Journal of Multivariate Analysis 136 (2015) 147–162

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Bayesian structure learning in graphical models

Sayantan Banerjee [a,*], Subhashis Ghosal [b]

[a] The University of Texas MD Anderson Cancer Center, United States

[b] North Carolina State University, United States

ARTICLE INFO

Article history:
Received 7 April 2014
Available online 23 January 2015

AMS subject classifications:
primary 62H12
secondary 62F12
62F15

Keywords:
Graphical lasso
Graphical models
Laplace approximation
Posterior convergence
Precision matrix

ABSTRACT

We consider the problem of estimating a sparse precision matrix of a multivariate Gaussian distribution, where the dimension p may be large. Gaussian graphical models provide an important tool in describing conditional independence through presence or absence of edges in the underlying graph. A popular non-Bayesian method of estimating a graphical structure is given by the graphical lasso. In this paper, we consider a Bayesian approach to the problem. We use priors which put a mixture of a point mass at zero and certain absolutely continuous distribution on off-diagonal elements of the precision matrix. Hence the resulting posterior distribution can be used for graphical structure learning. The posterior convergence rate of the precision matrix is obtained and is shown to match the oracle rate. The posterior distribution on the model space is extremely cumbersome to compute using the commonly used reversible jump Markov chain Monte Carlo methods. However, the posterior mode in each graph can be easily identified as the graphical lasso restricted to each model. We propose a fast computational method for approximating the posterior probabilities of various graphs using the Laplace approximation approach by expanding the posterior density around the posterior mode. We also provide estimates of the accuracy in the approximation.

1. Introduction

Statistical inference on a large covariance or precision matrix (inverse of covariance matrix) is a topic of growing interest in recent times. Often the dimension p grows with the sample size n and even p can exceed n. Data of this type are frequently encountered in fMRI, spectroscopy, gene array expressions and so on. Estimation of a covariance or precision matrix is of special interest because of its importance in methods like principal component analysis (PCA), linear discriminant analysis (LDA), etc. In cases where $p > n$, the sample covariance matrix is necessarily singular, and hence an estimator of the precision matrix cannot be obtained by inverting it. Therefore we need to resort to other techniques for handling the high-dimensional problems.

Regularization methods for estimation of the covariance or precision matrix have been proposed and studied in recent literature for high-dimensional problems. These include banding, thresholding, tapering and penalization based methods; for example, see Ledoit and Wolf [22], Huang et al. [18], Yuan and Lin [35], Bickel and Levina [4,5], Karoui [19], Friedman et al. [13], Rothman et al. [29], Lam and Fan [20], Rothman et al. [30], Cai et al. [8,7]; see also Banerjee and Ghosal [3] for a Bayesian method based on banding. The primary goal of these regularization based methods is to impose a sparsity structure in the

* Correspondence to: Department of Biostatistics, The University of Texas MD Anderson Cancer Center, 1400 Pressler Street, Houston, TX 77030, United States.

E-mail address: SBanerjee@mdanderson.org (S. Banerjee).

http://dx.doi.org/10.1016/j.jmva.2015.01.015

matrix. Most of these methods are applicable to situations where there is a natural ordering in the underlying variables, for example in data from time series, spatial data, etc., so that variables which are far off from each other have smaller correlations or partial correlations. In high-dimensional situations for data arising from genetics or econometrics, a natural ordering of the underlying variables may not always be readily available and hence estimation methods which are invariant to the ordering of the variables are desirable.

For estimation of a sparse inverse covariance matrix, graphical models [21] provide an excellent tool, as the conditional dependence between the component variables is captured by an undirected graph; see Dobra et al. [12], Meinshausen and Bühlmann [25], Yuan and Lin [35], Friedman et al. [13]. There are several methods in the frequentist literature for the estimation of the precision matrix through graphical models. These methods include minimization of the penalized log-likelihood of the data with a lasso type penalty on the elements of the precision matrix. Several algorithms have been developed in the literature to solve the above optimization problem, including coordinate descent based algorithm for the lasso, which is popularly known as the graphical lasso [25,13,2,35,16,33]. Other methods include the Sparse Permutation Invariant Covariance Estimator (SPICE) [29].

Frequentist behavior of Bayesian methods in the context of high dimensional covariance matrix estimation have been studied only by a few authors. Ghosal [14] studied asymptotic normality of posterior distributions for exponential families, which include the normal model with unknown covariance matrix, when the dimension $p \to \infty$, but restricting to $p \ll n$. Recently, Pati et al. [27] considered sparse Bayesian factor models for dimensionality reduction in high dimensional problems and showed consistency in the L_2-operator norm (also known as the spectral norm) by using a point mass mixture prior on the factor loadings, assuming such a factor model representation for the true covariance matrix.

Bayesian methods for inference using graphical models have also been developed, as in Roverato [31], Atay-Kayis and Massam [1], Letac and Massam [23]. A conjugate family of priors, known as the G-Wishart prior [31] have been developed for incomplete decomposable graphs. The equivalent prior on the covariance matrix is termed as the hyper inverse Wishart distribution in Dawid and Lauritzen [11]. Letac and Massam [23] introduced a more general family of conjugate priors for the precision matrix, known as the W_{P_G}-Wishart family of distributions, which also has the conjugacy property. The properties of this family of distributions, including expressions for the Bayes estimators were further explored in Rajaratnam et al. [28]. Recently Banerjee and Ghosal [3] studied posterior convergence rates for a G-Wishart prior inducing a banding structure, where the true precision matrix need not have the banding structure.

Wang [32] developed a Bayesian version of the graphical lasso, by putting Laplace priors on the off-diagonal elements of the precision matrix and exponential priors on the diagonals. Similar in lines with the Bayesian lasso [26], the posterior mode in this case coincides with the graphical lasso estimate. A block Gibbs sampler is also developed for sampling from the resulting posterior. However, the Bayesian graphical lasso does not introduce any sparsity in the graphical structure because of the absence of a point mass at zero in the prior distribution for the off-diagonal elements. On the other hand, if point masses are introduced, the resulting posterior distribution on the structure of the graph becomes extremely difficult to compute based on the traditional reversible jump Markov chain Monte Carlo method.

In this paper, we derive posterior convergence rates for the Bayesian graphical lasso prior in terms of the Frobenius norm under appropriate sparsity conditions when the dimension p grows with the sample size n. For computing the posterior distribution, we propose a Laplace approximation based method to compute the posterior probability of different graphical structures. Such Laplace approximations based methods have been developed for variable selection in regression models; for example, see Yuan and Lin [34], Curtis et al. [10]. The lasso type penalty on the elements lead to non-differentiability of the integrand, when the graphical lasso sets an off-diagonal entry to zero, but the model includes that off-diagonal entry as a free variable. We shall call such models non-regular following the terminology used by Yuan and Lin [34] for variable selection in linear regression models. We show that the posterior probability of non-regular models are substantially smaller than their regular counterparts and hence in comparison may be ignored from consideration. We also estimate the error in the Laplace approximation for regular models.

The paper is organized as follows. In the next section, we introduce notations and discuss preliminaries on graphical models required for the other sections of the paper. In Section 3, we state model assumptions and specify the prior distribution on the underlying parameters, derive the form of the posterior and obtain the posterior convergence rate using the general theory developed in Ghosal et al. [15]. In Section 4, we develop the approximation of the posterior probabilities for different graphical models and discuss the issue of non-regular graphical models. We also show that the error in approximation of the posterior probabilities using the Laplace approximation is asymptotically negligible under appropriate conditions. A simulation study is performed in the Section 5 followed by a real data example in Section 6. Proofs of main results and additional lemmas are included in the Appendix.

2. Notations and preliminaries

An undirected graph G comprises of a non-empty set V of p vertices indexing the components of a p-dimensional random vector along with an edge set $E \subset \{(i,j) \in V \times V : i < j\}$. Let $\boldsymbol{X} = (X_1, \dots, X_p)^T$ be distributed as $\mathrm{N}_p(\mathbf{0}, \boldsymbol{\Omega}^{-1})$, where the precision matrix $\boldsymbol{\Omega} = ((\omega_{ij}))$ is such that $(i,j) \notin E$ implies $\omega_{ij} = 0$. We then say that $\boldsymbol{X}$ follows a Gaussian graphical model (GGM) with respect to the graph G. Since the absence of an edge between i and j implies conditional independence of X_i and X_j given $(X_r : r \neq i, j)$, a GGM serves as an excellent tool in representing the sparsity structure in the precision matrix. Following the notation in Letac and Massam [23], the canonical parameter $\boldsymbol{\Omega}$ is restricted to $\mathscr{P}_G$, where $\mathscr{P}_G$ is the cone of