Inference and applications of sparse Markov models

Donald E.K. Martin *

August 31, 2022

Analysis of a categorical time series is facilitated by a model that captures the statistical properties of the sequence, while being simple enough so that statistical analysis is feasible. Markov models of an order of dependence m serve this purpose. However, the number of parameters in a Markov model is exponential in m, leading to variance considerations because many parameters need to be estimated from data. Thus first-order Markov chains ($m = 1$) are more frequently used. Yet using a lower-order model when higher-order dependence is called for leads to bias. Sparse Markov models help with this bias-variance trade-off. A sparse Markov model (SMM) is a higher-order Markov model for which conditioning m-tuple histories are grouped into classes such that the conditional probability distribution is constant over m-tuples in the same class. The clustering reduces the number of parameters.

In this research project, two algorithms have been derived for SMM fitting based on data, one a Bayesian method using Dirichlet priors for conjugacy, the other based on regularized regression with a penalty parameter determined in a data-driven manner. Also developed was a new model, the hidden sparse Markov model (HSMM), where latent states follow an SMM and generate observed data. Algorithms were derived for prediction of SMM data, computing distributions of pattern statistics in SMM and states of an HSMM, and conditional inference algorithms for HSMM including forward and backward algorithms and an algorithm for obtaining the most likely state sequence. Some progress was made on developing Central Limit Theorems for SMM. The methodology has been applied to such diverse problems as categorizing viruses, and analyses of fluctuations in S&P 500 data.

*North Carolina State University, Department of Statistics, 4272 SAS Hall, 2311 Stinson Drive, Raleigh, NC 27695-8203

Foreseen for the future are (i) comparisons of the two algorithms for model fitting mentioned above to determine when each should be used; (ii) quantification through simulated and real data of improvements in model fits and inference afforded by SMM and HSMM when compared to other models; (iii) the application of SMM and HSMM to the analysis of various categorical time series; (iv) the examination of the feasibility and application of sparse Markov decision processes, an extension of SMM to allow actions and rewards.