1 Introduction

Hidden Markov models (HMMs) are widely applied for the analysis of time series data with incomplete or noisy observations together with stochastic system dynamics (Capp, Moulines, and Rydn, 2006; Elliott, Aggoun, and Moore, 2008). HMMs are used in a diverse range of application domains, with recent attention in areas of speech recognition and natural language processing (Gales and Young, 2008). See MacDonald and Zucchini (1997) for a broad review of HMM applications in disciplines such as as medicine, finance, sociology, and climatology.

For a single discrete HMM, likelihood calculation involves summing over the distribution of a sequence of unknown latent states. This can be implemented either using standard direct filtering summations (*e.g.*, Elliott, Aggoun, and Moore, 2008, chapter 2) as part of either maximum likelihood or Bayesian analysis, or using Markov chain Monte Carlo (MCMC; Gilks, 2005; Brooks et al., 2011) for Bayesian analysis. In the case of MCMC, the unknown state variables are included in MCMC sampling. However, it is often the case that one or more HMMs are embedded in a larger hierarchical model, perhaps accounting for explanatory variables of state transition probabilities or shared variation among multiple time series. In such cases practitioners may rely on MCMC to perform a Bayesian analysis, but they face a quandary of computational efficiency. If they use standard MCMC software, they often have no choice to but to include the unknown latent state variables in MCMC sampling. For large models this can contribute hundreds or thousands of dimensions which require MCMC sampling, to the point of rendering this approach computationally impractical.

In theory there are computational tradeoffs between using MCMC and direct filtering summation when embedding HMMs in a larger hierarchical model, but these tradeoffs have not been explored to date. Here we do so, by considering combinations of several existing computational methods for fitting HMMs. These methods include direct filtering to remove latent variables, using a reduced representation of observational data, and dynamic blocking of model parameters to achieve efficient MCMC sampling. We demonstrate that for large