A Bayesian model for estimation and order selection in high order Markov chains

Matthew Heiner[†], Athanasios Kottas[†], and Stephan Munch[‡]

[†]Department of Applied Mathematics and Statistics, University of California, Santa Cruz, California, USA
[‡]Fisheries Ecology Division, Southwest Fisheries Science Center, National Marine Fisheries Service, NOAA, Santa Cruz, California, USA

April 7, 2018

Abstract

We develop a model for Bayesian selection in high order Markov chains through an extension of the mixture transition distribution of Raftery (1985). We demonstrate two uses for the model: parsimonious approximation of high order dynamics by mixing lower order transition models, and model selection through over-specification and shrinkage via priors for sparse probability vectors. We discuss properties of the model and demonstrate its utility with simulation studies. We further apply the model to a data analysis from the high-order Markov chain literature and a novel application to pink salmon abundance time series.

Keywords: high order Markov chain, mixture transition distribution, nonlinear dynamics, sparsity prior

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, crucial for replication and 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. While this avoids exponential growth in the parameter space, first order dynamics, or even selecting the incorrect lag, can miss important features in the data. As an illustration, consider a time series of alternating states $(1, 1, 2, 2, 1, 1, 2, 2, \ldots)$. A model conditioning on the first lag only would estimate all transition probabilities to be 0.5 for state 1 and 0.5 for state 2, but conditioning on the second lag reveals a deterministic system with perfect predictability. In this article, we propose a model to address our objectives of Bayesian estimation for the relevant time-delay coordinates: the Markovian order and important lags, as well as 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 R requires estimation of K^R such distributions, severely limiting consideration of all but low order models for moderate length time series. Typically, order (or lag) is selected by maximizing a (possibly penalized) likelihood (as in Raftery (1985), Prado and West (2010)), performing trans-dimensional MCMC (Green, 1995; Insua et al., 2012), using Bayes factors (Zucchini and MacDonald, 2009), predictive criteria, or goodness-of-fit tests (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.

The first general-purpose, parsimonious model for high order Markov chains was the mixture transition distribution (MTD) introduced by Raftery (1985). The MTD model was extended in Raftery and Tavaré (1994) and developed over the subsequent decade. A comprehensive review is given by Berchtold and Raftery (2002). In the 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. Although it has a simple form, the MTD is sufficiently flexible to capture features such as "outliers, bursts, and flat stretches in time series" (Le et al., 1990).

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 $K^R(K-1)$ parameters. These models can also account for exogenous sources of non-stationarity through covariates. Bayesian estimation and interpretability become problematic in these models when many lags are considered.

More recently, tree-based methods have allowed for parsimonious high order Markov chains. Variable-length Markov chains (VLMC, Ron et al. (1994); Bühlmann et al. (1999)) reduce the parameter space by clustering the K^R transition distributions via recursive pruning. Sparse Markov chains (SMC, Jääskinen et al. (2014)) partition the R-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.

Most recently, Sarkar and Dunson (2016) proposed a Bayesian nonparametric model for parsimonious representation of high order chains. They model the K^R transition distributions through tensor factorization and encourage sparsity in the high dimensional core mixing distribution by clustering its components with a Dirichlet process prior (Ferguson, 1973). By allowing variable dimensions along the core mixing distribution, the model further admits inferences for lag importance. This model enjoys a fully Bayesian, albeit complicated, implementation and performs well against the methods described above in forecasting when there are up to four states and ten lags.

Each of the reviewed models, with exception of the multinomial regression approach, can be represented as a mixture of core transition probability distributions and weights with varying complexity (Sarkar and Dunson, 2016). Our approach is to build on the simplicity and interpretability of the MTD model. We propose an extension which allows for higher-order interaction between lags and infers the Markovian order. The remainder of this paper is organized as follows. In Section 2, we develop the proposed extension of the MTD model together with an approach for Bayesian inference using different structured priors to aid the model's intended use. In Section 3, we test the model using two simulation scenarios which reflect our two objectives, demonstrating improved predictive performance over the original MTD. We then use the model to analyze a data set which appears in the preceding literature, as well as analyze an annual time series of pink salmon abundance in Alaska, U.S.A. in Section 4. Finally, we conclude with a summary in Section 5.

2 Model

In a full R-order, time-homogeneous Markov chain, the collection of all possible transition probabilities $\Pr(s_t = i_0 \mid s_{t-1} = i_1, \ldots, s_{t-R} = i_R)$ for $i_\ell \in \{1, \ldots, K\}$, $\ell \in \{1, \ldots, R\}$, $t \in \{R+1, \ldots, T\}$ can be arranged in a $(R+1)$-order tensor $(\boldsymbol{\Omega})_{i_R, i_{R-1}, \ldots, i_2, i_1, i_0}$. If we condition on the first R observations of the time series, the joint probability distribution for the remaining observations is given by $\Pr(\{s_t\}_{t=R+1}^{T} \mid \{s_t\}_{t=1}^{R}, \boldsymbol{\Omega}) = \prod_{t=R+1}^{T} (\boldsymbol{\Omega})_{s_{t-R}, \ldots, s_{t-1}, s_t}$, defining the conditional likelihood that we employ hereafter. We begin by specifying the original MTD model in Section 2.1 and motivating its extension, our proposed model, in Section 2.2. We discuss Bayesian estimation in Section 2.3 and introduce additional options for prior specification in Section 2.4.

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 row-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 R are given as

$$(\boldsymbol{\Omega})_{i_R, i_{R-1}, \ldots, i_2, i_1, i_0} \equiv \sum_{\ell=1}^{R} \lambda_\ell q_{i_\ell, i_0} \,, \tag{1}$$

where $q_{i,j} \equiv (\boldsymbol{Q})_{i,j}$, $0 \leq \lambda_\ell \leq 1$ and $\sum_{\ell=1}^{R} \lambda_\ell = 1$. Although this model is simple and 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.

Form (1) 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 near 0. Hence, inferences for $\boldsymbol{\lambda}$ potentially yield information about important lags for the Markov process. It is easy to show in the $K = 2$ and $R = 2$ case that if the rows of $\boldsymbol{Q}$ are unique, then the current state is conditionally independent of Lag ℓ if and only if $\lambda_\ell = 0$. 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). Heiner et al. (2018) use a single model, relying on inferences on $\boldsymbol{\lambda}$ for insight into lag importance. We adopt that approach here as well.

2.2 Multi-MTD model

The review in Berchtold and Raftery (2002) discusses a number of ways to extend the MTD model. One avenue for increasing modeling flexibility that has received some attention is to use multiple $\boldsymbol{Q}$ matrices, one for each lag. They also suggest, but do not pursue the possibility of mixing over higher order tensors. We propose and build a Bayesian framework for such an extension and refer to it here as the multi-mixture transition distribution (MMTD).

To define the MMTD model, let $J < R$ be a positive integer representing the highest order transition tensor over which we will mix. Then we have $\boldsymbol{Q}^{(1)}$, a $K \times K$ transition matrix, $\boldsymbol{Q}^{(2)}$, a $K \times K \times K$ transition tensor, and so forth up to $\boldsymbol{Q}^{(J)}$, a K^{J+1} transition tensor such that $\sum_{k=1}^{K}(\boldsymbol{Q}^{(J)})_{i_J,i_{J-1},\ldots,i_1,k} = 1$ for all $(i_J, i_{J-1}, \ldots, i_1) \in \{1, \ldots, K\}^J$. Next, introduce a mixing probability vector across orders $\boldsymbol{\Lambda} = (\Lambda_1, \ldots, \Lambda_J)$. The MMTD model for transition probabilities is then given by

$$
\begin{aligned}
&\Pr(s_t = i_0 \mid s_{t-1} = i_1, s_{t-2} = i_2, \ldots, s_{t-R} = i_R, \boldsymbol{\Lambda}, \{\boldsymbol{\lambda}^{(j)}\}_{j=1}^{J}, \{\boldsymbol{Q}^{(j)}\}_{j=1}^{J}) = (\boldsymbol{\Omega})_{i_R,i_{R-1},\ldots,i_2,i_1,i_0} \\
&\quad \equiv \Lambda_1 \sum_{\ell=1}^{R} \lambda_{\ell}^{(1)} Q^{(1)}(s_t = i_0 \mid s_{t-\ell} = i_\ell) + \\
&\qquad + \Lambda_2 \sum_{\ell_1<\ell_2} \sum \lambda_{(\ell_1,\ell_2)}^{(2)} Q^{(2)}(s_t = i_0 \mid s_{t-\ell_1} = i_{\ell_1}, s_{t-\ell_2} = i_{\ell_2}) + \ldots + \\
&\qquad + \Lambda_J \sum_{\ell_1<\ldots<\ell_J} \cdots \sum \lambda_{(\ell_1,\ldots,\ell_J)}^{(J)} Q^{(J)}(s_t = i_0 \mid s_{t-\ell_1} = i_{\ell_1}, \ldots, s_{t-\ell_J} = i_{\ell_J}),
\end{aligned}
\tag{2}
$$

where $\boldsymbol{\lambda}^{(j)}$ is a probability vector of length $\binom{R}{j}$ for $j = 1, \ldots, J$. 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 $\boldsymbol{Q}^{(j)}$. However this parameterization is more informative about important orders (via inference for $\boldsymbol{\Lambda}$) in addition to lags (via inference for $\boldsymbol{\lambda}^{(j)}$). If $\Lambda_1 = 1$, then $\Lambda_j = 0$ for $j > 1$ and we recover the original MTD model. Clearly, the transitions associated with $\boldsymbol{Q}^{(j)}$ fully allow nonlinear dynamics in j dimensions of the lag space. As a discrete mixture of probability distributions, this model produces a valid probability tensor.

The model in (2) is clearly over-parameterized, and consequently $\boldsymbol{\Lambda}$, $\{\boldsymbol{\lambda}^{(j)}\}_{j=1}^{J}$, and $\{\boldsymbol{Q}^{(j)}\}_{j=1}^{J}$ are not fully identified. Of primary concern is when the lags with substantial weight (λ high) in lower orders are a subset of the lags with substantial weight in higher orders. To see this, consider writing the model for $J = 2$, $R = 3$, $K = 2$. This case is not strictly included in our model definition, but can approximately occur in a model with $R = 3$ when $\lambda_3^{(1)}$, $\lambda_{1,3}^{(2)}$, and $\lambda_{2,3}^{(2)}$ are all approximately equal to 0. In this case, $\boldsymbol{Q}^{(2)}$ is sufficient to model this chain with full flexibility, yet there are effectively four additional, unnecessary parameters. We

Table 1: Free parameter count for MMTD model under different combinations of state space size K, largest possible lag R, and largest mixing order J. The total number of parameters estimated is the sum of the free Λ, λ, and Q parameters. The unrestricted total is the number of parameters required to estimate an unrestricted transition probability tensor of order R.

K	R	J	Λ	λ	Q	total	unrestricted
5	3	2	1	4	120	125	500
5	4	3	2	11	620	633	2,500
2	7	4	3	94	30	127	128
3	5	3	2	22	78	102	486
7	5	3	2	22	2,394	2,418	100,842
7	5	2	1	13	336	350	100,842

note, however, that the resulting transition tensor $\boldsymbol{\Omega}$ is still identifiable even if the individual parameters are not unique. It is also possible for a higher order $\boldsymbol{Q}$ to mimic a lower order tensor through repetition of transition probabilities across values of a certain lag. Absent complicated constraints, we note that these issues are detectable through inferences for $\boldsymbol{\Lambda}$, $\{\boldsymbol{\lambda}^{(j)}\}_{j=1}^{J}$, and $\{\boldsymbol{Q}^{(j)}\}_{j=1}^{J}$, and we recommend a modeler always check for them. We further address the identifiability issue with prior specification in Section 2.4.

We envision two primary uses for this model. The first is to uncover low-order structure from data whose practical lag dependence horizon is truly smaller than our over-specified R and the order is less than or equal to our selected J, 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 might fit the MMTD model with $R = 10$ and $J = 2$, anticipating Λ_2 to carry most posterior weight, and $\boldsymbol{\lambda}^{(2)}$ to identify the influential lags. In Section 2.4, we consider using priors which encourage sparsity to assist in shrinking back to the true model.

If the true order of dependence in the time series is greater than J, our second intended use for this model is analogous to the original MTD model, as we parsimoniously approximate higher order dependence by mixing lower order transition distributions. Adding these higher order $\boldsymbol{Q}$ tensors could be thought of as including "interaction" terms in the mixture.

Our proposed model formulation requires estimation of $J-1$ free Λ parameters, $\binom{R}{1}+\binom{R}{2}+\ldots+\binom{R}{J}-J$ free λ parameters, and $K(K-1)+K^2(K-1)+\ldots+K^J(K-1) = K(K^J-1)$ free parameters in the J base transition tensors $\boldsymbol{Q}$. Table 1 calculates the total number of parameters to estimate for different combinations of K, R, and J. The model grows primarily with K and J, which should be kept relatively small in applications.

2.3 Bayesian inference and computation

As noted earlier, all inferences are conditional on the first R observations in the time series $\{s_t\}_{t=1}^T$. To define a Bayesian model, we first break the mixture in (2) by introducing latent indicators Z_t such that $\Pr(Z_t = j) = \Lambda_j$ independently for each time point $t = R+1$ to $t = T$. Then conditional on Z_t, introduce latent indicators $\boldsymbol{z}_t$ such that $\Pr(\boldsymbol{z}_t = (\ell_1, \ldots, \ell_j) \mid Z_t = j) = \lambda^{(j)}_{(\ell_1,\ldots,\ell_j)}$, independently for each $t = R+1, \ldots, T$. The hierarchical formulation for this model is given as follows. For $t = R+1, \ldots, T, i_0, i_\ell \in \{1, \ldots, K\}$, $\ell \in \{1, \ldots, R\}$, $\ell_1 < \ldots < \ell_j$, $j \in \{1, \ldots, J\}$, we have

$$
\begin{aligned}
&(\boldsymbol{Q}^{(j)})_{i_j, i_{j-1}, \ldots, i_1, \cdot} \sim \mathrm{Dir}(\boldsymbol{\alpha}_{\boldsymbol{Q}^{(j)}}) \quad \forall \quad (i_j, i_{j-1}, \ldots, i_1) \in \{1, \ldots, K\}^j \,, \\
&\boldsymbol{\Lambda} \sim \mathrm{Dir}(\boldsymbol{\alpha}_\Lambda) \,, \qquad\qquad \boldsymbol{\lambda}^{(j)} \sim \mathrm{Dir}(\boldsymbol{\alpha}_{\lambda^{(j)}}) \,, \\
&\Pr(Z_t = j \mid \boldsymbol{\Lambda}) = \Lambda_j \,, \qquad \Pr(\boldsymbol{z}_t = (\ell_1, \ldots, \ell_j) \mid Z_t = j, \boldsymbol{\lambda}^{(j)}) = \lambda^{(j)}_{(\ell_1,\ldots,\ell_j)} \,, \\
&\Pr(s_t = i_0 \mid s_{t-1} = i_1, s_{t-2} = i_2, \ldots, s_{t-R} = i_R, Z_t = j, z_t^{(j)} = (\ell_1, \ldots, \ell_j), \boldsymbol{Q}^{(j)}) = (\boldsymbol{Q}^{(j)})_{i_{\ell_j}, i_{\ell_{j-1}}, \ldots, i_{\ell_1}, i_0} \,,
\end{aligned}
\tag{3}
$$

where all quantities without explicit dependence are considered independent a priori. Given this structure, all posterior inference can be accomplished entirely through closed-form Gibbs sampling. To simplify computation, we uniquely map all Z_t and $\boldsymbol{z}_t$ pairs into a single variable $\zeta_t \in \left\{1, \ldots, \left[\binom{R}{1} + \binom{R}{2} + \ldots + \binom{R}{J}\right]\right\}$ whose prior probability under the model is equal to the product of the corresponding Λ and λ. Full conditional distributions for $\boldsymbol{\Lambda}$, each $\boldsymbol{\lambda}^{(j)}$, and each $(\boldsymbol{Q}^{(j)})_{i_j, i_{j-1}, \ldots, i_1, \cdot}$ are exactly analogous to multinomial-Dirichlet conjugate updates where Z_t, $\boldsymbol{z}_t$, and observed data transitions supply the multinomial counts. Full conditional updates for Z_t and $\boldsymbol{z}_t$ (equivalently ζ_t) require calculation and sampling from a discrete distribution.

As is common with mixture models, the joint posterior distribution resulting from (3) is multimodal and the Gibbs sampler described above is prone to poor mixing. We improve mixing in the sampler by integrating $\{\boldsymbol{Q}^{(j)}\}_{j=1}^J$ from the joint posterior, sampling the collapsed conditional distributions for $\boldsymbol{\Lambda}$, each $\boldsymbol{\lambda}^{(j)}$, and $\{\zeta_t\}$. These are supported by the tractable marginal distributions reported in Appendix A. Additionally, to encourage occasional jumps between modes of the posterior, we include a hybrid independence Metropolis step which jointly proposes $\boldsymbol{\Lambda}$, each $\boldsymbol{\lambda}^{(j)}$, and $\{\zeta_t\}$ from their joint prior every 25 iterations of MCMC.

To obtain results in Sections 3.1, 3.2, and 4 that follow, each model was initialized with random draws from the Dirichlet priors for $\boldsymbol{\Lambda}$ and each $\boldsymbol{\lambda}^{(j)}$, and discrete uniform for $\{\zeta_t\}$. Random initialization in these models necessitated long burn-in periods, on the order of tens to hundreds of thousands of iterations. In

our analyses, 100,000 burn-in iterations were followed by another 400,000 iterations, producing (unless otherwise noted) stable chains suitable for inference. Reported posterior quantities were calculated using a thinned sample retaining every 10^{th} iteration. Posterior sampling was conducted using the *Julia* scientific computing language (Bezanson et al., 2017). Details for the MCMC algorithm are given in Appendix B.

2.4 Order estimation with sparse probability vectors

One intended primary use of this model is to uncover low-order structure from a high-order model by over-specifying R and possibly J. If the true model is contained within the full MMTD specification, our goal is to identify that structure through inferences for $\boldsymbol{\Lambda}$ and $\boldsymbol{\lambda}^{(j)}$, which should concentrate posterior mass on particular components. In this case, $\boldsymbol{\Lambda}$ and $\boldsymbol{\lambda}^{(j)}$ should be sparse, which we can encourage by replacing the Dirichlet priors in (3) by the sparse Dirichlet mixture (SDM) or stick-breaking mixture (SBM) priors proposed by Heiner et al. (2018).

The SDM prior is a fixed-weight mixture of Dirichlet densities, each featuring a "boost" of equivalent sample size β in one of the categories. If $\boldsymbol{\theta}$ is a K-length probability vector, the SDM density is given as

$$p_{\text{SDM}}(\boldsymbol{\theta}; \boldsymbol{\alpha}, \beta) = \sum_{k=1}^{K} \frac{w_k}{\sum_{j=1}^{K} w_j} \text{Dir}(\boldsymbol{\theta}; \boldsymbol{\alpha} + \beta \mathbf{e}_k)\,, \tag{4}$$

where $w_k = \prod_{j=1}^{K} \Gamma(\alpha_j + \beta 1_{(j=k)})$ and $\mathbf{e}_k$ is a vector of 0s with a 1 in the k^{th} position, and $\Gamma(\cdot)$ is the gamma function. For small sample sizes and relatively large β, the SDM can be characterized as a "winner-takes-all" prior.

The 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,

$$\theta_1 = X_1,\ \theta_k = X_k \prod_{j=1}^{k-1}(1 - X_j) \text{ for } k = 2, \ldots, K-1, \text{ and } \theta_K = 1 \cdot \prod_{j=1}^{K-1}(1 - X_j)\,, \tag{5}$$

with X_k independently drawn from a mixture of two beta distributions, $X_k \overset{\text{ind.}}{\sim} \pi\text{Beta}(1, \eta) + (1 - \pi)\text{Beta}(\gamma, \delta)$. This allows us to encourage sparsity by setting η large, in which case the first component corresponds to small probabilities in $\boldsymbol{\theta}$. While the second mixture component should favor larger values of X_k, we recommend leaving them relatively non-informative, as the second beta component must be sufficiently flexible to break various sizes of the "remaining stick" (i.e., $\prod_{j=1}^{k-1}(1 - X_j)$) to provide the non-negligible elements of $\boldsymbol{\theta}$. Depending on the desired level of sparsity, we recommend using values of π

in the 0.5 to 0.95 range. The value of π should be chosen with care due to the asymmetry and truncation of the SBM prior, as discussed in Heiner et al. (2018).

If the hyperparameters of the SDM and SBM priors are fixed, as is usually the case with the original Dirichlet priors, incorporating these into the model requires minimal effort since posterior Gibbs sampling can proceed with tractable conditional distributions, leaving the updates in Appendix B structurally unchanged.

In scenarios where the modeler believes only one order/lag combination should dominate, but is unsure which it is, we recommend replacing the Dirichlet prior for $\boldsymbol{\Lambda}$ and potentially each $\boldsymbol{\lambda}^{(j)}$ with the SDM prior. Because the SBM prior can accommodate more than one non-negligible probability, it may also provide a compromise between the two objectives of shrinking an over-specified model and mixing in the spirit of the MTD. Furthermore, if R and J are low and the transition probability tensor is known to be sparse, it *may potentially* be advantageous to replace the Dirichlet priors on the transition distributions in $\{\boldsymbol{Q}^{(j)}\}_{j=1}^{J}$ with independent SBM priors. However, this should be done with care, as the prior produces strongly biased estimates when no transition counts are observed.

3 Simulation studies

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 different 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 contribute to transition probabilities. In both models, each distribution in the transition tensor $\boldsymbol{\Omega}$ 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 retained as 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 $p_{t'}^{(k)} = \Pr(s_{t'} = k \mid s_{t'-1}, \ldots, s_{t'-R}) = (\boldsymbol{\Omega})_{s_{t'-R},\ldots,s_{t'-1},k}$ 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 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 $100 \times \sum_{t'} L_{t'}/(KT')$ where $T' = 1000$ is the number of validation points.

We fit the MMTD to each training set with various settings. Let MMTD(R, J) denote a model fit with specified maximum lag horizon R and maximum order J, and MTD(R) = MMTD$(R, 1)$. Unless otherwise specified, all MMTD models follow (3) with all symmetric Dirichlet priors in which all shape parameters sum to unity. Specifically, the shape parameters for the Dirichlet prior on $\boldsymbol{\Lambda}$ are all equal to $1/J$; for $\boldsymbol{\lambda}^{(j)}$, the shape parameters are all equal to $1/\binom{R}{j}$; and for all K^j Dirichlet priors for $\boldsymbol{Q}^{(j)}$, the shape parameters are all equal to $1/K$. We denote with SDM$(\boldsymbol{\Lambda})$ use of the SDM prior on $\boldsymbol{\Lambda}$. In all cases, the base shape parameters in the SDM were set identical to those used in the corresponding Dirichlet priors. We also used $\beta = T/4$ in all cases to encourage sparsity. Results were not sensitive to the alternate specification $\beta = T/2$.

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

3.1 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 L_1 loss across the 1,000 validation points are given in Table 2. In addition to transition probability estimation, we are interested in inferences for Markovian order and important lags afforded by the MTD and MMTD models. With exception of the MTD only, we see improved estimation with the larger sample size across all models.

3.1.1 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 results in good estimation, but clearly this model enjoys an unfair advantage. It may, however, result from an iterative model selection process. We emphasize here that the MMTD does not require a model selection process if the modeler specifies the maximum lag horizon R and

Table 2: Simulation results for transition probability estimation under various models and model settings using two sample sizes of Simulation 1 data, $T = 200$ and $T = 500$. The reported loss is 100 times the mean L_1 loss, computed across 1,000 validation time points. Within each sample size group, the lowest mean loss is highlighted with bold font.

$T = 200$		$T = 500$	
model	loss	model	loss
LogitMC(6, 1)	18.48	LogitMC(6, 1)	16.78
LogitMC(6, 2)	36.61	LogitMC(6, 2)	18.88
LogitMC(3, 1), Lags 1, 3, 4 only	17.03	LogitMC(3, 1), Lags 1, 3, 4 only	16.39
LogitMC(3, 2), Lags 1, 3, 4 only	13.44	LogitMC(3, 2), Lags 1, 3, 4 only	10.42
LogitMC(3, 3), Lags 1, 3, 4 only	**12.29**	LogitMC(3, 3), Lags 1, 3, 4 only	7.60
VLMC	19.01	VLMC	15.24
MTD(6)	17.17	MTD(6)	17.33
MTD(6), SDM($\boldsymbol{\lambda}$)	17.13	MTD(6), SDM($\boldsymbol{\lambda}$)	17.11
MMTD(6, 2)	14.71	MMTD(6, 2)	13.95
MMTD(6, 2), SDM($\boldsymbol{\Lambda}$), SDM($\boldsymbol{\lambda}$)	14.53	MMTD(6, 2), SDM($\boldsymbol{\Lambda}$), SDM($\boldsymbol{\lambda}$)	13.91
MMTD(6, 3)	14.30	MMTD(6, 3)	7.44
MMTD(6, 3), SDM($\boldsymbol{\Lambda}$), SDM($\boldsymbol{\lambda}$)	14.41	MMTD(6, 3), SDM($\boldsymbol{\Lambda}$), SDM($\boldsymbol{\lambda}$)	7.28
MMTD(6, 4)	14.71	MMTD(6, 4)	7.50
MMTD(6, 4), SDM($\boldsymbol{\Lambda}$), SDM($\boldsymbol{\lambda}$)	13.77	MMTD(6, 4), SDM($\boldsymbol{\Lambda}$), SDM($\boldsymbol{\lambda}$)	7.30
MMTD(6, 5)	15.20	MMTD(6, 5)	7.45
MMTD(6, 5), SDM($\boldsymbol{\Lambda}$), SDM($\boldsymbol{\lambda}$)	16.26	MMTD(6, 5), SDM($\boldsymbol{\Lambda}$), SDM($\boldsymbol{\lambda}$)	**7.26**

maximum order J, as order and lag inferences are built-in.

The variable length Markov chain model offers no improvement over the older methods, possibly because Lag 2 plays no role in Simulation 1 and VLMC branches must include Lag 2 because more distant lags are important, causing it to miss an opportunity at greater parsimony (Jääskinen et al., 2014).

The MTD model offers little help in this scenario because Simulation 1 is third order with nonlinear interactions. Posterior densities for $\boldsymbol{\lambda}$ (not shown) reveal that λ_3 is favored under the Dirichlet prior and dominates with the SDM prior. The latter effectively produces a first order Markov chain dependent on the third lag.

Several MMTD models were fit with increasing maximum order J ranging from 2 to 5. As expected, estimation performance generally stops improving when J exceeds the true order 3. Surprisingly, the SDM priors help in only two of four cases. In the $J = 2$ model, posterior mass slightly favors second order with the Lag 3,4 combination receiving most posterior weight. Adding the SDM priors on $\boldsymbol{\Lambda}$ and each $\boldsymbol{\lambda}^{(j)}$ for $j = 1, 2$, more strongly supports the same conclusion. In the $J = 3$ model, posterior mass slightly favors third order over first and second, with the correct Lag 1, 3, and 4 combination receiving most posterior weight. The Lag 3, 4, and 5 combination is also favored over the others. Adding the SDM priors led to

strong selection of second order dynamics in both of two independent runs, yielding results nearly identical to those of the $J = 2$ model with SDM priors. The $J = 4$ model favors fourth order and Lags 1, 3, 4, and 5. Here the model overestimates the order, underlining the need for a modeler to examine the estimate of $\boldsymbol{Q}^{(4)}$ for redundancy along one of the lag dimensions, as we demonstrate in Section 4.2. Adding the SDM prior tends to favor second order with Lags 3 and 4, although different MCMC runs explore different secondary modes (first order in one and fourth order in another). The $J = 5$ model mimics the $J = 4$ model, except the fifth order receives some posterior mass under the Dirichlet prior and most mass under the SDM. This explains the degraded estimation performance. We generally do not advocate considering such high orders when $T = 200$, even with a relatively small state space. Overall, the MMTD consistently produces the most faithful estimates of transition probabilities from a single model without requiring iterative model selection.

3.1.2 Sample size 500

In the $T = 500$ case, even the multinomial logistic models with an unfair fit to the correct lags fails to outperform the MMTD with $J \geq 3$. With a larger sample size, the VLMC model is more competitive, but the MTD is unable to capitalize. The MTD model mixes over Lags 2, 4, and 5 with a Dirichlet prior on $\boldsymbol{\lambda}$ and Lags 3 and 4 with the SDM prior.

In all cases of the MMTD, the SDM prior on order and lag structure appears to improve estimation accuracy. The inferences from $J = 2$ are similar to those of the $J = 2$ models fit to $T = 200$ observations under their respective prior specifications. The increased sample size especially assists in identifying lag structure beginning with the $J = 3$ models, where third order is the clear preference and the correct Lags 1, 3, and 4 are always identified. Under the model with the SDM prior, posterior densities for $\boldsymbol{\Lambda}$ and each $\boldsymbol{\lambda}^{(j)}$ exhibit strong peaks concentrated at the true values. These results hold in the $J = 4$ and $J = 5$ cases as well. Among the models considered in this scenario, the MMTD consistently produces the most faithful estimates of transition probabilities.

3.2 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 six lags, which we use where possible to promote equitable comparisons. Results of the mean L_1 loss across the 1,000 validation points are given in Table 3. Again, we examine order and lag inferences from MTD and MMTD models in addition to estimation performance.

Table 3: Simulation results for transition probability estimation under various models and model settings using three sample sizes of Simulation 2 data, $T = 100$, $T = 200$ and $T = 500$. The reported loss is 100 times the mean L_1 loss, computed across 1,000 validation time points. Within each sample size group, the lowest mean loss is highlighted with bold font. Here SDM($\boldsymbol{\Lambda}$, $\boldsymbol{\lambda}$) indicates that both $\boldsymbol{\Lambda}$ and all $\boldsymbol{\lambda}$ parameters have SDM priors.

$T = 100$		$T = 200$		$T = 500$	
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
VLMC	21.28	VLMC	14.37	VLMC	12.00
MTD(7)	24.73	MTD(7)	22.47	MTD(7)	19.83
MTD(7), SDM($\boldsymbol{\lambda}$)	24.20	MTD(7), SDM($\boldsymbol{\lambda}$)	23.56	MTD(7), SDM($\boldsymbol{\lambda}$)	21.85
MMTD(7, 4)	21.65	MMTD(7, 4)	14.76	MMTD(7, 4)	12.60
MMTD(7, 4), SDM($\boldsymbol{\Lambda}$)	20.51	MMTD(7, 4), SDM($\boldsymbol{\Lambda}$)	14.73	MMTD(7, 4), SDM($\boldsymbol{\Lambda}$)	13.32
MMTD(7, 4), SDM($\boldsymbol{\Lambda}$, $\boldsymbol{\lambda}$)	20.62	MMTD(7, 4), SDM($\boldsymbol{\Lambda}$, $\boldsymbol{\lambda}$)	14.54	MMTD(7, 4), SDM($\boldsymbol{\Lambda}$, $\boldsymbol{\lambda}$)	12.31
MMTD(7, 7)	20.67	MMTD(7, 7)	12.80	MMTD(7, 7)	7.47
MMTD(7, 7), SDM($\boldsymbol{\Lambda}$)	18.80	MMTD(7, 7), SDM($\boldsymbol{\Lambda}$)	11.29	MMTD(7, 7), SDM($\boldsymbol{\Lambda}$)	7.34
MMTD(7, 7), SDM($\boldsymbol{\Lambda}$, $\boldsymbol{\lambda}$)	**17.84**	MMTD(7, 7), SDM($\boldsymbol{\Lambda}$, $\boldsymbol{\lambda}$)	**9.81**	MMTD(7, 7), SDM($\boldsymbol{\Lambda}$, $\boldsymbol{\lambda}$)	**7.26**

3.2.1 Sample size 100

In the $T = 100$ case, high order interactions are not supported by the multinomial logistic model. The VLMC model is competitive with the MMTD and outperforms the standard models, presumably because Simulation 2 features no gap in important lags.

Because the simulation follows fifth order dynamics, the MTD(7) and MMTD(7, 4) models are under-specified and must rely on a lower order sub-model and/or mixing across lags to produce estimates. The MTD models mix over Lags 2 and 4. The MMTD(7, 4) model with Dirichlet priors switches between all four orders without favoring a clear lag pattern. Adding the SDM prior on $\boldsymbol{\Lambda}$ only tends to favor third and fourth orders while remaining agnostic to lag pattern. Adding the SDM prior on each $\boldsymbol{\lambda}^{(j)}$ leads the model to favor first and fourth orders.

The over-specified MMTD(7, 7) does not outperform the $J = 4$ models without SDM priors encouraging shrinkage, although it does weakly favor fifth order and the correct configuration of Lags 1 through 5. The model remains primarily agnostic under the SDM prior on $\boldsymbol{\Lambda}$ only, although the correct structure begins to emerge. Adding the SDM prior for each $\boldsymbol{\lambda}^{(j)}$ further clarifies the inference.

3.2.2 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 standard models.

As before, the MTD is unable to leverage increased sample size to the extent that the other models can. The MTD models mix primarily over Lags 1 and 5. The MMTD(7, 4) model with Dirichlet priors clearly favors fourth order with Lags 1, 2, 3, and 5. Adding the SDM prior on $\mathbf{\Lambda}$ only tends to favor third and fourth orders, with the same lag pattern as before. Adding the SDM prior on each $\boldsymbol{\lambda}^{(j)}$ results in clear preference for fourth order and the same lag pattern as the original MMTD(7, 4).

With this sample size, the correct structure emerges in MMTD(7, 7). With Dirichlet priors, $\mathbf{\Lambda}_4$ carries some posterior weight. As SDM priors are added, the corresponding inferences are sharpened in support of the truth (Λ_5 and $\lambda^{(5)}_{(1,2,3,4,5)}$ are close to 1). The final model, with the best estimation performance, is essentially equivalent to fitting a fifth order chain to the correct lags.

3.2.3 Sample size 500

The fifth order binary chain in Simulation 2 has 32 total (univariate) transition distributions which are easily identified 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) models lag noticeably behind in estimation performance.

The MMTD(7, 4) behaves similarly to the same model in the $T = 200$ case, and offers performance on par with the VLMC. The MMTD(7, 7) decisively identifies the correct order and lag structure with one exception: the model with the SDM prior on $\mathbf{\Lambda}$ only occasionally supports sixth order. Notice that with the larger sample size, the advantage of the SDM priors is less pronounced. We require less shrinkage of an over-specified model to the true subset. As before, we conclude that the MMTD consistently produces the most faithful estimates of transition probabilities. Furthermore, it offers insight into the lag dependence structure.

4 Data illustrations

We now apply the MMTD to two data analyses. The first was analyzed with the original MTD and subsequent literature. The second is a novel analysis of pink salmon population dynamics over the second half of the twentieth century. We illustrate the use of inferences on order and lag importance available

from the MMTD model.

4.1 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 Bayesian information criterion (BIC) to ultimately select a MTD with eight lags. Their estimation yields λ_8 with the greatest magnitude. We note that the MTD model used in Berchtold and Raftery (2002) allows negative values in $\boldsymbol{\lambda}$. To do so requires a complex set of constraints not practical for our approach in which each $\boldsymbol{\lambda}^{(j)}$ is a probability vector.

This time series was revisited and fit using the methods in Sarkar and Dunson (2016), who report a model of maximal order 8 to be most likely, with Lag 8 having the highest posterior inclusion probability. In contrast with Berchtold and Raftery (2002), they find Lag 1 to be the second most important. They report the posterior mode for the number of important lags to be three.

In light of these two analyses, we fit the MMTD to the seizure data with $R = 10$ and $J = 3$, and various priors, each with hyperparameter settings equivalent to those in the simulation studies. Trace plots (not shown) indicate that the posterior over $\boldsymbol{\Lambda}$ and each $\boldsymbol{\lambda}^{(j)}$ is multimodal, suggesting that many combinations of lags could model the dynamics with similar accuracy, but also that the dynamics probably lack homogeneity (no seizures were reported in the last 29 days).

The MMTD(10, 3) model with all Dirichlet priors mixes across all orders, with the most posterior weight going to Λ_3 which has a posterior mean of 0.425. Among the first order transitions, Lags 4, 8, and 9 are favored with posterior means for their respective $\lambda^{(1)}$ parameters in the 0.12 to 0.17 range. Among the second order transitions, no lag combination stands out. However, at least one of Lags 4, 8, and 9 appear in the lag pairs with the 12 highest posterior means, accounting for .42 of the posterior mean of the $\boldsymbol{\lambda}^{(2)}$ vector. The third order component of the model likewise does not select a most important lag combination. However, Lag 8 appears in lag combinations that account for approximately 0.46 of the posterior mean of $\boldsymbol{\lambda}^{(3)}$, with Lags 9 and 4 taking 0.41 and 0.34, respectively. The MMTD(10, 3) model with SDM priors on $\boldsymbol{\Lambda}$ and the $\boldsymbol{\lambda}^{(j)}$s more heavily favors third order dynamics with Λ_3 having a posterior mean of 0.626, consistent with the findings of Sarkar and Dunson (2016). In contrast with their results, however, Lag 1 is less important in our model, which favors the combination of Lags 3, 4, and 8 (with 0.22 being the posterior mean for the corresponding $\lambda^{(3)}$).

We can also assess overall lag importance by computing a lag inclusion probability as the sum of all

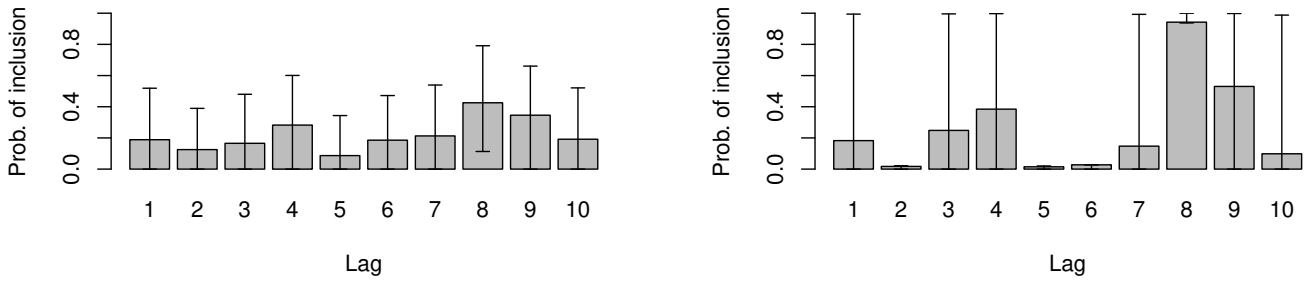

Figure 1: Posterior mean (with 95% credible interval) probability of inclusion for each lag in the seizure analysis using the Dirichlet priors (left) and SDM priors (right).

products $\Lambda_j \times \lambda_r^{(j)}$ for which lag ℓ appears in the lag configuration $\boldsymbol{z}_r$. We compute this for each lag at each MCMC sample. These inclusion plots for the models fit to the seizure data with Dirichlet priors and SDM priors are shown in Figure 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 probability for all lags except Lag 8, especially in the SDM model. We further note that the plot on the right associated with the SDM priors resemble a plot with similar interpretation in Figure 6 (e) of Sarkar and Dunson (2016), with exception that Lags 1 and 2 have lower inclusion probability and Lag 9 has higher inclusion probability in our model. All four analyses, including our two, agree that Lag 8 is the most important in determining the transition probability.

4.2 Pink salmon data

We analyze a time series of annual pink salmon abundance (escapement) in Alaska, U.S.A.[1] from 1932 to 1985. Population dynamics for pink salmon provide a testing opportunity for out 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 2. As with the seizure data, we suspect non-stationarity and lack of homogeneity with long-term trends. It appears that the even and odd year populations diverged in the late 1940s, with perhaps some intervention in the early 1960s. Despite these trends, we expect to extract informative inferences on lag dependence.

Prior to analysis, we imputed the missing value for 1962 as the geometric mean of escapements in 1960

[1]Data obtained through personal correspondence.

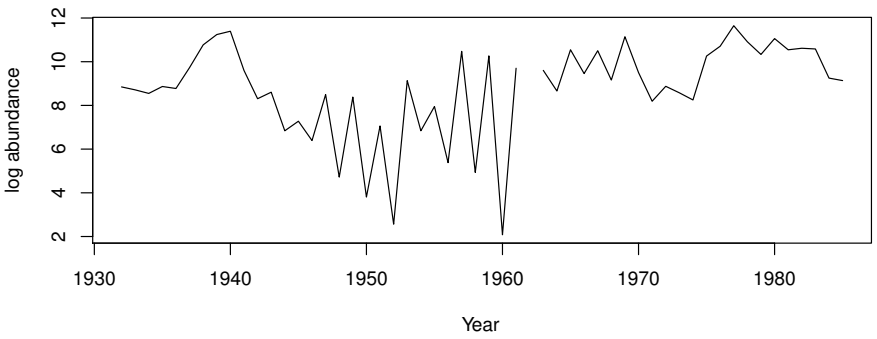

Figure 2: Time series plot (left) and bivariate lag plots for the natural logarithm of pink salmon abundance from 1932 to 1985.

and 1964 (the mean of the logarithm for these years). Although this choice will have some impact on model inferences, it represents only one year of the 54 observations and can be justified given the pink salmon life cycle. After discretizing the data into sets of $K = 4$ and $K = 7$ quantile-based bins using all 54 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 transformations such as the natural logarithm.

MTD(7) models, both with Dirichlet and SDM priors on $\boldsymbol{\lambda}$, favor Lag 4 in the $K = 4$ case. In the $K = 7$ models $\boldsymbol{\lambda}$ mixes over several lags in model with the Dirichlet prior on $\boldsymbol{\lambda}$ and strongly favors Lag 2 under the SDM prior. When we consider higher orders ($J = 3$), the $K = 4$ discretization favors second or third order without strongly favoring any particular lag combination. However, Lag 2 has the highest posterior mean probability of inclusion at 0.43 (<0.001, 0.89), with Lag 4 next at 0.38 (<0.001, 0.85) and Lag 6 close behind at 0.37 (<0.001, 0.86). Under the SDM prior, Lag 2 has a higher posterior mean of inclusion at 0.78, but the credible interval is not informative because the marginal posterior is bimodal. The $K = 7$ case, while spreading the data thin, produces interesting results. Both sets of priors favor second order with the (2, 3) and (2, 4) lag combinations emerging as important. Marginal posterior density plots for $\boldsymbol{\Lambda}$ and a few entries in $\boldsymbol{\lambda}^{(2)}$ in the Dirichlet case are shown in Figure 3. The SDM priors yield the similar results, with more posterior weight concentrated on Order 2 and the same lag combinations. Lag inclusion plots for both models are shown in Figure 4. Again, the model which shrinks down the over-specified MMTD with SDM priors yields more confident inferences for inclusion. Clearly Lag 2 is active, as expected, and Lags 3 and 4 appear to have a role.

It is important to examine the estimate of $\boldsymbol{Q}^{(2)}$ 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 $\boldsymbol{Q}^{(2)}$ would repeat across the second lag index (in this case representing Lag 3 and/or 4). The posterior

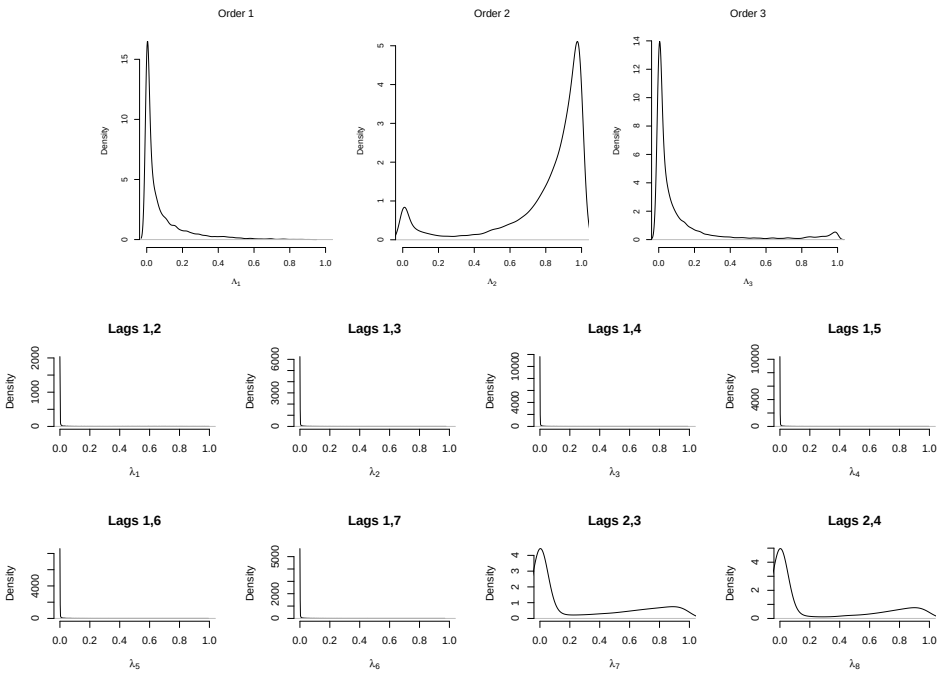

Figure 3: Marginal posterior density plots for $\boldsymbol{\Lambda}$ (top) and selected entries of $\boldsymbol{\lambda}^{(2)}$ (bottom) in the salmon analysis with $K = 7$ using Dirichlet priors.

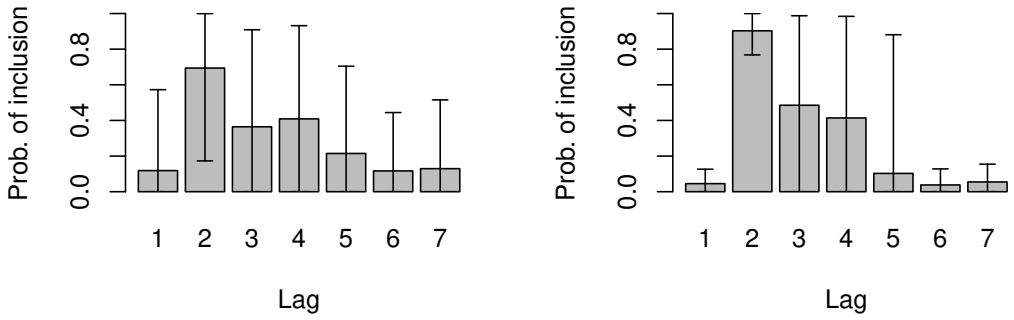

Figure 4: Posterior mean (with 95% credible interval) probability of inclusion for each lag in the salmon analysis with $K = 7$ using the Dirichlet priors (left) and SDM priors (right).

mean point estimate of $\boldsymbol{Q}^{(2)}$, shown in Figure 5 for both sets of priors, appears not to have this problem, as consecutive 7×7 sub-matrices appear not to repeat, except perhaps after the first and second. This is consistent under both priors, Lag 2 may not be the only important predictor, and that the second-order effects are not additive (and positively associated, as required by the positivity of $\boldsymbol{\lambda}$).

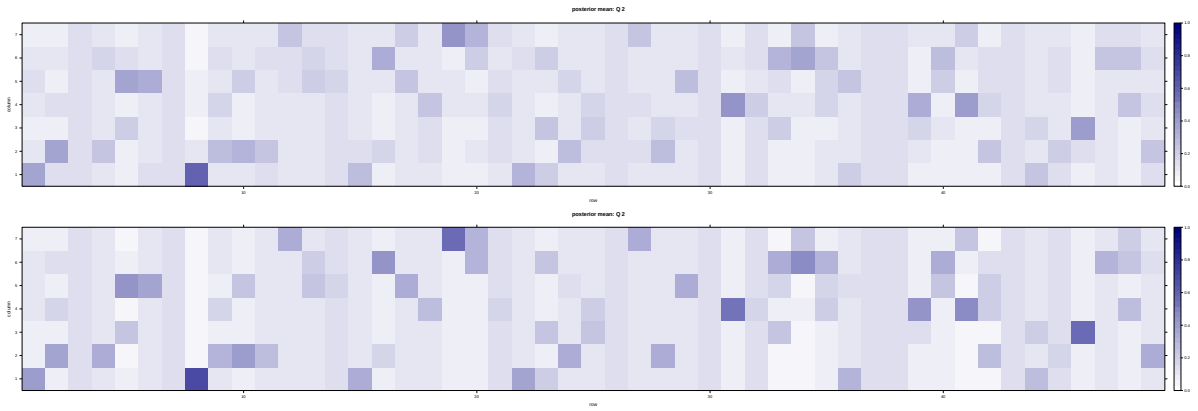

Figure 5: Posterior mean point estimate of the matricized $\boldsymbol{Q}^{(2)}$ for the salmon analysis with $K = 7$ using the Dirichlet priors (top) and SDM priors (bottom) on order and lag weights. Columns (along the y-axis) represent states to which the transition occurs, and rows (along the x-axis) represent the lag combinations, with the most recent lag changing index first. Hence the first row represents State 1 of the first selected lag and State 1 of the second selected lag, and the second row represents State 2 of the first selected lag and State 1 of the second selected lag. With $K = 7$ and $j = 2$, a second order tensor would come from stacking each consecutive 7×7 sub-matrix plotted here.

5 Summary

We have explored an extension of the original mixture transition distribution model for high order Markov chains which captures higher order interactions and can potentially yield useful inferences for order and lag importance. To accomplish the latter, the multi-mixture transition distribution model often over-specifies the true data-generating mechanism. We then dispatch sparsity-inducing priors with the goal of shrinking back to the truth in a single model without necessitating iterative model selection. Furthermore, our MCMC algorithm allows us to evaluate uncertainty about the model structure and transition probabilities. We demonstrated that our model can outperform a few 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 be immediately manifest from inferences for the mixture weight parameters. More often, lag importance should be aggregated and extracted in post-processing as we demonstrated in Section 4.1. 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.

The MMTD can approximate high order dynamics by exploiting additivity among lower-order transition probabilities. However, when this is not the case, a full jump to the next order is required. For example, in the salmon data analysis with seven states, the first-order mixture has seven components, whereas the second-order mixture has 21. A parsimonious compromise might rely on some factorization of the second-order tensor, as in Sarkar and Dunson (2016). We do not pursue this here, but rather choose to emphasize the straightforward and interpretable structure of our model (2) which showcases a model-averaging flavor, with added flexibility.

A 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 B.2.

Consider a sequence of independent random variables $\{s_t\} \in \{1, \ldots, K\}^T$ with common distribution $\boldsymbol{\theta} \equiv (\theta_1, \ldots, \theta_K)$. 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. If the ordering t is not important, the probability is multiplied by the multinomial coefficient $T!/(n_1! \cdots n_K!)$.

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

$$p(\{s_t\}) = \int p(\{s_t\} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \frac{\prod_k \Gamma(\alpha_k + n_k)}{\Gamma(\sum_k \alpha_k + n_k)} = \frac{\mathrm{MVB}(\boldsymbol{\alpha} + \boldsymbol{n})}{\mathrm{MVB}(\boldsymbol{\alpha})}\,, \tag{6}$$

where $\mathrm{MVB}(\cdot)$ denotes the multivariate beta function and the integral is defined on the simplex supporting $\boldsymbol{\theta}$.

If $\boldsymbol{\theta}$ follows a SDM distribution with parameters $\boldsymbol{\alpha}$ and β, and $w_k = \prod_{j=1}^K \Gamma(\alpha_j + \beta 1_{(j=k)})$, then the marginal distribution of $\{s_t\}$ is given by

$$p(\{s_t\}) = \sum_{k=1}^K \frac{w_k}{\sum_{j=1}^K w_j} \frac{\mathrm{MVB}(\boldsymbol{\alpha} + \beta \boldsymbol{e}_k + \boldsymbol{n})}{\mathrm{MVB}(\boldsymbol{\alpha} + \beta \boldsymbol{e}_k)}\,, \tag{7}$$

where $\boldsymbol{e}_k$ denotes a vector of 0s with a 1 in the k^{th} entry.

Under the generalized Dirichlet distribution (Connor and Mosimann, 1969) in which the stick-breaking latent variables are $X_k \overset{\text{iid}}{\sim} \mathrm{Beta}(a_k, b_k)$, we have $p(\{s_t\}) = \prod_{k=1}^{K-1} g_k(a_k, b_k, \boldsymbol{n})$ where

$$g_k(a_k, b_k, \boldsymbol{n}) = \frac{\Gamma(a_k + b_k)}{\Gamma(a_k^* + b_k^*)} \frac{\Gamma(a_k^*)}{\Gamma(a_k)} \frac{\Gamma(b_k^*)}{\Gamma(b_k)}\,,$$

with $a_k^* = a_k + n_k$, and $b_k^* = b_k + \sum_{j=k+1}^K n_j$. Using a similar approach, it can be shown that under the SBM model with parameters π, η, γ, and δ, we have

$$p(\{s_t\}) = \prod_{k=1}^{K-1} \left[\pi g_k(1, \eta, \boldsymbol{n}) + (1 - \pi) g_k(\gamma, \delta, \boldsymbol{n})\right]. \tag{8}$$

B MCMC algorithm details

Following the hierarchical MMTD model outlined in (3), the joint posterior distribution of all unknown parameters is given up to proportionality:

$$p\left(\boldsymbol{\Lambda}, \{\boldsymbol{\lambda}^{(j)}\}, \{\boldsymbol{Q}^{(j)}\}, \{Z_t\}, \{\boldsymbol{z}_t\} \mid \{s_t\}\right) \propto \mathrm{Dir}(\boldsymbol{\Lambda}) \prod_j \left[\mathrm{Dir}(\boldsymbol{\lambda}^{(j)}) \prod_{r=1}^{K^j} \mathrm{Dir}(\boldsymbol{Q}_{r,\cdot}^{(j)}) \right] \prod_t \left[\Lambda_{Z_t} \, \lambda_{\boldsymbol{z}_t}^{(Z_t)} \, Q^{(Z_t)}_{\varrho_{Z_t}(\boldsymbol{s}_{(t-\boldsymbol{z}_t)}, s_t)} \right], \tag{9}$$

where $\boldsymbol{Q}_{r,\cdot}^{(j)}$ denotes row r from a matricized version of $\boldsymbol{Q}^{(j)}$, $\varrho_j(\cdot)$ uniquely maps each possible combination of lags to a row of the matrcized $\boldsymbol{Q}^{(j)}$, and $\boldsymbol{s}_{(t-\boldsymbol{z}_t)}$ indicates the lags of s_t selected by $\boldsymbol{z}_t$.

B.1 Original algorithm

MCMC for the original hierarchical MMTD structure entirely with Gibbs updates. This is also the case when substituting in the SDM and/or SBM priors. A Gibbs sampler cycles through the parameters, drawing updates from the conditional distributions given below.

- $p(\boldsymbol{\Lambda} \mid \cdots) \propto \mathrm{Dir}(\boldsymbol{\Lambda}) \prod_t \Lambda_{Z_t}$, a standard Dirichlet/multinomial update using the counts of Z_t in each of $\{1, \ldots, J\}$. A SDM update for $\boldsymbol{\Lambda}$ is also trivial, as the full conditional remains SDM with the multinomial counts added to $\boldsymbol{\alpha}$, analogous to the Dirichlet full conditional.

- $p(\boldsymbol{\lambda}^{(j)} \mid \cdots) \propto \mathrm{Dir}(\boldsymbol{\lambda}^{(j)}) \prod_{t:Z_t=j} \lambda_{\boldsymbol{z}_t}^{(j)}$ independently for $j \in \{1, \ldots, J\}$. Again, this is a standard Dirichlet/multinomial update using the counts of lag configurations $\boldsymbol{z}_t$ within order j. SDM updates are also trivial, as the full conditionals remain SDM with the multinomial counts added to $\boldsymbol{\alpha}$, analogous to the Dirichlet full conditionals.

- $p(\boldsymbol{Q}_{r,\cdot}^{(j)} \mid \cdots) \propto \mathrm{Dir}(\boldsymbol{Q}_{r,\cdot}^{(j)}) \prod_{t:\varrho_j(\boldsymbol{s}_{(t-\boldsymbol{z}_t)})=r} Q_{r,s_t}^{(j)}$ independently for $j \in \{1, \ldots, J\}$ and $r \in \{1, \ldots, K^j\}$. Again, this is a standard Dirichlet/multinomial update using the transition counts. If a SBM prior is used here, the full conditionals are still available for convenient sampling using the SBM/multinomial update described in Heiner et al. (2018).

- $\Pr(\zeta_t = \ell \mid \cdots) \propto \Lambda_{Z_t} \, \lambda_{\boldsymbol{z}_t}^{(Z_t)} \, Q^{(Z_t)}_{\boldsymbol{s}_{(t-\boldsymbol{z}_t)}, s_t}$ independently for each $t \in \{R+1, \ldots, T\}$ with $\ell \in \left\{1, \ldots, \left[\binom{R}{1} + \binom{R}{2} + \ldots + \binom{R}{J}\right]\right\}$, where Z_t and $\boldsymbol{z}_t$ are uniquely mapped to ζ_t.

B.2 Modified algorithm

Iterated full-conditional sampling of both $\{\zeta_t\}$ and $\boldsymbol{Q}^{(j)}$ slows exploration of the joint posterior. To improve mixing, we instead integrate each $\boldsymbol{Q}^{(j)}$ out of the joint posterior (9) and conduct Gibbs sampling between $\{\zeta_t\}$, $\boldsymbol{\Lambda}$ and each $\boldsymbol{\lambda}^{(j)}$. At each iteration, it is then straightforward to draw each $\boldsymbol{Q}^{(j)}$ from the conditional distributions given in B.1.

For each $j \in \{1, \ldots, J\}$, let $\boldsymbol{N}^{(j)}$ be a matrix containing transition counts for which the (r, k) entry is the cardinality of $\{t : \varrho_j(\boldsymbol{s}_{(t-\boldsymbol{z}_t)}) = r \text{ and } s_t = k\}$. Integrating all $\boldsymbol{Q}^{(j)}$ from the full joint posterior proportional to (9) yields

$$p\left(\boldsymbol{\Lambda}, \{\boldsymbol{\lambda}^{(j)}\}, \{Z_t\}, \{\boldsymbol{z}_t\} \mid \{s_t\}\right) \propto \mathrm{Dir}(\boldsymbol{\Lambda}) \prod_j \left[\mathrm{Dir}(\boldsymbol{\lambda}^{(j)})\right] \prod_t \left[\Lambda_{Z_t} \lambda^{(Z_t)}_{\boldsymbol{z}_t}\right] \prod_j \left[\prod_{r=1}^{K^j} p\left(\boldsymbol{N}^{(j)}_{r,\cdot} \mid \{Z_t\}, \{\boldsymbol{z}_t\}\right)\right], \tag{10}$$

where $p\left(\boldsymbol{N}^{(j)}_{r,\cdot} \mid \{Z_t\}, \{\boldsymbol{z}_t\}\right)$ takes the form of (6) if the rows of matricized $\boldsymbol{Q}^{(j)}$ are independent Dirichlet, (7) if they are independent SDM, and (8) if they are independent SBM.

The modified algorithm then proceeds with the standard updates for $\boldsymbol{\Lambda}$ and each $\boldsymbol{\lambda}^{(j)}$ given in B.1. Each ζ_t is then updated individually using its collapsed conditional

$$p\left(\zeta_t \mid \cdots, -\{\boldsymbol{Q}^{(j)}\}\right) \propto \Lambda_{Z_t} \lambda^{(Z_t)}_{\boldsymbol{z}_t} \prod_j \left[\prod_{r=1}^{K^j} p\left(\boldsymbol{N}^{(j)}_{r,\cdot} \mid \{Z_t\}, \{\boldsymbol{z}_t\}\right)\right], \tag{11}$$

by modifying $\{\boldsymbol{N}^{(j)}\}$ to reflect each possible value of $\zeta_t \in \left\{1, \ldots, \left[\binom{R}{1} + \binom{R}{2} + \ldots + \binom{R}{J}\right]\right\}$.

References

Berchtold, A., and Raftery, A. E. (2002), "The mixture transition distribution model for high-order Markov chains and non-Gaussian time series," *Statistical Science*, 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.

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

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.

Fahrmeir, L., and Kaufmann, H. (1987), "Regression models for non-stationary categorical time series," *Journal of time series Analysis*, 8, 147–160.

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

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

Heard, W. R. (1991), "Life history of pink salmon (Oncorhynchus gorbuscha)," *Pacific salmon life histories*, 119–230.

Heiner, M., Kottas, A., and Munch, S. (2018), "Structured priors for sparse probability vectors with application to model selection in Markov chains," Tech. Rep. UCSC-SOE-18-06, Jack Baskin School of Engineering, University of California, Santa Cruz.

Insua, D., Ruggeri, F., and Wiper, M. (2012), *Bayesian analysis of stochastic process models* (Vol. 978), John Wiley & Sons.

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

Le, N., Martin, R., and Raftery, A. (1990), "Modeling Outliers, Bursts and Flat Stretches in Time Series Using Mixture Transition Distribution (MTD) Models," Tech. rep., Technical Report 194, University of Washington, Dept. of Statistics.

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

MacDonald, I. L., and Zucchini, W. (1997), *Hidden Markov and other models for discrete-valued time series* (Vol. 110), CRC Press.

Maechler, M. (2015), *VLMC: Variable Length Markov Chains ('VLMC') Models*, R package version 1.4-1.

Prado, R., and West, M. (2010), *Time series: modeling, computation, and inference*, CRC Press.

Raftery, A., and Tavaré, S. (1994), "Estimation and modelling repeated patterns in high order Markov chains with the mixture transition distribution model," *Applied Statistics*, 179–199.

Raftery, A. E. (1985), "A model for high-order Markov chains," *Journal of the Royal Statistical Society. Series B (Methodological)*, 528–539.

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*, ACM, pp. 35–46.

Sarkar, A., and Dunson, D. B. (2016), "Bayesian nonparametric modeling of higher order Markov chains," *Journal of the American Statistical Association*, 111, 1791–1803.

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*, 121–130.

Zucchini, W., and MacDonald, I. L. (2009), *Hidden Markov models for time series: an introduction using R* (Vol. 22), CRC press Boca Raton.