Bayesian Model Selection for Markov Chains using Sparse Probability Vectors

Matthew Heiner[1], Athanasios Kottas[1], and Stephan Munch[2]

[1]Department of Applied Mathematics and Statistics, University of California, Santa Cruz, California, USA

[2]Fisheries Ecology Division, Southwest Fisheries Science Center, National Marine Fisheries Service, NOAA, Santa Cruz, California, USA

Objectives

We seek a flexible Bayesian time series model to infer

1. Nonlinear dynamics.
2. Order and lag structure up to a pre-determined time horizon $R > 1$.

We extend a well-known mixture model for high order Markov chains and develop two novel priors for probability vectors. These priors, in contrast with the popular Dirichlet distribution, retain sparsity properties in the presence of data.

Mixture Transition Distribution model

The mixture transition distribution (MTD) model was introduced by Raftery (1985) (see Berchtold and Raftery (2002)) for high-order Markov chains. Consider a categorical time series $s_t \in \{1, \ldots, K\}$ at $t = 1, \ldots, T$. The R^{th} order MTD transition probabilities are

$$\Pr(s_t = i_0 \mid s_{t-1} = i_1, \ldots, s_{t-R} = i_R) \equiv \sum_{\ell=1}^{R} \lambda_\ell Q_{i_\ell, i_0},$$

with transition matrix $\boldsymbol{Q}$, $0 \leq \lambda_\ell \leq 1$ and $\sum_{\ell=1}^{R} \lambda_\ell = 1$. Important lags have relatively large λ_ℓ. Now let $J < R$ represent the highest-order "interaction" of lags. Introduce $\boldsymbol{Q}^{(1)}$, a $K \times K$ transition matrix, $\boldsymbol{Q}^{(2)}$, a $K \times K \times K$ transition tensor, and so forth. Next, introduce a mixing probability vector across orders $\boldsymbol{\Lambda} = (\Lambda_1, \ldots, \Lambda_J)$. The the Multi-MTD or MMTD(R, J) model for transition probabilities is then given by

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

where $\boldsymbol{\lambda}^{(j)}$ is a probability vector of length $\binom{R}{j}$. Inferences for $\boldsymbol{\Lambda}$ and corresponding $\boldsymbol{\lambda}^{(j)}$s can yield direct insight into lag importance.
We introduce two tractable sparsity priors for $\boldsymbol{\Lambda}$ and each $\boldsymbol{\lambda}^{(j)}$ to shrink down to a single (or a few) component(s), effectively performing model selection.

Priors for sparse probability vectors

Sparse Dirichlet Mixture (SDM)

The sparse Dirichlet mixture (SDM) prior model is a fixed-weight mixture of Dirichlet densities, each with a "boost" of equivalent sample size β in one of the categories. The density for a probability vector $\boldsymbol{\theta}$ is given as

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

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. For small sample sizes and relatively large β, the SDM can be characterized as a "winner-takes-all" prior.

Stick Breaking Mixture (SBM)

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

$$\theta_1 = V_1, \ \theta_k = V_k \prod_{j=1}^{k-1} (1 - V_j) \text{ for } k = 2, \ldots, K-1, \text{ and } \theta_K = 1 \cdot \prod_{j=1}^{K-1} (1 - V_j),$$

with V_k independently drawn from a mixture of two beta distributions, $V_k \overset{\text{ind.}}{\sim} \pi\text{Beta}(1, \eta) + (1 - \pi)\text{Beta}(\gamma, \delta)$. We encourage sparsity by setting η large, so that the first mixture component corresponds to small probabilities in $\boldsymbol{\theta}$.

Multinomial data illustration

With $K = 3$ categories, we can visualize the posterior density for $\boldsymbol{\theta}$ under each prior type as in Figure 1. Note the repelling effect of the proposed priors.

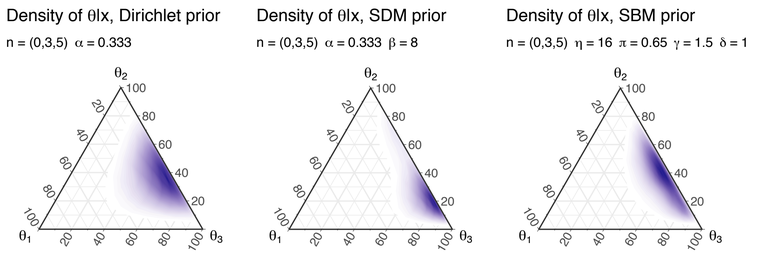

FIGURE 1: Kernel density estimated posterior densities for probability vector $\boldsymbol{\theta}$ under multinomial data and three prior models.

Application: pink salmon abundance

We illustrate with a time series of annual pink salmon abundance (escapement) in Alaska, U.S.A. from 1932 to 1961 shown in Figure 2 (Alaska Fisheries Science Center, 2018). Because pink salmon have a strict two-year life cycle, we expect even lags to have the most influence in predicting the current year's population.

FIGURE 2: Time series of the logarithm of annual abundance of pink salmon.

The observations were discretized into $K = 7$ bins by quantiles and two MMTD(7, 3) models were fit (via MCMC) using two prior specifications: Dirichlet for $\boldsymbol{\Lambda}$ and each $\boldsymbol{\lambda}^{(j)}$; and SDM priors for the same. Both sets of priors favor second order with the (2, 3) and (2, 4) lag combinations emerging as important. Shrinking down the over-specified MMTD with SDM priors yields more decisive inference for lag relevance, as seen in Figure 3.

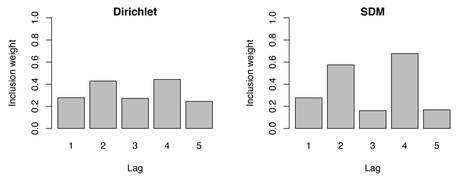

FIGURE 3: Posterior lag inclusion weights.

References

Alaska Fisheries Science Center (2018). AFSC/ABL: Pink salmon data collected at sashin creek weir 1934-2002. NOAA National Centers for Environmental Information, `https://inport.nmfs.noaa.gov/inport/item/17256`.

Berchtold, A. and Raftery, A. E. (2002). The mixture transition distribution model for high-order Markov chains and non-Gaussian time series. *Statistical Science*, pages 328–356.

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(325):194–206.

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

National Science Foundation under award DMS 1310438; San Francisco Bay Area Chapter of ASA Student Travel Award for JSM 2018.