality that we are working with component ℓ, so that we can drop the ℓ index on each parameter. Let n_ℓ count the cardinality of $\{t : z_t = \ell\}$ and partition $\boldsymbol{f}$ into $\boldsymbol{f}^i$ and $\boldsymbol{f}^o$, indexed by $z_t = \ell$ and $z_t \neq \ell$, respectively. The joint full conditional density for this component is

$$\begin{aligned} p(\mu, \sigma^2, \boldsymbol{f}, \kappa, \psi \mid \{z_t\}, \nu_\kappa, \kappa_0, \nu_\psi, \psi_0, \{y_t\}) \propto & \prod_{t:z_t=\ell} \left[\mathrm{N}(y_t \mid \mu + f_{t,\ell}, \sigma^2)\right] \times \\ & \mathrm{N}(\mu \mid m_0, v_0)\ \mathrm{IG}(\sigma^2 \mid \nu_\sigma/2, \nu_\sigma s_0/2)\ \mathrm{N}\left(\boldsymbol{f} \mid \boldsymbol{0}, \kappa\sigma^2 \boldsymbol{R}(\psi)\right) \times \\ & \mathrm{IG}(\kappa \mid \nu_\kappa/2, \nu_\kappa\kappa_0/2)\ \mathrm{IG}(\psi \mid \nu_\psi/2, \nu_\psi\psi_0/2)\,, \end{aligned} \tag{C.1}$$

where the normal density for $\boldsymbol{f}$ has dimension $T - L$ and correlation matrix $\boldsymbol{R}(\psi)$ which can also be partitioned into active and inactive parts $\boldsymbol{R}^{ii}$, $\boldsymbol{R}^{oo}$, $\boldsymbol{R}^{io}$, and $\boldsymbol{R}^{oi} = (\boldsymbol{R}^{io})'$. We begin by marginalizing $\boldsymbol{f}$ out of (C.1), resulting in a n_ℓ-variate Gaussian density for the vector $\boldsymbol{y}^i$ containing $\{y_t : z_t = \ell\}$ given by $\mathrm{N}\left(\boldsymbol{y}^i \mid \mathbf{1}\mu, \sigma^2\boldsymbol{W}\right)$, where $\boldsymbol{W} = (\kappa\boldsymbol{R}(\psi)^{ii} + \boldsymbol{I})$ and $\boldsymbol{I}$ is the conforming identity matrix. Keep in mind that $\boldsymbol{W}$ is dependent on ψ. Now let $\hat{\mu} = (\mathbf{1}'\boldsymbol{W}^{-1}\mathbf{1})^{-1}\mathbf{1}'\boldsymbol{W}^{-1}\boldsymbol{y}^i = \sum_{j=1}^{n_\ell}(\boldsymbol{W}^{-1}\boldsymbol{y}^i)_j/w$ where $w = \mathbf{1}'\boldsymbol{W}^{-1}\mathbf{1}$, and $s = (\boldsymbol{y}^i - \mathbf{1}\hat{\mu})'\boldsymbol{W}^{-1}(\boldsymbol{y}^i - \mathbf{1}\hat{\mu})$. The joint density for $\boldsymbol{y}^i$ can then be factored as

$$p(\boldsymbol{y}^i \mid \cdots, -\boldsymbol{f}) \propto \det(\boldsymbol{W})^{-1/2}\ (\sigma^2)^{-n_\ell/2} \exp\left[-\frac{w(\hat{\mu}-\mu)^2 + s}{2\sigma^2}\right]. \tag{C.2}$$

Now using the prior for μ, we can further marginalize to obtain

$$\begin{aligned} p(\boldsymbol{y}^i \mid \cdots, -\boldsymbol{f}, -\mu) &\propto \int p(\boldsymbol{y}^i \mid \cdots, -\boldsymbol{f})\ \mathrm{N}(\mu \mid m_0, v_0)\,\mathrm{d}\mu \\ &\propto \det(\boldsymbol{W})^{-1/2}\ (\sigma^2)^{-n_\ell/2} \exp\left[-\frac{s}{2\sigma^2}\right]\left(\frac{\sigma^2}{w}\right)^{1/2} c\,, \end{aligned} \tag{C.3}$$

where

$$\begin{aligned} c_0 &= \int \mathrm{N}(\hat{\mu} \mid \mu, \sigma^2/w)\ \mathrm{N}(\mu \mid m_0, v_0)\,\mathrm{d}\mu \\ &\propto (\sigma^2/w + v_0)^{-1/2}\ \exp\left[-\frac{(\hat{\mu}-m_0)^2}{2(\sigma^2/w+v_0)}\right]\ \int \mathrm{N}(\mu \mid m_1, v_1)\,\mathrm{d}\mu \\ &= (\sigma^2/w + v_0)^{-1/2}\ \exp\left[-\frac{(\hat{\mu}-m_0)^2}{2(\sigma^2/w+v_0)}\right] = c\,, \end{aligned} \tag{C.4}$$

with $v_1 = \left(v_0^{-1} + w/\sigma^2\right)^{-1}$ and $m_1 = v_1(m_0/v_0 + w\hat{\mu}/\sigma^2)$.

A full Gibbs scan for $(\mu, \sigma^2, \boldsymbol{f}, \kappa, \psi)_\ell$ then proceeds as follows:

1. Perform a random-walk Metropolis update of (κ, ψ) with their joint collapsed conditional density proportional to $p(\boldsymbol{y}^i \mid \cdots, -\boldsymbol{f}, -\mu)\ p(\kappa \mid \nu_\kappa, \kappa_0)\ p(\psi \mid$

$\nu_\psi, \psi_0)$ where the first density is given in (C.3) and the remaining two are the inverse-gamma densities in (C.1). Gaussian proposals are drawn on the logarithmic scale, requiring a Jacobian adjustment by multiplying the collapsed conditional density by $\kappa\psi$ when computing the acceptance probability.

2. Draw μ from its collapsed conditional distribution $\mathrm{N}(m_1, v_1)$.

3. Draw σ^2 from its collapsed conditional with density proportional to $p(\boldsymbol{y}^i \mid \cdots, -\boldsymbol{f})\, p(\sigma^2 \mid \nu_\sigma, s_0)$, where the first density is given in (C.2) and the second is the inverse-gamma density in (C.1). The result is another inverse-gamma density with shape $(\nu_\sigma + n_\ell)/2$ and scale $(\nu_\sigma s_0 + w(\hat{\mu} - \mu)^2 + s)/2$.

4. Introduce $\boldsymbol{f}^i$ with $\boldsymbol{f}^o$ still marginalized and draw from $p(\boldsymbol{f}^i \mid \cdots, -\boldsymbol{f}^o) \propto \mathrm{N}(\boldsymbol{y}^i - \mathbf{1}\mu \mid \boldsymbol{f}^i, \sigma^2\boldsymbol{I})\ \mathrm{N}(\boldsymbol{f}^i \mid \mathbf{0}, \kappa\sigma^2\boldsymbol{R}^{ii})$, a standard conditionally conjugate multivariate Gaussian update with covariance matrix $\boldsymbol{\Sigma} = \sigma^2 \left(\kappa^{-1}(\boldsymbol{R}^{ii})^{-1} + \boldsymbol{I}\right)^{-1}$ and mean vector $\boldsymbol{\Sigma}\left(\boldsymbol{y}^i - \mathbf{1}\mu\right)/\sigma^2$. Following Rasmussen and Williams (2006, p. 46), the positive definite matrix $\boldsymbol{\Sigma}$ is computed in a numerically stable way with the matrix inversion lemma as $\sigma^2\boldsymbol{K}(\boldsymbol{I} - \tilde{\boldsymbol{K}})$ where $\boldsymbol{K} = \kappa\boldsymbol{R}^{ii}$ and $\tilde{\boldsymbol{K}}$ is the solution to $(\boldsymbol{K} + \boldsymbol{I})\tilde{\boldsymbol{K}} = \boldsymbol{K}$.

5. Finally, draw $\boldsymbol{f}^o$ from its full conditional distribution. Let $\boldsymbol{C} = \kappa\sigma^2\boldsymbol{R}$. Then we have $p(\boldsymbol{f}^o \mid \cdots) = \mathrm{N}\left[\boldsymbol{f}^o \mid \boldsymbol{C}^{oi}(\boldsymbol{C}^{ii})^{-1}\boldsymbol{f}^i, \boldsymbol{C}^{oo} - \boldsymbol{C}^{oi}(\boldsymbol{C}^{ii})^{-1}\boldsymbol{C}^{io}\right]$.

C.2 Gibbs sampler for GPMTD

The full Gibbs sampler for the GPMTD model proceeds as follows:

1. Draw z_t from the discrete full conditional distribution given in (4.4) independently for $t = L + 1, \ldots, T$.

2. Calculate the current mixture allocation counts $\boldsymbol{n} = (n_0, n_1, \ldots, n_L)$ where $n_\ell = \sum_t 1_{(z_t=\ell)}$ and draw $\boldsymbol{\lambda}$ from the SBM-multinomial full conditional distribution outlined in Section 2.2.2. A Dirichlet or sparse Dirichlet mixture prior for $\boldsymbol{\lambda}$ could also be trivially accommodated in this model, with this full conditional update corresponding to the conjugate model for multinomial data.

3. Draw μ_0 from the full conditional distribution $\mathrm{N}\left(m_1^{(0)}, v_1^{(0)}\right)$ where $v_1^{(0)} = \left((v_0^{(0)})^{-1} + n_0/\sigma_0^2\right)^{-1}$ and $m_1^{(0)} = v_1^{(0)}\left(m_0^{(0)}/v_0^{(0)} + \sum_{t:z_t=0} y_t/\sigma_0^2\right)$.

4. Draw σ_0^2 from the full conditional inverse-gamma distribution with shape $\left(\nu_\sigma^{(0)} + n_0\right)/2$ and scale $\left(\nu_\sigma^{(0)} s_0^{(0)} + \sum_{t:z_t=0}(y_t - \mu_0)^2\right)/2$.

5. Perform the scan for $(\mu, \sigma^2, \boldsymbol{f}, \kappa, \psi)_\ell$ described in Appendix C.1, independently for $\ell = 1, \ldots, L$.

6. Draw ν_κ and ν_ψ from their discrete full conditional distributions
$p(\nu_\kappa \mid \ldots) \propto \prod_{\{\ell>0:n_\ell>0\}} \left[\mathrm{IG}(\kappa_\ell \mid \nu_\kappa/2, \nu_\kappa\,\kappa_0/2)\right] 1_{(\nu_\kappa \in \mathcal{V}_\kappa)}$ and
$p(\nu_\psi \mid \ldots) \propto \prod_{\{\ell>0:n_\ell>0\}} \left[\mathrm{IG}(\psi_\ell \mid \nu_\psi/2, \nu_\psi\,\psi_0/2)\right] 1_{(\nu_\psi \in \mathcal{V}_\psi)}$.

7. Draw κ_0 and ψ_0 from their full conditional gamma distributions. In the former case, if we let $n^* = \sum_{\ell=1}^{L} 1_{(n_\ell>0)}$ and $\tilde{\kappa} = \sum_{\{\ell>0:n_\ell>0\}} \kappa_\ell^{-1}$, we have $p(\kappa_0 \mid \cdots) \propto \kappa_0^{a_\kappa + \nu_\kappa n^*/2 - 1} \exp\left[-(b_\kappa + \nu_\kappa\tilde{\kappa}/2)\kappa_0\right] \propto \mathrm{Ga}(\kappa_0 \mid a_\kappa + \nu_\kappa n^*/2, b_\kappa + \nu_\kappa\tilde{\kappa}/2)$. The full conditional distribution for ψ_0 is analogous.

Appendix D

Slice Sampler for Stick-Breaking Weights in the Nonparametric Model

We present the hyperrectangle slice sampler from Neal (2003), applied to the full conditional distribution for the stick-breaking weights in Section 5.2.3. 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 full conditional density (5.11) 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 $z \sim \text{Unif}(0, g(\boldsymbol{v}_0))$.

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 $U_h \overset{\text{ind.}}{\sim} \text{Unif}(0,1)$, 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 $\tilde{U}_h \overset{\text{ind.}}{\sim} \text{Unif}(0,1)$, 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 $h = 1, \ldots, H-1$.

Bibliography

Agresti, A. and Hitchcock, D. B. (2005), "Bayesian inference for categorical data analysis," *Statistical Methods & Applications*, 14, 297–330.

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`.

Albert, J. H. and Gupta, A. K. (1982), "Mixtures of Dirichlet Distributions and Estimation in Contingency Tables," *The Annals of Statistics*, 10, 1261–1268.

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.

Armagan, A., Dunson, D. B., and Lee, J. (2013), "Generalized double Pareto shrinkage," *Statistica Sinica*, 23, 119.

Atchison, J. and Shen, S. M. (1980), "Logistic-normal distributions: Some properties and uses," *Biometrika*, 67, 261–272.

Azat, J. (2016), "GrandTab.2016.04.11," California Central Valley Chinook Population Database Report. California Department of Fish and Wildlife, URL `http://www.calfish.org/ProgramsData/Species/CDFWAnadromousResourceAssessment.aspx`.

Azzalini, A. (1985), "A Class of Distributions Which Includes the Normal Ones," *Scandinavian Journal of Statistics*, 12, 171–178.

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.

Bacallado, S. (2011), "Bayesian analysis of variable-order, reversible Markov chains," *The Annals of Statistics*, 39, 838–864.

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.

Bartlett, M. S. (1951), "The frequency goodness of fit test for probability chains," *Mathematical Proceedings of the Cambridge Philosophical Society*, 47, 86–95.

Basson, M. and Fogarty, M. J. (1997), "Harvesting in discrete-time predator-prey systems," *Mathematical Biosciences*, 141, 41–74.

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.

Berchtold, A. and Raftery, A. E. (2002), "The Mixture Transition Distribution Model for High-Order Markov Chains and Non-Gaussian Time Series," *Statistical Science*, 17, 328–356.

Besag, J. and Mondal, D. (2013), "Exact Goodness-of-Fit Tests for Markov Chains," *Biometrics*, 69, 488–496.

Bezanson, J., Edelman, A., Karpinski, S., and Shah, V. B. (2017), "Julia: A Fresh Approach to Numerical Computing," *SIAM Review*, 59, 65–98.

Bouguila, N. and Ziou, D. (2004), "Dirichlet-based probability model applied to human skin detection [image skin detection]," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 5, IEEE.

Box, G. and Jenkins, G. (1976), *Time Series Analysis: Forecasting and Control*, Holden-Day series in time series analysis and digital processing, Holden-Day.

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.

Bühlmann, P., Wyner, A. J., et al. (1999), "Variable length Markov chains," *The Annals of Statistics*, 27, 480–513.

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

Cancho, V. G., Dey, D. K., Lachos, V. H., and Andrade, M. G. (2011), "Bayesian nonlinear regression models with scale mixtures of skew-normal distributions: Estimation and case influence diagnostics," *Computational Statistics & Data Analysis*, 55, 588–602.

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.

Chen, R. and Tsay, R. S. (1993), "Nonlinear Additive ARX Models," *Journal of the American Statistical Association*, 88, 955–967.

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

Connor, R. J. and Mosimann, J. E. (1969), "Concepts of Independence for Proportions with a Generalization of the Dirichlet Distribution," *Journal of the American Statistical Association*, 64, 194–206.

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

De Iorio, M., Müller, P., Rosner, G. L., and MacEachern, S. N. (2004), "An ANOVA Model for Dependent Random Measures," *Journal of the American Statistical Association*, 99, 205–215.

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.

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.

Dietz, L. (2010), "Directed Factor Graph Notation for Generative Models," Technical report, Max Planck Institute for Informatics.

Dunson, D. B. and Park, J.-H. (2008), "Kernel stick-breaking processes," *Biometrika*, 95, 307–323.

Duvenaud, D. K., Nickisch, H., and Rasmussen, C. E. (2011), "Additive Gaussian Processes," in Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q. (editors), *Advances in Neural Information Processing Systems 24*, Curran Associates, Inc.

Ecker, M. D. and Gelfand, A. E. (1999), "Bayesian Modeling and Inference for Geometrically Anisotropic Spatial Data," *Mathematical Geology*, 31, 67–83.

Elfadaly, F. G. and Garthwaite, P. H. (2017), "Eliciting Dirichlet and Gaussian copula prior distributions for multinomial models," *Statistics and Computing*, 27, 449–467.

Fahrmeir, L. and Kaufmann, H. (1987), "Regression Models for Non-stationary Categorical Time Series," *Journal of Time Series Analysis*, 8, 147–160.

Fan, T.-H. and Tsai, C.-A. (1999), "A Bayesian method in determining the order of a finite state Markov chain," *Communications in Statistics-Theory and Methods*, 28, 1711–1730.

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.

George, E. I. and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889.

— (1997), "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7, 339–373.

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.

Good, I. J. (1976), "On the Application of Symmetric Dirichlet Distributions and their Mixtures to Contingency Tables," *The Annals of Statistics*, 4, 1159–1189.

Green, P. J. (1995), "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, 82, 711–732.

Gregorčič, G. and Lightbody, G. (2009), "Gaussian process approach for modelling of nonlinear systems," *Engineering Applications of Artificial Intelligence*, 22, 522–533.

Gutjahr, T., Ulmer, H., and Ament, C. (2012), "Sparse Gaussian Processes with Uncertain Inputs for Multi-Step Ahead Prediction," *IFAC Proceedings Volumes*, 45, 107–112.

Hamilton, N. (2017), *ggtern: An Extension to 'ggplot2', for the Creation of Ternary Diagrams*, URL `https://CRAN.R-project.org/package=ggtern`. R package version 2.2.1.

Hansen, B. E. (1994), "Autoregressive Conditional Density Estimation," *International Economic Review*, 35, 705–730.

Hare, S. R. and Mantua, N. J. (2001), "An historical narrative on the Pacific Decadal Oscillation, interdecadal climate variability and ecosystem impacts," Report of a talk presented at the 20th NE Pacific Pink and Chum workshop, Seattle, WA.

Hastie, T. and Tibshirani, R. (1990), *Generalized Additive Models*, London; New York: Chapman and Hall, 1st edition.

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.

Henze, N. (1986), "A Probabilistic Representation of the 'Skew-normal' Distribution," *Scandinavian Journal of Statistics*, 13, 271–275.

Hirata, Y., Judd, K., and Kilminster, D. (2004), "Estimating a generating partition from observed time series: Symbolic shadowing," *Physical Review E*, 70, 016215.

Hjort, N. L. (1996), "Bayesian approaches to non- and semiparametric density estimation," in *Bayesian Statistics*, volume 5, Oxford University Press.

Huang, J. Z. and Yang, L. (2004), "Identification of non-linear additive autoregressive models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66, 463–477.

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.

Insua, D., Ruggeri, F., and Wiper, M. (2012), *Bayesian Analysis of Stochastic Process Models*, Hoboken, New Jersey: Wiley.

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

Jääskinen, V., Xiong, J., Corander, J., and Koski, T. (2014), "Sparse Markov Chains for Sequence Data," *Scandinavian Journal of Statistics*, 41, 639–655.

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.

Kalli, M., Griffin, J. E., and Walker, S. G. (2011), "Slice sampling mixture models," *Statistics and Computing*, 21, 93–105.

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.

Kantz, H., Holstein, D., Ragwitz, M., and Vitanov, N. K. (2004), "Markov chain model for turbulent wind speed data," *Physica A: Statistical Mechanics and its Applications*, 342, 315–321.

Kantz, H. and Schreiber, T. (2004), *Nonlinear Time Series Analysis*, Cambridge University Press, 2nd edition.

Katz, R. W. (1981), "On Some Criteria for Estimating the Order of a Markov Chain," *Technometrics*, 23, 243–249.

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.

Kocijan, J., Murray-Smith, R., Rasmussen, C. E., and Likar, B. (2003), "Predictive control with Gaussian process models," in *The IEEE Region 8 EUROCON 2003. Computer as a Tool.*, volume 1, IEEE.

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.

Le, N. D., Martin, R. D., and Raftery, A. E. (1996), "Modeling Flat Stretches, Bursts Outliers in Time Series Using Mixture Transition Distribution Models," *Journal of the American Statistical Association*, 91, 1504–1515.

Lèbre, S. and Bourguignon, P.-Y. (2008), "An EM algorithm for estimation in the mixture transition distribution model," *Journal of Statistical Computation and Simulation*, 78, 713–729.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008), "Mixtures of g Priors for Bayesian Variable Selection," *Journal of the American Statistical Association*, 103, 410–423.

Liang, K.-Y. and Zeger, S. L. (1986), "Longitudinal data analysis using generalized linear models," *Biometrika*, 73, 13–22.

Lin, T. I., Lee, J. C., and Yen, S. Y. (2007), "Finite Mixture Modelling Using the Skew Normal Distribution," *Statistica Sinica*, 17, 909–927.

Lochner, R. H. (1975), "A Generalized Dirichlet Distribution in Bayesian Life Testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, 37, 103–113.

MacDonald, I. L. and Zucchini, W. (1997), *Hidden Markov and Other Models for Discrete-valued Time Series*, CRC Press.

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

Maechler, M. (2015), *VLMC: Variable Length Markov Chains ('VLMC') Models*, URL `https://CRAN.R-project.org/package=VLMC`. R package version 1.4-1.

Martin, R. D. and Raftery, A. E. (1987), "Comment: Robustness, Computation, and Non-Euclidean Models," *Journal of the American Statistical Association*, 82, 1044–1050.

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., Quintana, F. A., Jara, A., and Hanson, T. (2015), *Bayesian Nonparametric Data Analysis*, Springer.

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. (1996), *Bayesian Learning for Neural Networks*, Springer-Verlag New York.

— (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.

Park, T. and Casella, G. (2008), "The Bayesian Lasso," *Journal of the American Statistical Association*, 103, 681–686.

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.

Plotly Technologies Inc. (2015), "Collaborative data science," URL `https://plot.ly`.

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

Quinn, T. J. and Deriso, R. B. (1999), *Quantitative Fish Dynamics*, Oxford University Press.

R Core Team (2016), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, URL `https://www.R-project.org/`.

Raftery, A. and Tavaré, S. (1994), "Estimation and Modelling Repeated Patterns in High Order Markov Chains with the Mixture Transition Distribution Model," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 43, 179–199.

Raftery, A. E. (1985), "A Model for High-Order Markov Chains," *Journal of the Royal Statistical Society: Series B (Methodological)*, 47, 528–539.

Rasmussen, C. E. and Williams, C. K. (2006), *Gaussian Processes for Machine Learning*, MIT Press Cambridge, MA.

Raye, J. (2005), "Using nonlinear dynamics to predict Old Faithful," *Mathematical and Computer Modelling*, 41, 679–687.

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 Dunson, D. B. (2011), "Nonparametric Bayesian models through probit stick-breaking processes," *Bayesian Analysis*, 6, 145–177.

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

Ron, D., Singer, Y., and Tishby, N. (1994), "Learning Probabilistic Automata with Variable Memory Length," in *Proceedings of the Seventh Annual Conference on Computational Learning Theory*, Association for Computing Machinery.

Sarkar, A. and Dunson, D. B. (2016), "Bayesian Nonparametric Modeling of Higher Order Markov Chains," *Journal of the American Statistical Association*, 111, 1791–1803.

Satterthwaite, W. H., Carlson, S. M., and Criss, A. (2017), "Ocean Size and Corresponding Life History Diversity among the Four Run Timings of California Central Valley Chinook Salmon," *Transactions of the American Fisheries Society*, 146, 594–610.

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.

Shi, J. Q., Murray-Smith, R., and Titterington, D. M. (2003), "Bayesian regression and classification using mixtures of Gaussian processes," *International Journal of Adaptive Control and Signal Processing*, 17, 149–161.

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.

Sobol, I. M. (2001), "Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates," *Mathematics and Computers in Simulation*, 55, 271–280.

Sugihara, G., Grenfell, B., May, R. M., Chesson, P., Platt, H., and Williamson, M. (1990), "Distinguishing error from chaos in ecological time series [and discussion]," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 330, 235–251.

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

Takens, F. (1981), "Detecting strange attractors in turbulence," in *Dynamical Systems and Turbulence, Warwick 1980*, volume 898 of *Lecture Notes in Mathematics*, Springer Berlin Heidelberg.

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.

Tank, A., Fox, E. B., and Shojaie, A. (2017), "Granger Causality Networks for Categorical Time Series," *arXiv preprint arXiv:1706.02781*.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, 58, 267–288.

Tong, H. (1990), *Non-linear Time Series: A Dynamical System Approach*, Oxford: Clarendon Press.

Tucker, L. R. (1966), "Some mathematical notes on three-mode factor analysis," *Psychometrika*, 31, 279–311.

Wade, S., Dunson, D. B., Petrone, S., and Trippa, L. (2014a), "Improving Prediction from Dirichlet Process Mixtures via Enrichment," *The Journal of Machine Learning Research*, 15, 1041–1071.

Wade, S., Walker, S. G., and Petrone, S. (2014b), "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.

West, M. and Harrison, J. (1997), *Bayesian Forecasting and Dynamic Models*, Springer Series in Statistics, Springer-Verlag New York, 2nd edition.

Wong, C.-m. and Kohn, R. (1996), "A Bayesian Approach to Estimating and Forecasting Additive Nonparametric Autoregressive Models," *Journal of Time Series Analysis*, 17, 203–220.

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.

Wong, T.-T. (1998), "Generalized Dirichlet distribution in Bayesian analysis," *Applied Mathematics and Computation*, 97, 165–181.

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

Wu, Y., Ghosal, S., et al. (2008), "Kullback Leibler property of kernel mixture priors in Bayesian density estimation," *Electronic Journal of Statistics*, 2, 298–331.

Yang, Y. and Dunson, D. B. (2016), "Bayesian Conditional Tensor Factorizations for High-Dimensional Classification," *Journal of the American Statistical Association*, 111, 656–669.

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.

Yee, T. W. et al. (2010), "The VGAM Package for Categorical Data Analysis," *Journal of Statistical Software*, 32, 1–34.

Zeger, S. L. and Liang, K.-Y. (1986), "Longitudinal Data Analysis for Discrete and Continuous Outcomes," *Biometrics*, 42, 121–130.

Zellner, A. (1986), "On Assessing Prior Distributions and Bayesian Regression Analysis With g-Prior Distributions," in Goel, P. K. and Zellner, A. (editors), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, Elsevier North-Holland, 233–243.

Zhang, H. (2004), "Inconsistent Estimation and Asymptotically Equal Interpolations in Model-Based Geostatistics," *Journal of the American Statistical Association*, 99, 250–261.

Zucchini, W. and MacDonald, I. L. (2009), *Hidden Markov Models for Time Series: An Introduction Using R*, CRC Press, 2nd edition.