Novel Statistical Frameworks for Analysis of Structured Sequential Data

Abstract:
We are developing a broad array of novel statistical frameworks for analyzing complex sequential data sets. Our research is primarily motivated by a collaboration with neuroscientists trying to understand the neurological, genetic and evolutionary basis of human communication using bird and mouse models. The data sets comprise structured sequences of syllables or 'songs' produced by animals from different genotypes under different experimental conditions. Simple first order Markov chains are insufficiently flexible to learn complex serial dependency structures and systematic patterns in the vocalizations, an important goal in these studies. To this end, we have developed a sophisticated nonparametric Bayesian approach to higher order Markov chains building on probabilistic tensor factorization techniques. Our proposed method is of very broad utility, with applications not limited to analysis of animal vocalizations, and provides new insights into the serial dependency structures of many previously analyzed sequential data sets arising from diverse application areas. Our method has appealing theoretical properties and practical advantages, and achieves substantial gains in performance compared to previously existing methods. Our research also paves the way to advanced automated methods for more sophisticated dynamical systems, including higher order hidden Markov models that can accommodate more general data types.