Latent Graphical Model Selection: Efficient Methods for Locally Tree-like Graphs

Animashree Anandkumar
UC Irvine
a.anandkumar@uci.edu

Ragupathyraj Valluvan
UC Irvine
rvalluva@uci.edu

Abstract

Graphical model selection refers to the problem of estimating the unknown graph structure given observations at the nodes in the model. We consider a challenging instance of this problem when some of the nodes are latent or hidden. We characterize conditions for tractable graph estimation and develop efficient methods with provable guarantees. We consider the class of Ising models Markov on locally tree-like graphs, which are in the regime of correlation decay. We propose an efficient method for graph estimation, and establish its structural consistency when the number of samples n scales as $n = \Omega(\theta_{\min}^{-\delta\eta(\eta+1)-2} \log p)$, where $\theta_{\min}$ is the minimum edge potential, δ is the depth (i.e., distance from a hidden node to the nearest observed nodes), and η is a parameter which depends on the minimum and maximum node and edge potentials in the Ising model. The proposed method is practical to implement and provides flexibility to control the number of latent variables and the cycle lengths in the output graph. We also present necessary conditions for graph estimation by any method and show that our method nearly matches the lower bound on sample requirements.

Keywords: Graphical model selection, latent variables, quartet methods, locally tree-like graphs.

1 Introduction

It is widely recognized that the process of fitting observed data to a statistical model needs to incorporate latent or hidden factors, which are not directly observed. Learning latent variable models involves mainly two tasks: discovering structural relationships among the observed and hidden variables, and estimating the strength of such relationships. One of the simplest models is the *latent class model* (LCM), which incorporates a single hidden variable and the observed variables are conditionally independent given the hidden variable. Latent tree models extend this model class to incorporate many hidden variables in a hierarchical fashion. Latent trees have been effective in modeling data in a variety of domains, such as phylogenetics [1]. Their computational tractability: upon learning the latent tree model, enables the inference to be carried out efficiently through *belief propagation*. There has been extensive work on learning latent trees, including some of the recent works, e.g. [2–4], demonstrate efficient learning in high dimensions. However, despite the advantages, the assumption of an underlying tree structure may be too restrictive. For instance, consider the example of topic-word models, where topics (which are hidden) are discovered using information about word co-occurrences. In this case, a latent tree model does not accurately represent the hierarchy of topics and words, since there are many common words across different topics. Here, we relax the latent tree assumption to incorporate cycles in the latent graphical model while retaining many advantages of latent tree models, including tractable learning and inference. Relaxing the tree constraint leads to many challenges: in general, learning these models is NP-hard, even when there are no latent variables, and developing tractable methods for such models is itself an area of active research, e.g. [5–7]. We consider structure estimation in latent graphical models Markov on locally