The Annals of Statistics
2013, Vol. 41, No. 2, 401–435
DOI: 10.1214/12-AOS1070

LEARNING LOOPY GRAPHICAL MODELS WITH LATENT VARIABLES: EFFICIENT METHODS AND GUARANTEES

BY ANIMASHREE ANANDKUMAR[1] AND RAGUPATHYRAJ VALLUVAN[2]

University of California, Irvine

The problem of structure estimation in graphical models with latent variables is considered. We characterize conditions for tractable graph estimation and develop efficient methods with provable guarantees. We consider models where the underlying Markov graph is locally tree-like, and the model is in the regime of correlation decay. For the special case of the Ising model, the number of samples n required for structural consistency of our method scales as $n = \Omega(\theta_{\min}^{-\delta\eta(\eta+1)-2} \log p)$, where p is the number of variables, $\theta_{\min}$ 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 bounds on node and edge potentials in the Ising model. Necessary conditions for structural consistency under any algorithm are derived and our method nearly matches the lower bound on sample requirements. Further, 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.

1. Introduction.

Learning latent variable models from observed samples involves mainly two tasks: discovering relationships between the observed and hidden variables, and estimating the strength of such relationships. One of the simplest latent variable models is the so-called *latent class model* or *näive Bayes model*, where the observed variables are conditionally independent given the state of the latent factor. An extension of these models are *latent tree models* with many hidden variables forming a tree hierarchy. Latent tree models have been effective in modeling data in a variety of domains, such as the evolutionary process which gave rise to the present-day species in bio-informatics (popularly known as *phylogenetics*) [21, 43], for financial and topic modeling [17] and for modeling contextual information for object recognition in computer vision [16]. Prior works on learning latent tree models (e.g., [17, 23, 35]), demonstrate that latent tree models can be learned efficiently in high dimensions. In other words, the number of samples required for consistent learning is much smaller than the number of variables at hand. Moreover, inference in latent tree models is computationally tractable by means of simple algorithms such as *belief propagation*.

Received March 2012; revised October 2012.

[1]Supported in part by NSF Award CCF-1219234, AFOSR Award FA9550-10-1-0310 and ARO Award W911NF-12-1-0404.

[2]Supported by the ONR Award N00014-08-1-1015.

MSC2010 subject classifications. Primary 62H12; secondary 05C12.

Key words and phrases. Graphical model selection, latent variables, quartet methods.