Research statement

1 Introduction

Analysis of a categorical time series $X_1, \ldots, X_n$ is facilitated by a model that captures the statistical properties of the sequence, while being simple enough so that statistical analysis is feasible. Markovian structure, which assumes that conditioning on the last m observations is sufficient for computing the probability of the current one, supplies such a model for many categorical time series. A Markov chain is when $m = 1$, and independence is the special Markov chain with conditional probabilities not depending on the last state at all.

The goals of this research are improved modeling and efficiency of inference algorithms for categorical time series that have Markovian structure.

2 Sparse Markov models

Whereas higher-order Markov models (orders of dependence $m > 1$) provide a good approximation to probabilities associated with many categorical time series, the number of associated transition probability parameters is $|\Sigma|^m(|\Sigma| - 1)$, exponential in m, where $|\Sigma|$ is the number of possible values for each X_i. Therefore, low-order models are typically used in applications, even when higher-order dependence is called for. There is also a dearth in possible model choices, since the number of parameters must increase by the factor $|\Sigma|$.

Garcia and Gonzalez-Lopez (2010) introduced a model that they called *minimal Markov models*, but that go by other names, including *sparse Markov models* (SMM). An SMM of order m is a clustering of m-tuples histories $\mathcal{S} = \{\gamma_1, \gamma_2, \ldots, \gamma_\eta\}$, along with the conditional probability distributions for the η groups. Within groups γ_j, conditional probability distributions given each of the clustered m-tuples are the same. SMM are both flexible and parsimonious, allowing a better trade-off of bias that arises from having a model of an order that is too low, and variance from having many transition probabilities to estimate, when compared to general mth-order Markov models. As the model is relatively new, there is space for work relative to it. Two of my former students studied fitting SMM as part of their PhD work. Bennett et al. (2022) carried out SMM model fitting through Gibbs sampling; Majumder et al. (2024) used convex clustering of m-tuples through regularized regression. Martin (2020) gave an algorithm for obtaining pattern distributions in SMM. In those three papers, SMM were applied to modeling wind speeds and a DNA sequence, classifying viruses, and coverage of spaced seeds for DNA sequence alignment. Also, Martin and Bennett (2024) have developed hidden sparse Markov models, discussing the fitting of the model and some related inference algorithms. The paper is currently under revision.

Future work relative to SMM includes the development of an algorithm for fitting SMM through sequential hypothesis testing, the comparison of model fitting algorithms in terms of their asymptotic and finite-sample properties, prediction of future values for an SMM, and the development and application of sparse Markov decision processes.

3 Auxiliary Markov chains and the computation of pattern distributions

Much of my research has dealt with the computation of distributions of pattern statistics in Markovian data through an auxiliary Markov chain (AMC). The AMC maps data strings into its states, achieving data reduction, and the Markov chain is then used to obtain the desired distribution. The method was forwarded as far back as Brookner (1966), and was popularized in Fu and Koutras (1994).

Since then, improvements in computational efficiency have been attained for complicated patterns in biosequences. For example, deterministic finite automata (DFA) minimization has been used (see, e.g., Lladser et al. 2008) to turn an AMC into a smaller version that still facilitates exact computation. Yet setting up a large state space before minimizing it can still be prohibitive. Nuel (2008) got around that problem through non-deterministic DFA, and, in novel work, Martin (2019) characterized equivalent states so that extraneous ones can be identified and deleted during the process of sequentially forming an AMC state space.

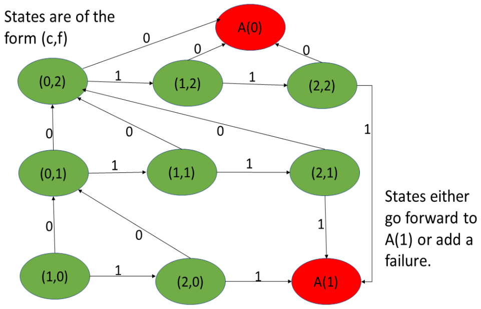
To highlight the effectiveness of these breakthroughs, methods of the latter two papers allowed the computation of distributions related to occurrences of structured motifs in DNA of the form $w_1\Sigma^{16:18}w_2$, where w_1 and w_2 are strings from $\Sigma = \{A, C, G, T\}$, and $\Sigma^{16:18}$ denotes possible gaps of lengths 16 to 18 between w_1 and w_2, gap lengths that are realized in practice. The number of patterns under search is $4^{16} + 4^{17} + 4^{18} \approx 90$ billion, so that exact distributions for occurrences of such structured motifs had not been computed previously in the literature. However, exact distributions were obtained in Martin (2019) for 15 different sets of w_1 and w_2 in 0.22s to 1.46s. This illustrates that the developed methodology has made exact computation accessible for important pattern structures.

A problem that has come into my purview for future work is as follows. In some studies, the researcher desires pattern distributions for many input probabilities. Using AMC, different input probabilities mean that a new Markov chain needs to be set up. Mak and Benson (2009) gave a way around this dilemma, noting that independent binary trials with the same number of successes have the same probability. They then counted sequences in ”probability equivalence classes” and with desired properties relative to patterns. The advantage is that the counts only need be obtained once, and can then be used in an equation at the end by inserting the many input probabilities to obtain the desired distributions. The trade-off for obtaining counts is the extra storage needed in terms of sufficient statistics. My plan is to extend the use of counts to higher-order Markovian sequences and various types of patterns. However, as illustrated by the following problem, flexibility is needed.

Suppose that a lawnmower is tested to determine its start-up reliability. The unit is accepted by a customer if k consecutive successes occur before d total failures, and rejected otherwise. The maximum number of trials in such tests is $n = kd$. Transitions of an AMC for this problem with $k = 3$ and $d = 3$ are shown in Figure 1. States are of the form (c, f), with c denoting consecutive successes, and f total failures. Two absorbing states, A(0) and A(1), symbolizing rejection and acceptance of the unit, are added. The computation is initialized at time 1 in either state (0,1) or (1,0). We seek the probability in A(1) after time n.

A simple recursion can be set up to solve this problem. Let $\Phi(h)$ denote the conditional probability of acceptance of the unit from state $(0, h)$, $h = d - 1, d - 2, \ldots, 1$, or state (1,0) for $h = 0$. Then $\Phi(d-1)$ can be obtained directly as k consecutive successes must occur for acceptance, and $\Phi(h-1)$ may be obtained in terms of $\Phi(h)$ since the system either transitions directly to $A(1)$, or an hth failure occurs. The many input probabilities may be entered into the d recursive equations at the end, and the unconditional probability of $A(1)$ is obtained by conditioning on the state at time 1 and using the ”Law of Total Probability.” Thus while my plan is to compare Benson's counting method to using an AMC in terms of computational efficiency, it is also clear that the best solution may not fit either the AMC or counting paradigms that have been presented.

Figure 1: Auxiliary Markov chain transitions for $k = 3$, $d = 3$



States are of the form (c,f)

States either go forward to A(1) or add a failure.

References

1. Bennett, I., Martin, D.E.K. and Lahiri, S. (2023). ”Fitting sparse Markov models through a collapsed Gibbs sampler,” *Computational Statistics*, 38, 1977-1994.

2. Brookner, E. (1966), “Recurrent events in a Markov chain,” *Information and Control*, 9, 215-229.

3. Fu, J.C., and Koutras, M.V. (1994), “Distribution theory of runs: a Markov chain approach,” *Journal of the American Statistical Association*, 89, 1050-1058.

4. García, J.E. and González-López, V.A. (2010). Minimal Markov models. arXiv:1002.0729.

5. Lladser, M.E., Betterton, M.D., and Knight, R. (2008), “Multiple pattern matching: a Markov chain approach,” Journal of Mathematical Biology, 56, 51-92.

6. Mak, D.Y.F. and Benson, G. (2009), “All hits, all the time: parameter-free computation of spaced seed sensitivity,” *Bioinformatics*, 25(3), 302-308.

7. Majumder, T., Lahiri, S., and Martin, D.E.K. Fitting sparse Markov models to categorical time series using regularization (under review).

8. Martin, D.E.K. (2019), “Minimal auxiliary Markov chains through sequential elimination of states,” *Communication in Statistics, Simulation and Computation*, 48(4), 1040-1054, DOI: 10.1080/03610918.2017.1406505.

9. Nuel, G. (2008), “Pattern Markov chains: Optimal Markov chain embedding through deterministic finite automata,” *Journal of Applied Probability*, 45(1), 226-243.