Bayesian model determination in complex systems

Mohammadi, Abdolreza

Chapter 1

Introduction

1.1 Motivation

One of the main challenges in modern science is to model the complex systems. The difficulty of modeling complex systems lies partly in their topology and how they form rather complex networks. For example, in neuroscience we are interested to understand how various regions of the brain interact with one another. In genetics, the cell is a network of chemicals linked by chemical reactions and we are interested to model this complex network. From this perspective, our interest to modeling networks is part of a broader current of research on complex systems. The main contribution of this thesis is to develop Bayesian statistical methods that jointly model the underlying network (or graph) and its structure among variables in the system.

Graphical models provide potential tools to model and make statistical inference regarding complex relationships among variables. The key feature of graphical models is the close relationship between their probabilistic properties and the topology of the underlying graphs represents, as it allows an intuitive understanding of complex systems.

1.2 Bayesian model determination in graphical models

In graphical models, nodes represent variables and edges represent pairwise dependencies, with the edge set defining the global conditional independence structure of the distribution. The methodological issues faced, as the dimension grows, include questions of the nature and consistency of prior specification (priors over graph space, and parameters on any single, specified graph). Then the challenging problem is searching over the space of

graphs to identify subsets of interest under the theoretically implied posterior distributions. This represents a complex model selection problem.

The analysis challenge is inherently one of model uncertainty and model selection: we are interested in exploring graph space and identifying graphs that are most appropriate for a given dataset. Therefore, inference on variable dependencies and prediction is then based on parametric inferences within a set of selected graphs.

In Bayesian paradigm, we are interested in exploring graph space and identifying graphs that are most appropriate for a given data. In this regard, we need to calculate the posterior distribution of the graph G conditional on data

$$Pr(G|\text{data}) = \frac{Pr(G)Pr(\text{data}|G)}{\sum_{G\in\mathcal{G}} Pr(G)Pr(\text{data}|G)}, \tag{1.1}$$

in which $\mathcal{G}$ is a graph space. Computing this posterior distribution is computationally unfeasible, since in the denominator we require the sum over all possible graph space. The graph space supper-exponentially increases according to the dimension of variables. p nodes in a graph mean $p(p-1)/2$ possible edge, and hence we have $2^{p(p-1)/2}$ different possible graphs corresponding to all combinations of individual edges being in or out of the graph. For example, for the graph with only 10 variables, we have more than 35×10^{12} possible different graphical models.

This motivates us to develop effective search algorithms for exploring graphical model uncertainty. In order to be accurate and scalable, the main key is to design search algorithms which are able to quickly move towards high posterior probability region, and also to take advantage of local computation. One solution is the trans-dimensional MCMC methodology (4).

1.3 Trans-dimensional Markov chain Monte Carlo

This subsection is more narrowly focused on Bayesian model selection via Markov chain Monte Carlo (MCMC) methods for what can be called *trans-dimensional* problems; those where the dynamic variable of the simulation, the *unknowns* in the Bayesian set-up, does not have fixed dimension. One conceptually elegant method is reversible jump Markov chain Monte Carlo (RJMCMC), which was proposed by (3), also termed trans-dimensional MCMC, in which the model itself is conceived as another unknown and the MCMC algorithm is enlarged to allow 'jumps' between all possible models. A prior is required over the model space, but with judicious selection of jumps the number of models does not need to

be specified in advance and each model does not require separate estimation. These moves then require (reversible) bridges to be built between parameters of models in different dimensions. The posterior probability of a model is then estimated by the proportion of times that the particular model is accepted in the MCMC run. This method has been employed and discussed for model selection in many contexts. Specially, (2) used this approach for model selection in decomposable undirected Gaussian graphical models. More recently, (1, 6, 5) used this method for model selection in more general case, non-decomposable graphical models.

An alternative trans-dimensional MCMC approach is the birth-death MCMC (BDMCMC) algorithm, which is based on a continuous time Markov birth-death process. In this method, the time between jumps to a larger dimension (birth) or a smaller one (death) is taken to be a random variable with a specific rate. The choice of birth and death rates determines the birth-death process and is made in such a way that the stationary distribution is precisely the posterior distribution of interest. Contrary to the RJMCMC approach, moves between models are always accepted, which makes the BDMCMC approach extremely efficient. In the context of finite mixture distributions with variable dimension, this method has been used (15).

1.4 Outline of thesis contribution

In this thesis we consider the problem of Bayesian inference in the following statistical models: graphical models (Chapters 2, 3, 4), exponential random graph models (Chapter 5) and queuing systems (Chapter 6).

In **Chapter 2** we introduce a novel and efficient Bayesian framework for Gaussian graphical model determination. We cover the theory and computational details of the proposed method. We carry out the posterior inference by using an efficient sampling scheme which is a trans-dimensional MCMC approach based on birth-death process. It is easy to implement and computationally feasible for high-dimensional graphs. We show our method outperforms alternative Bayesian approaches in terms of convergence and computing time. Unlike frequentist approaches, it gives a principled and, in practice, sensible approach to structure learning. We apply the method to large-scale real applications from human and mammary gland gene expression studies to show its empirical usefulness. The result of this chapter is published in (13).

The method that we propose in Chapter 2 is limited only to the data that follows the Gaussianity assumption. In **Chapter 3** we propose a Bayesian approach for graphical

model determination based on a Gaussian copula approach that can deal with continuous, discrete, or mixed data. We embed a graph selection procedure inside a semi-parametric Gaussian copula. We carry out the posterior inference by using an efficient sampling scheme which is a trans-dimensional MCMC approach based on the birth-death process. We implement our approach to discovering potential risk factors related to Dupuytren disease. The contents of this chapter corresponded to the manuscripts (7, 8) and the paper (11).

In **Chapter 4** we introduce an R package BDgraph (12) which contains functions to perform Bayesian structure learning in high-dimensional graphical models with either continuous or discrete variables. This package efficiently performs the Bayesian approaches that proposed in Chapters 2 and 3. The core of the BDgraph package efficiently implemented in C++ to maximize computational speed. The contents of this chapter corresponded to the manuscript (14).

In **Chapter 5** we introduce a comprehensive Bayesian graphical modeling for new features of exponential random graph models (ERGM). The method increases the range and applicability of the ERGM as a potential tool for the statistical inference in network structure learning.

In **Chapter 6** we introduce a Bayesian framework in an M/G/1 queuing system with an optional second service. The semi-parametric model based on a finite mixture of Gamma distributions is considered to approximate both the general service and re-service times densities in this queuing system. We estimate system parameters, predictive densities and some performance measures related to this queuing system such as stationary system size and waiting time. The result of this chapter is published in (9, 10).

References

[1] Dobra, A., Lenkoski, A., and Rodriguez, A. (2011). Bayesian inference for general gaussian graphical models with application to multivariate lattice data. *Journal of the American Statistical Association*, 106(496):1418–1433.

[2] Giudici, P. and Green, P. (1999). Decomposable graphical gaussian model determination. *Biometrika*, 86(4):785–801.

[3] Green, P. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732.

[4] Green, P. J. (2003). Trans-dimensional markov chain monte carlo. *Oxford Statistical Science Series*, pages 179–198.

[5] Lenkoski, A. (2013). A direct sampler for g-wishart variates. *Stat*, 2(1):119–128.

[6] Lenkoski, A. and Dobra, A. (2011). Computational aspects related to inference in gaussian graphical models with the g-wishart prior. *Journal of Computational and Graphical Statistics*, 20(1):140–157.

[7] Mohammadi, A., Abegaz Yazew, F., van den Heuvel, E., and Wit, E. C. (2015). Bayesian modeling of dupuytren disease using gaussian copula graphical models. *Arxiv preprint arXiv:1501.04849v2*.

[8] Mohammadi, A., Abegaz Yazew, F., and Wit, E. C. (2014). Bayesian copula gaussian graphical modelling. *(IWSM'14) Proceedings of the 29th International Workshop on Statistical Modelling*, 1:225–230.

[9] Mohammadi, A., Salehi-Rad, M., and Wit, E. (2013). Using mixture of gamma distributions for bayesian analysis in an m/g/1 queue with optional second service. *Computational Statistics*, 28(2):683–700.

[10] Mohammadi, A. and Salehi-Rad, M. R. (2012). Bayesian inference and prediction in an m/g/1 with optional second service. *Communications in Statistics-Simulation and Computation*, 41(3):419–435.

[11] Mohammadi, A. and Wit, E. (2014). Contributed discussion on article by finegold and drton. *Bayesian Analysis*, 9(3):577–579.

[12] Mohammadi, A. and Wit, E. (2015a). *BDgraph: Graph Estimation Based on Birth-Death MCMC Approach*. R package version 2.17.

[13] Mohammadi, A. and Wit, E. C. (2015b). Bayesian structure learning in sparse gaussian graphical models. *Bayesian Analysis*, 10(1):109–138.

[14] Mohammadi, A. and Wit, E. C. (2015c). Bdgraph: Bayesian structure learning of graphs in r. *arXiv preprint arXiv:1501.05108v2*.

[15] Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods. *Annals of Statistics*, 28(1):40–74.