

Learning Bayesian networks from postgenomic data with an improved structure MCMC sampling scheme

Marco Grzegorzcyk and Dirk Husmeier
BIOSS

February 22, 2008

1 Abstract

Our paper contributes to recent research on sampling Bayesian network structures from the posterior distribution with MCMC. Two principled paradigms have been applied in the past. Structure MCMC, first proposed by Madigan and York, defines a Markov chain in the space of graph structures by applying basic operations to individual edges of the graph, like the creation, deletion or reversal of an edge. Alternatively, order MCMC, proposed by Friedman and Koller, defines a Markov chain in the space of node orders. While the second approach has been found to substantially improve the mixing and convergence of the Markov chain, it does not allow an explicit specification of the prior distribution over graph structures or, to phrase this differently, it incurs a distortion of the specified prior distribution as a consequence of the marginalization over node orders. This distortion can lead to problems for applications in systems biology, where owing to the limited number of experimental conditions the integration of biological prior knowledge into the inference scheme becomes desirable. Different approaches and modifications have been developed in the literature to address this shortcoming (e.g. by Ellis, Eaton and Murphy). Unfortunately, these methods incur extra computational costs and are not practically viable for inferring large networks with more than 20 to 30 nodes. There have been suggestions of how to improve the classical structure MCMC approach by using the concept of the inclusion boundary, as proposed by Castelo and Kocka, but these methods only partially address the convergence and mixing problems. In the present paper we propose a novel structure MCMC scheme, which augments the classical structure MCMC method of Madigan and York with a novel edge reversal move. The idea of the new move is to resample the parent sets of the two nodes involved in such a way that the selected edge is reversed subject to the acyclicity constraint. The proposal of the new parent sets is done effectively by adopting ideas from importance sampling; in this way faster convergence is effected. For methodological consistency, and in contrast to inclusion-driven MCMC, we have properly derived the Hastings factor, which is a function of various partition functions that are straightforward to compute. The resulting Markov chain is reversible, satisfies the condition of detailed balance, and is hence guaranteed to theoretically converge to the desired posterior distribution. For our empirical evaluation, we have tested our method on various data sets from the UCI repository, such as Vote, Flare, Boston Housing, and Alarm, which have previously been used by Friedman and Koller to demonstrate that order MCMC outperforms structure MCMC. Our experimental results show that integrating the novel edge reversal move yields a substantial improvement of the resulting MCMC sampler over classical structure MCMC, with convergence and mixing properties that are similar to those of order MCMC. To demonstrate the avoidance of the distortional effect incurred with order MCMC, we have extended our empirical evaluation by analysing flow cytometry protein concentrations from the Raf-Mek-Erk signalling pathway. The experimental results show that the novel MCMC scheme can lead to a slight yet significant performance improvement over order MCMC when explicit prior knowledge is integrated into the learning scheme. This suggests that the avoidance of any systematic distortion of the prior probability distribution on network structures renders our improved structure MCMC sampler

preferable to order MCMC, especially for those contemporary systems biology applications where the number of experimental conditions relative to the complexity of the investigated system, and hence the weight of the likelihood, is relatively low, and explicit prior knowledge about network structures from publicly accessible data bases is included.