ABSTRACT

BANERJEE, SAYANTAN. Bayesian Inference for High Dimensional Models: Convergence Properties and Computational Issues. (Under the direction of Subhashis Ghoshal.)

This dissertation focuses on Bayesian inference for high-dimensional models, including estimation of the mean in different regression models, and estimation of precision matrices for high dimensional random variables. Along with studying theoretical properties of posterior distributions, we also develop computational methods for efficient and fast model assessment.

In Chapter 2, we consider a fast Bayesian variable selection method for generalized additive partial linear models. The functions in the non-parametric additive part of the model are expanded in a B-spline basis and multivariate Laplace prior put on the coefficients with point mass at zero. The coefficients corresponding to the strictly linear components are assigned a univariate Laplace prior with point mass at zero. The prior times the likelihood is mathematically intractable but we find an approximation by expansion around the posterior mode, which is the group lasso solution in generalized linear model setting for the choice of prior. We thus completely avoid Markov Chain Monte Carlo (MCMC) or any other time consuming sampling based methods, hence leading to quick assessment of various posterior model probabilities. This technique is applied to the high-dimensional situation where the number of parameters may exceed the number of observations. We evaluate the performance of the Bayesian method by conducting simulation studies and real data analyses.

In Chapter 3, we consider Bayesian estimation of a $p \times p$ precision matrix, when p can be much larger than the available sample size n. It is well known that consistent estimation in such ultra-high dimensional situations requires regularization such as banding, tapering or thresholding. We consider a banding structure in the model and induce a prior distribution on a banded precision matrix through a Gaussian graphical model, where an edge is present

only when two vertices are within a given distance. For a proper choice of the order of graph, we obtain the convergence rate of the posterior distribution and Bayes estimators based on the graphical model in the L_∞-operator norm uniformly over a class of precision matrices, even if the true precision matrix may not have a banded structure. Along the way to the proof, we also compute the convergence rate of the maximum likelihood estimator (MLE) under the same set of conditions, which is of independent interest. The graphical model based MLE and Bayes estimators are automatically positive definite, which is a desirable property not possessed by some other estimators in the literature. We also conduct a simulation study to compare finite sample performance of the Bayes estimators and the MLE based on the graphical model with that obtained by using a Cholesky decomposition of the precision matrix. Finally, we discuss a practical method of choosing the order of the graphical model using the marginal likelihood function.

In Chapter 4, we consider a similar problem of estimating a sparse precision matrix of a multivariate Gaussian distribution, including the case where the dimension p exceeds the sample size n, but now without the assumption of a banding structure in the model. A popular non-Bayesian method of estimating a graphical structure is given by the graphical lasso. In this chapter, 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. The posterior distribution of different graphical models is extremely cumbersome to compute. We propose a fast computational method for approximating the posterior probabilities of various graphs using the Laplace approximation method by expanding the posterior density around the posterior mode, which is the graphical lasso by our choice of the prior distribution. We also provide estimates of the accuracy in the approximation.

Bayesian Inference for High Dimensional Models: Convergence Properties and Computational Issues

by
Sayantan Banerjee

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2014

APPROVED BY:

Peter Bloomfield	Soumendra Nath Lahiri
Brian Reich	Subhashis Ghoshal Chair of Advisory Committee

DEDICATION

To Adrija,

to my parents,

and

to my city, Chandannagar

BIOGRAPHY

Sayantan Banerjee was born to two wonderful parents, Chandan Kumar Banerjee and Malobika Banerjee, in the city of Chandannagar, West Bengal, India. He started schooling in The Study Home, before getting admitted to Sri Aurobindo Vidyamandir, one of the best schools in the state. He spent the next ten years (1993-2003) of his academic life there, gradually growing to the man of today. After spending two years for higher secondary studies at Kanailal Vidyamandir, he pursued his undergraduate studies in Statistics at St Xavier's College, Kolkata (2005-2008). Sayantan got admission at the prestigious Indian Statistical Institute, Kolkata for his Master's in Statistics (2008-2010). With a motto to build a career in research, he joined the PhD program in the Department of Statistics at North Carolina State University in August 2010. Apart from research work, Sayantan takes interest in a variety of other activities, ranging from literature to sports. He is an avid reader, writes occasionally, plays football (soccer, in US) and cricket, and wanders off often with his camera in hand, for his love of photography and travel. He loves to eat, is a tea-freak, and takes a passionate interest in railways. He met his soulmate Adrija Chatterjee in 2011, who remain his best treasure since then. Sayantan will be joining the University of Texas MD Anderson Cancer Center at Houston, Texas as a post-doctoral scholar from July 2014.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Professor Subhashis Ghoshal, for his immense help in the entire journey of my research career here at NC State. Without his help, advice, and mentoring, this wouldn't have been possible. His patience and sincerity makes him one of the best advisors one could ever have. The experience of working with him has been great, and I have learnt a lot from the one-to-one conversations we have had over the past few years. He has been most accommodating, caring, and had taken the relationship beyond the walls of academia, mixing it with good humor. Our discussions went beyond the titbits of epsilons and deltas, bringing in chats ranging from cricket to food and much more.

I would also like to thank all the other members in my thesis committee, namely, Professors Peter Bloomfield, Soumendra Nath Lahiri and Brian Reich, for careful review and providing valuable suggestions in improving the contents of this thesis. They have always been very generous to provide their valuable time and energy in the entire process. Also I would like to extend my sincere thanks to all the other Professors in the department for giving a platform of high quality research and training environment through their professional activities. Not only the Professors, I should thank all the staff, especially Alison McCoy and Adrian Blue, for their excellent service to the department and for making our lives easier.

Life as a graduate student is always difficult. Mine was no exception. The first journey to US itself had been a remarkable story for me. Nevertheless, the first few months would have been much difficult if there weren't the helping hands of Dr Sujit Ghosh and his wife Chumki di. I would forever remain indebted to their help and care. Also I would like to thank Dr Arnab Maity for all the valuable guidance he has provided from the very beginning.

My days here in the US would have been much dim but for the presence of two of the most wonderful persons I have ever come across – Samarpan Majumder and Rinku Majumder. Their

presence itself provided a sense of security. I have always felt blessed with their ever extending hands of love and care. Thank you!

I remain indebted towards all the teachers I have come across in my entire academic life. I extend my heartiest gratitude to these people who laid the foundations of success during my school days. Specially I should mention the roles of Shanti Ranjan Pal and Srikumar Ghosh, the two people who inspired me to choose a career in Statistics, remaining a guiding light throughout; and two of my most favorite teachers, Patralekha Basu and Gunamay Burman, for all I have learnt from them. I was also blessed with the best of the teachers during my college days. I would specially like to thank Amit Ghosh and Surupa Chakraborty at St Xavier's, and Tapas Samanta, Sumitra Purkayastha, Smarajit Bose at ISI Kolkata, for providing utmost help and guidance. And last but not the least, I would like to thank the great Partho Sarathi Chakrabarti from RKM Narendrapur for showing the path of effectively teaching Statistics at the undergraduate level. His style of teaching, ability to think about a statistical problem still remain a benchmark for me to follow as a teacher in the future.

I have been blessed with friends in this great department, for whom the department continues to be a home away from home. My sincere gratitude to all of them, including Weining Shen, Seung Jun Shin, Meng Li, Bo Zhang, Woo Sung Jang, Clemontina Davenport, Kasturi Talapatra, Cassie Kozyrkov – for all the good days. I take the opportunity to bid all of them good-bye with a big thanks through here. I wish all of them a huge success in life. Not only in my department, I also have been lucky enough to be amongst a handsome number of friends in the locality from other departments and also from neighboring universities like UNC Chapel Hill and Duke University, who continued to provide ample support throughout the days here – Lopamudra Kundu, for her deep caring love as a sister; Anjishnu Banerjee, for the brain-storming Bayesian sessions and cricket; Shuva Gupta, for all the help and concerns; Pratyaydipta Rudra, for being my anytime helpline; Pourab Roy, Sayan Dasgupta, Abhishek

Pal Majumdar, Sujatro Chakladar, Priyam Das, Pradip Tarafdar, Chiranjit Mukherjee, for extending their hands of friendship always – in normal times, and in times of crisis. And I should not forget to mention the role of Ayan Das Gupta, who treated me more than a brother and provided constant support (along with his culinary skills) during the two years he spent here at Raleigh.

School days have remained the best of the days from yesteryear. My childhood was awesome for my wonderful school Sri Aurobindo Vidyamandir, where I have spent the most vital 10 years of my life as a student. I will be ever grateful for all the love, support, care and fun my school had provided me. It would be a great mistake if I do not acknowledge the role of my friends in this regard, for accompanying me growing to the adult of today. I owe a lot to all my school friends; especially I would like to mention Nayan Banerjee, Arghadip Samaddar, Debanjan Mitra, Dipanjan Das, Arpan Sinha, Anindya Ganguly, Soham Mukherjee, Avinandan Sthanpati, Soumyadip Neogy, Sourav Ghosh and Soham Das, for all their love and support. I would also like to thank my friends from my college days. The five years I have spent at St Xavier's and ISI would have been incomplete without them – Abirbhab Bandyopadhyay, for being more than a best friend, and without the continued mental support of whom I wouldn't have sustained the journey this far; Soumya Banerjee, Shankhadeep Senchowdhury, Subhajyoti Sen, for always being the fun guys around; Sarani Sengupta, for all the care and support; Parichoy Pal Chowdhury, aka PPC, for just being himself; Sandipan Roy, Soumalya Mukhopadhyay, Projjwal Das, Sohini Dasgupta, for defining an epitome of eternal friendship. And I would also like to thank two of my junior friends, Debankur Mukherjee and Abhijoy Saha, for standing by my side always. I have learnt a lot from you two.

There are no shortcuts to success without working hard. I hereby express my heartiest gratitude to two people who have always remained a role-model to me in this regard – Satyajit Ray and Sachin Tendulkar. I can never express in words what they mean to me, and what

role they have played in motivating at times when everything seemed to fall apart. I turned to them in happiness, and in times of sorrow. And I would like to add a little vote of thanks for the Indian Cricket team and Manchester United Football Club for all the happiness they have provided since childhood.

This thesis wouldn't have seen light if not for the wonderful family I have, who have constantly motivated me throughout my life. Words fall short for the contribution of my father, Chandan Kumar Banerjee, who laid the foundation of whatever interest I have in mathematical sciences. The same goes for my mother, Malobika Banerjee, who taught me to read and write, and without whose love, care and blessings I couldn't have come this far. Thanks 'Mesomashai', Dr Chittaranjan Mukherjee, for truly respecting and supporting my decision to pursue a research career in Statistics, instilling the confidence in me when some significant others doubted. I would also thank all my uncles, aunts, cousins and grandparents for being supportive always. Their good wishes will remain the motivating forces for the days to come. A lot of thanks goes to my mother-in-law, Santasri Chatterjee, for all her love and blessings; and my father-in-law Tara Prasad Chatterjee, for all his best wishes. Amongst all the happiness, I spare some moments for my maternal grandparents, Late Narayan Chandra Chattopadhyay and Late Mira Chattopadhyay, and my very dear brother Late Arindam Mukherjee, who couldn't live to see this day. I miss you.

Finally, my heartiest thanks goes to the one without whom I am incomplete; my better half, Adrija Chatterjee. She has been prolific in providing more than her soul to help me breeze through the entire journey. She might have lived miles away most of the times, but never let me feel alone even for a moment. Without her constant encouragement and love, the stressful me couldn't have managed to deliver the best. I hereby gift this work to her.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

High-dimensional statistical inference has recently become one of the most important problems in statistics, in which the number of unknown parameters p is much larger than the available sample size n, that is, $p \gg n$. Such kind of high dimensional problems include regression, classification, clustering, multiple testing, random matrices, and graphical models. This work focuses on inference for high dimensional regression models including both parametric and non-parametric mean functions, and large random matrices like precision (inverse covariance) matrices using graphical models. We discuss some preliminaries on high dimensional regression and then introduce the concept of graphical models required for analysis of large matrices. We introduce classical and Bayesian methods available for inference on large matrices. We give a brief account of Bayesian asymptotics which we will be used in obtaining convergence rates in the subsequent chapters.

1.1 High dimensional regression models

Consider a linear regression model $Y_i = \boldsymbol{X}_i^T\boldsymbol{\beta} + \varepsilon_i$, $i = 1, \ldots, n$. Here $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the vector of regression parameters and $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ is the set of p predictors for the dependent variable Y. The random errors are $\varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$. One of the main goals in this context is consistent estimation of the regression coefficients. As the dimension p grows to infinity, the classical methods for estimation of the unknown parameters behave poorly. For instance, the least square estimator is not uniquely defined and has high variability. In such situations, we need to choose an effective set of predictors of dimension much smaller than p, the problem better known as variable selection problem. Variable selection methods aim at identifying a set of predictors for which the coefficients β_j are non-zero. Efficient variable selection methods help to improve accuracy in estimation, reduce computational complexity and make the models more interpretable. In such cases, reasonable estimation is possible if $\boldsymbol{\beta}$ is sparse in some sense. Generally, a condition like

$$\log p \times \{\text{sparsity}(\boldsymbol{\beta})\} \ll n, \tag{1.1}$$

depending on the type of sparsity specification on $\boldsymbol{\beta}$ will allow reasonable estimation.

1.1.1 Classical variable selection

A number of variable selection and estimation procedures for linear regression models abound the literature; see, for example Miller (2002) and George (2000). Most of these procedures use the idea of penalization of the negative log-likelihood. Of these, one of the most significant and widely used techniques is the Least Absolute Shrinkage and Selection Operator, or the lasso, which is immensely popular for its simplicity, prediction accuracy, ease in computation and

applicability in a variety of regression models. We briefly discuss the concept of lasso below.

The linear regression model above can be written in the matrix and vector notation as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.2}$$

where $\boldsymbol{Y}$ is the $n \times 1$ vector of responses, $\boldsymbol{X}$ is the $n \times p$ design matrix, $\boldsymbol{\beta}$ is the vector of regression parameters and $\boldsymbol{\varepsilon}$ is the $n \times 1$ vector of random errors. The lasso estimator is defined as

$$\hat{\boldsymbol{\beta}}(\lambda) = \text{argmin}_{\boldsymbol{\beta}} \left(\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right), \tag{1.3}$$

where $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$ and $\lambda \geq 0$ is the penalty parameter. For $\lambda = 0$, the solution becomes the ordinary least square estimate. Depending on the choice of λ, the lasso shrinks the least square solutions towards zero. The most important aspect of the lasso is that it produces exact zero solutions for some components of the parameter, that is, for some j, $\hat{\beta}_j(\lambda) = 0$, so that variable selection is automatically done. The tuning parameter λ may be chosen by cross-validation or using model selection criteria like the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC). These techniques perform well for prediction purposes, but a larger value of the penalty parameter is often necessary for effective variable selection (Bühlmann and van de Geer, 2011). Consider the set of variables selected by the lasso for a given λ, given by

$$\hat{S}(\lambda) = \{j \colon \hat{\beta}_j \neq 0,\, j = 1, \ldots, p\}, \tag{1.4}$$

and the true set of effective variables be $S_0 = \{j \colon \beta_j^0 \neq 0,\, j = 1, \ldots, p\}$, where β_j^0, $j = 1, \ldots, p$, are the true values of the regression coefficients. Then, under certain conditions on

the parameter space, it can be shown that for $\lambda = \lambda_n \gg n^{-1/2}(\log p)^{1/2}$,

$$\mathrm{P}(\hat{S}_\lambda = S_0) \to 1, \tag{1.5}$$

as $n \to \infty$ (Meinshausen and Bühlmann, 2006). This result implies that the lasso procedure performs consistent variable selection under appropriate sparsity assumptions on the true model.

Many other variable selection approaches are variations on this penalized regression theme and typically differ from the lasso by varying the form of the penalty; see, for example, Breiman (1995); Fan and Li (2001); Zou and Hastie (2005); Zou (2006); Bondell and Reich (2008); Hwang et al. (2009) and so on.

In some situations, the underlying predictors form natural groupings, for example, sets of dummy variables for factors, or in additive models where the effect of the predictors are expanded in a basis. Yuan and Lin (2006) presented a variable selection technique, called the group lasso in this context. Similar to the lasso, the group lasso is also a penalized least-squares method that uses a special form of penalty, called the group-L_1 penalty, to eliminate redundant variables from the model simultaneously in a pre-specified groups of variables.

Let $\boldsymbol{Y}$ be an $n \times 1$ vector of responses, $\boldsymbol{X}_j$ is an $n \times m_j$ matrix of variables associated with the jth predictor (which may be fixed or random) and $\boldsymbol{\beta}_j$ is an $m_j \times 1$ vector of coefficients. Then the group lasso minimizes

$$\mathrm{argmin}_{\boldsymbol{\beta}} \left\| \boldsymbol{Y} - \sum_{j=1}^{g} \boldsymbol{X}_j \boldsymbol{\beta}_j \right\|_2^2 + \lambda \sum_{j=1}^{g} \|\boldsymbol{\beta}_j\|_2, \tag{1.6}$$

where $\boldsymbol{\beta} = ({\boldsymbol{\beta}_1}^T, \ldots, {\boldsymbol{\beta}_g}^T)^T$ and g is the number of groups. For appropriate choices of the tuning parameter λ, the group lasso solution includes $\boldsymbol{\beta}_j = \mathbf{0}$ for some subset of $j = 1, \ldots, g$.

1.1.2 Bayesian variable selection

One drawback of most classical variable selection methods is that they do not provide a measure of model uncertainty and typically give one model as the best, without giving some measurement of uncertainty for this estimated model. The exceptions to this are methods that follow the Bayesian paradigm. They provide a measure of model uncertainty by computing the posterior probabilities of models, which are typically estimated by the proportions of visits to these models in the posterior draws from a Markov chain Monte Carlo (MCMC) simulation (George and McCulloch, 1993). However, MCMC methods are computationally expensive when a large number of variables are involved and it can be hard to assess convergence when MCMC methods must traverse a space of differing dimensions. In fact, when the model dimension is quite high, most MCMC based methods break down.

Consider the linear regression set-up as in equation (1.2), and define the p-dimensional vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^T$ such that $\gamma_j = 1$ if X_j is included in the model and $\gamma_j = 0$ otherwise. Let $\boldsymbol{X}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ respectively denote the matrix of covariates and the vector of regression coefficients corresponding to the non-zero elements of $\boldsymbol{\gamma}$. In a Bayesian variable selection approach, a prior distribution is assigned to all possible models $\boldsymbol{\gamma}$ along with prior distributions for the unknown parameters $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and σ^2. Prior distributions are usually specified in a hierarchical fashion, as

$$p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \sigma^2) = p(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma}, \sigma^2) p(\boldsymbol{\gamma}) p(\sigma^2). \tag{1.7}$$

The marginal posterior distribution of a model $\boldsymbol{\gamma}$ is then given by

$$p(\boldsymbol{\gamma} \mid \boldsymbol{Y}) \propto p(\boldsymbol{Y} \mid \boldsymbol{\gamma}) p(\boldsymbol{\gamma}), \tag{1.8}$$

where

$$p(\boldsymbol{Y} \mid \boldsymbol{\gamma}) = \int p(\boldsymbol{Y} \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \sigma^2) p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2 \mid \boldsymbol{\gamma}) d\boldsymbol{\beta}_{\boldsymbol{\gamma}}\, d\sigma^2. \tag{1.9}$$

Posterior probabilities of models can be used to determine the model with highest posterior probability. Barbieri and Berger (2004) showed that the median probability model, defined as the collection of all variables whose marginal inclusion probabilities are at least $1/2$, has better prediction properties compared with the highest posterior probability model. In addition to this, the posterior probabilities of various models can be used to perform Bayesian model averaging (BMA), which incorporates model uncertainty and is typically preferred for prediction purposes.

Choice of prior distribution on regression coefficients plays a vital role in model assessment. In many instances, independent priors are put on the regression coefficients as a mixture of two components conditioned on $\boldsymbol{\gamma}$, given by

$$p(\beta_j \mid \gamma_j) = (1 - \gamma_j) p^{(0)}(\beta_j) + \gamma_j p^{(1)}(\beta_j). \tag{1.10}$$

Mitchell and Beauchamp (1988) proposed the 'spike and slab' prior for β_j by choosing the priors $p^{(0)}(\beta_j) = \mathbb{1}_{\{0\}}(\beta_j)$ and $p^{(1)}(\beta_j) = \mathbb{1}_{[-a,a]}(\beta_j)/2a$. Note that this prior has positive prior probability for $\beta_j = 0$ so that variable selection can be done using model posterior probabilities. George and McCulloch (1993) used a mixture prior for $\boldsymbol{\beta}$ given by $p^{(0)}(\beta_j) = \mathrm{N}(0, \tau_j^2)$ and $p^{(1)}(\beta_j) = \mathrm{N}(0, c_j^2 \tau_j^2)$. Here the constant τ_j^2 is chosen to be very small so that if $\gamma_j = 0$, then β_j has very negligible variance and may be dropped from the selected model. On the other hand, the constant c_j^2 is chosen to be large, so that if $\gamma_j = 1$, β_j is included in the final model.

In most of the applications for variable selection, the dimension p is large, and hence evaluation of all 2^p possible models is not feasible. George and McCulloch (1993) applied a Gibbs sampling algorithm (Geman and Geman, 1984; Gelfand and Smith, 1990) to stochastically

search the model space (SSVS; also see George and McCulloch, 1997). A related MCMC based approach is the MC^3 algorithm (Raftery et al., 1997) based on random-walk Metropolis sampler. Other stochastic search variable selection methods include the Laplace approximation based method by Yuan and Lin (2005) and the rescaled spike and slab prior proposed by Ishwaran and Rao (2005).

Dependent priors on $\boldsymbol{\beta}$ are also used for variable selection. One of the most widely used priors in this regard is Zellner's g-prior (Zellner, 1986) given by

$$p(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \sigma^2, g) \sim \mathrm{N}(\mathbf{0}, g\sigma^2(\boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X}_{\boldsymbol{\gamma}})^{-1}). \tag{1.11}$$

The constant g is used to control the uncertainty in the prior relative to the variance of the observations around the mean, and the $(\boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X}_{\boldsymbol{\gamma}})^{-1}$ term provides a prior covariance structure for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$. The posterior model probabilities are analytically expressible in this case, and hence are computationally simple. To resolve some inconsistencies arising from using a fixed g, Liang et al. (2008) used mixtures of Zellner's g-prior, such as the Zellner-Siow Cauchy prior (Zellner and Siow, 1980) or the hyper-g prior (Liang et al., 2008).

Penalized regression methods in classical variable selection can also be viewed from the Bayesian perspective. For instance, the lasso can be characterized as the posterior mode when the regression parameters are assigned a common Laplace prior independent of the coefficients. Park and Casella (2008) designed MCMC schemes for computing the posterior distribution by representing the Laplace distribution as scale mixture of normals, and termed the resulting procedure the Bayesian lasso. It may be noted that the Bayesian lasso does not incorporate any sparsity unless it is complemented by an external thresholding procedure. Yuan and Lin (2005) used the Laplace prior for the regression coefficients along with a point mass at zero, and evaluated approximate posterior probabilities of various models using Laplace approximations

by expanding the prior times the likelihood around the posterior mode, which in this case is identical with the lasso.

Similar to the lasso, the group lasso can be viewed as the posterior mode with respect to some appropriate multivariate Laplace prior, where the grouping structure is induced by the dependence in the components. Curtis et al. (2014) used such a kind of prior for variable selection and developed a fast approximate method of evaluating posterior probabilities of different models in the context of nonparametric additive regression. We shall extend their technique for variable selection to cover generalized additive partial linear models.

1.2 Undirected graphical models

We first introduce some preliminary concept on graphs, and then proceed to the notion of graphical models.

An *undirected* graph G consists of a finite non-empty set V of p points, called vertices, and a set of edges $E = \{(i, j) \in V \times V; i < j\}$. A *subgraph* $G' = (V', E')$ of $G = (V, E)$ is a graph such that $V' \subseteq V, E' = \{(i, j) \in V' \times V'; i < j\}$.

In a *graphical model*, the vertex set $V = \{1, \ldots, p\}$ of a graph G corresponds to a p-dimensional random variable $\boldsymbol{X} = (X_1, \ldots, X_p)^T \sim P$. The graph G equipped with the probability distribution P is referred to as a graphical model (G, P). For an undirected graph, the graphical model is called an undirected graphical model. We shall consider all the related concepts introduced henceforth for undirected graphical models.

1.2.1 Markov properties of a graphical model

Related to an undirected graphical model (G, P), we define Markov properties of the graphical model as follows.

Definition 1. *The probability distribution P satisfies the pairwise Markov property with respect to an undirected graph G if for any distinct vertices j and k such that $(j,k) \notin E$,*

$$X_j \perp X_k \mid X_{V\setminus\{j,k\}}$$

A stronger version of the pairwise Markov property is the global Markov property which we define below.

Definition 2. *The probability distribution P satisfies the global Markov property with respect to an undirected graph G if for disjoint subsets A, B, C such that C separates A and B, we have,*

$$X_A \perp X_B \mid X_C$$

Note that the global Markov property implies the pairwise Markov property. The two become equivalent under some special cases.

Proposition 1. *If the distribution P has a positive and continuous density with respect to a Lebesgue measure, then the pairwise and global Markov properties of P are equivalent.*

A *Conditional Independence Graph* (CIG) is a graphical model for which the pairwise Markov property holds.

1.2.2 Gaussian graphical models

Consider a p-dimensional Gaussian random variable $\boldsymbol{X} = (X_1, \ldots, X_p)^T \sim \mathrm{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$. A conditional independence graph with a Gaussian distribution on the random variables indexed by the vertices of the graph is called a *Gaussian graphical model* (GGM). Due to Gaussian assumptions, the global Markov property is also satisfied. For a GGM, the presence or absence

of an edge is reflected in the zeroes of the precision matrix $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ of the random variables. Thus we have,

$$(j,k) \notin E \iff X_j \perp X_k \mid X_{V\setminus\{j,k\}} \iff \omega_{jk} = 0, \tag{1.12}$$

where ω_{jk} is the (j,k)th element of $\mathbf{\Omega}$. Thus, the conditional independence of any two components of the random variable $\boldsymbol{X}$ is reflected by a zero in the corresponding position of the precision matrix.

1.3 Inference on large matrices

Estimating a covariance matrix or a precision matrix (inverse covariance matrix) is one of the most important problems in multivariate analysis. Of special interest are situations where the number of underlying variables p is much larger than the sample size n. Situations like this are common in gene expression data, fMRI data and in several other modern applications. Special care needs to be taken for tackling such high-dimensional scenarios. Conventional estimators like the sample covariance matrix or maximum likelihood estimator behave poorly when the dimensionality is much higher than the sample size.

1.3.1 Classical methods for estimating large matrices

Different regularization based methods have been proposed and developed in the recent years for dealing with high-dimensional data. These include banding, thresholding, tapering and penalization based methods to name a few; see, for example, Ledoit and Wolf (2004); Huang et al. (2006); Yuan and Lin (2007); Bickel and Levina (2008a,b); Karoui (2008); Friedman et al. (2008); Rothman et al. (2008); Lam and Fan (2009); Rothman et al. (2009); Cai et al. (2010, 2011). Most of these regularization based methods for high dimensional models impose a

sparse structure in the covariance or the precision matrix, as in Bickel and Levina (2008a), where a rate of convergence was derived for the estimator obtained by "banding" the sample covariance matrix, or by banding the Cholesky factor of the inverse sample covariance matrix, as long as $n^{-1} \log p \to 0$. Cai et al. (2010) obtained the minimax rate under the operator norm and constructed a tapering estimator which attains the minimax rate over a smoothness class of covariance matrices. Cai and Liu (2011) proposed an adaptive thresholding procedure. More recently, Cai and Yuan (2012) introduced a data-driven block-thresholding estimator which is shown to be optimally rate adaptive over some smoothness classes of covariance matrices.

For estimation of a sparse inverse covariance matrix, graphical models (Lauritzen, 1996) provide an excellent tool, as the conditional dependency between the variables is captured by means of an undirected graph; see Dobra et al. (2004); Meinshausen and Bühlmann (2006); Yuan and Lin (2007); Friedman et al. (2008). There are several methods in the frequentist literature for 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 (Meinshausen and Bühlmann, 2006; Friedman et al., 2008; Banerjee et al., 2008; Yuan and Lin, 2007; Guo et al., 2011; Witten et al., 2011). Other methods include the Sparse Permutation Invariant Covariance Estimator (SPICE) as developed by Rothman et al. (2008).

1.3.2 Bayesian methods for estimating large matrices

There are only a few relevant work in Bayesian inference for such kind of problems. Ghosal (2000) studied asymptotic normality of posterior distributions for exponential families when

the dimension $p \to \infty$, but restricting to $p \ll n$. Recently, Pati et al. (2012) 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 of the true covariance matrix.

Bayesian methods for inference using graphical models have also been developed, as in Roverato (2000); Atay-Kayis and Massam (2005); Letac and Massam (2007). For a complete graph corresponding to the saturated model, clearly the Wishart distribution is the conjugate prior for the precision matrix $\mathbf{\Omega}$. For an incomplete decomposable graph, a conjugate family of priors is given by the G-Wishart prior (Roverato, 2000). The equivalent prior on the covariance matrix is termed as the hyper inverse Wishart distribution in Dawid and Lauritzen (1993). Letac and Massam (2007) 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 distribution were further explored by Rajaratnam et al. (2008), who also obtained expressions for Bayes estimators under different loss functions. Wang (2012) developed a Bayesian version of the graphical lasso, 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 (Park and Casella, 2008), 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 distribution.

1.4 Asymptotics in Bayesian inference

In this section, we discuss and review the asymptotic properties of posterior distributions. We will start with discussion on posterior consistency. Let us consider a sequence of statistical

experiments parametrized by a parameter θ, possibly of infinite dimension, taking values in a separable metric space. Let Π be the prior distribution on θ and for data of size n denoted by $\boldsymbol{X}^{(n)}$, the posterior distribution is given by $\Pi(\cdot \mid \boldsymbol{X}^{(n)})$. The joint density of observations is denoted by $p_{\theta,n}(\boldsymbol{X}^{(n)})$. The posterior distribution is said to be *consistent* at a given θ_0, if

$$\Pi(d(\theta, \theta_0) > \epsilon \mid \boldsymbol{X}^{(n)}) \to 0 \,\text{in}\, P_{\theta_0}^{(n)}\text{probability}, \tag{1.13}$$

as $n \to \infty$ for some distance metric d and any $\epsilon > 0$ when θ_0 is the true value of the parameter.

Consistency is almost guaranteed for finite-dimensional parameter spaces, if the support of the prior is in the neighborhoods of the true parameter. But for infinite-dimensional spaces, this alone is not sufficient to ensure posterior consistency. Some counterexamples in this context were discussed in Freedman (1963), Diaconis and Freedman (1986), Kim and Lee (2001) and James (2008). These examples show that more conditions are required on the prior to ensure consistency.

If explicit forms of the posterior distribution is available, then posterior consistency results may be proved using Chebychev-type inequalities. Examples of such include Bayesian survival analysis problems where posterior conjugacy holds for priors described by an independent increment process, and also in the case of posterior distributions arising from Dirichlet process priors. Posterior consistency also holds for Bayesian estimation of a c.d.f, using tail-free priors. All of these involve much restrictive cases, and better approaches need to be exploited for checking consistency in the general case. The theory by Schwartz (1965) is useful in this respect, when the family of densities is dominated. The theory provides two sufficient conditions required for consistency. The first condition is to construct strictly unbiased tests for $\theta = \theta_0$ against the alternative $\theta \in B$, B being the complement of a neighborhood U of θ_0. The existence of such tests ensures that the type I and type II error probabilities are going to zero

exponentially fast. The second condition, known as Schwartz's prior positivity condition, or the *Kullback-Leibler property* of the prior, requires that $\Pi(\theta\colon K(p_{\theta_0}; p_\theta) < \epsilon) > 0$ for all $\epsilon > 0$, where $K(p_{\theta_0}; p_\theta) = \int p_{\theta_0} \log(p_{\theta_0}/p_\theta) d\nu$ is the Kullback-Leibler divergence between the two densities p_{θ_0} and p_θ (with respect to a dominating measure ν). This condition is a crucial one for establishing consistency.

However, the above procedure faces difficulties in infinite-dimensional spaces with stronger topologies, unless the parameter space is compact. Ghosal et al. (1999) considered a technique based on truncating the parameter space depending on the sample size. For this, they considered a sequence of subsets of the parameter space, called sieve, and show that by applying Schwartz's testing criterion on a carefully chosen sieve, posterior consistency can be obtained by bounding the metric entropy of the sieve and the prior probability of the complement of the sieve. A similar result under stronger condition using bracketing entropy was given by Barron et al. (1999).

We now discuss convergence rates of posterior distributions, which is another important aspect of Bayesian asymptotics. For a data $\boldsymbol{X}^{(n)}$ generated by a model $P_\theta^{(n)}$, a sequence $\epsilon_n \to 0$ is called the *convergence rate of the posterior distribution* $\Pi(\cdot \mid X^{(n)})$ at the true parameter θ_0 with respect to a pseudo-metric d, if

$$\Pi(\theta\colon d(\theta, \theta_0) \geq M_n \epsilon_n) \to 0 \,\text{in}\, P_{\theta_0}^{(n)}\text{probability}, \tag{1.14}$$

for any sequence $M_n \to \infty$.

The convergence rate in regular parametric families is well known to be $n^{-1/2}$, whereas for infinite-dimensional models, the rate is generally slower than $n^{-1/2}$. If the posterior converges at the rate ϵ_n, then there exists point estimator which converges to the true parameter at rate ϵ_n in the frequentist sense (Ghosal et al., 2000). Thus the posterior convergence rate can never

be better than the minimax rate in a class, for which attaining the latter can be regarded as the goal. Posterior convergence rate may be obtained in some situations using Chebychev-type inequalities, in problems where the parameter space is equipped with the L_2-norm and explicit expressions for posterior mean and posterior variance are available. Also for infinite-dimensional problems, convergence rate may be obtained by explicit calculations in situations involving conjugate normal priors for normal mean. General results on obtaining posterior convergence rates was given by Ghosal et al. (2000). We briefly summarize their idea here. For i.i.d. observations $X_1, \ldots, X_n \sim p_0$, the posterior probability of the set $B = \{p\colon d(p, p_0) \geq \epsilon_n\}$ is expressed as a ratio

$$\Pi(\theta \in B \mid X_1, \ldots, X_n) = \frac{\int_B p_{\theta,n}(X_1, \ldots, X_n)/p_{\theta_0,n}(X_1, \ldots, X_n) d\Pi(\theta)}{\int p_{\theta,n}(X_1, \ldots, X_n)/p_{\theta_0,n}(X_1, \ldots, X_n) d\Pi(\theta)}. \tag{1.15}$$

Then, the desired convergence rate may be established if it can be shown that, (i) the numerator is upper bounded by $e^{-cn\epsilon_n^2}$, where $c > 0$ can be chosen sufficiently large, and (ii) the denominator is lower bounded by $e^{-bn\epsilon_n^2}$. For satisfying the second assertion, the Kullback-Leibler positivity condition in the consistency result is replaced by a stronger one given by

$$\Pi(B(p_0, \epsilon_n)) \geq e^{-b_1 n\epsilon_n^2}, \tag{1.16}$$

where $B(p_0, \epsilon_n) = \{p\colon K(p_0; p) \leq \epsilon_n^2, V(p_0; p) \leq \epsilon_n^2\}$, and $V(p_0; p) = \int p_0(\log(p_0/p))^2$. Thus the prior distribution is required to have sufficient level of concentration around the true density p_0 in terms of the first and second moments of the log-likelihood ratio. This ϵ_n is termed as the *prior concentration rate*. For bounding the numerator of (1.15), the testing approach of Schwartz (1965) is used. We find tests for the null hypothesis $p = p_0$ against the alternative $\{p\colon d(p_0, p) \geq \epsilon_n\}$, such that the type I and type II error probabilities are exponentially small.

Unless the parameter space is compact, such tests are constructed by the technique of sieves. The sieve $\mathcal{P}_n$ is covered using balls of size $\bar{\epsilon}_n/2$ and the number of such balls need to be controlled by the metric entropy condition

$$\log N(\bar{\epsilon}_n/2, \mathcal{P}_n, d) \leq c_1 n \bar{\epsilon}_n^2. \tag{1.17}$$

In addition to this, we also need to ensure that the prior probability on the complement of the sieve is exponentially small, that is, $\Pi(\mathcal{P}_n^c) \leq e^{-c_2 n \epsilon_n^2}$. The posterior convergence rate is obtained by choosing the maximum of the prior concentration rate ϵ_n and $\bar{\epsilon}_n$. For obtaining the minimax rate, the two rates need to match.

1.5 Notations

We now describe the notations to be used in this work. By $t_n = O(\delta_n)$ (respectively, $o(\delta_n)$), we mean that t_n/δ_n is bounded (respectively, $t_n/\delta_n \to 0$ as $n \to \infty$). For a random sequence X_n, $X_n = O_P(\delta_n)$ (respectively, $X_n = o_P(\delta_n)$) means that $\mathrm{P}(|X_n| \leq M\delta_n) \to 1$ for some constant M (respectively, $\mathrm{P}(|X_n| < \epsilon\delta_n) \to 1$ for all $\epsilon > 0$). For numerical sequences r_n and s_n, by $r_n \ll s_n$ (or, $r_n \gg s_n$) we mean that $r_n = o(s_n)$, while by $s_n \gtrsim r_n$ we mean that $r_n = O(s_n)$. By $r_n \asymp s_n$, we mean that $r_n = O(s_n)$ and $s_n = O(r_n)$, while $r_n \sim s_n$ stands for $r_n/s_n \to 1$. The indicator function is denoted by $\mathbb{1}$.

We represent vectors in bold lowercase English or Greek letters. The components of a vector are represented by the corresponding non-bold letters, that is, for $\boldsymbol{x} \in \mathbb{R}^p$, $\boldsymbol{x} = (x_1, \ldots, x_p)^T$. We define the following norms for a vector $\boldsymbol{x} \in \mathbb{R}^p$: $\|\boldsymbol{x}\|_r = \left(\sum_{j=1}^p |x_j|^r\right)^{1/r}$, $\|\boldsymbol{x}\|_\infty = \max_j |x_j|$. If $r = 2$, this reduces to the usual Euclidean norm in which case we typically drop the subscript and simply write $\|\boldsymbol{x}\|$. Matrices are denoted in bold uppercase

English or Greek letters, like $\boldsymbol{A} = (\!(a_{ij})\!)$, where a_{ij} stands for the (i,j)th entry of $\boldsymbol{A}$. The identity matrix of order p will be denoted by $\boldsymbol{I}_p$. If $\boldsymbol{A}$ is a symmetric $p \times p$ matrix, let $\mathrm{eig}_1(\boldsymbol{A}) \leq \ldots \leq \mathrm{eig}_p(\boldsymbol{A})$ stand for its eigenvalues and let the trace of $\boldsymbol{A}$ be denoted by $\mathrm{tr}(\boldsymbol{A})$. Viewing $\boldsymbol{A}$ as a vector in $\mathbb{R}^{p^2}$, we define L_r, $1 \leq r < \infty$ and L_∞-norms on $p \times p$ matrices as

$$\|\boldsymbol{A}\|_r = \left(\sum_{i=1}^{p} \sum_{j=1}^{p} |a_{ij}|^r \right)^{1/r}, \ 1 \leq r < \infty, \quad \|\boldsymbol{A}\|_\infty = \max_{1 \leq i,j \leq p} |a_{ij}|.$$

Note that $\|\boldsymbol{A}\|_2 = \sqrt{\mathrm{tr}(\boldsymbol{A}^T\boldsymbol{A})}$, the Frobenius norm. Viewing $\boldsymbol{A}$ an operator from $(\mathbb{R}^p, \|\cdot\|_r)$ to $(\mathbb{R}^p, \|\cdot\|_s)$, where $1 \leq r, s \leq \infty$, we can also define, $\|\boldsymbol{A}\|_{(r,s)} = \sup(\|\boldsymbol{A}\boldsymbol{x}\|_s \colon \|\boldsymbol{x}\|_r = 1)$. We refer to the norm $\|\cdot\|_{(r,r)}$ as the L_r-operator norm. This gives

$$\|\boldsymbol{A}\|_{(1,1)} = \max_j \textstyle\sum_i |a_{ij}|, \quad \|\boldsymbol{A}\|_{(\infty,\infty)} = \max_i \textstyle\sum_j |a_{ij}|,$$
$$\|\boldsymbol{A}\|_{(2,2)} = \{\max(\mathrm{eig}_i(\boldsymbol{A}^T\boldsymbol{A}) \colon 1 \leq i \leq p)\}^{1/2},$$

and that for symmetric matrices, $\|\boldsymbol{A}\|_{(2,2)} = \max\{|\mathrm{eig}_i(\boldsymbol{A})| \colon 1 \leq i \leq p\}$, and $\|\boldsymbol{A}\|_{(1,1)} = \|\boldsymbol{A}\|_{(\infty,\infty)}$. We state the following lemma involving relations between different matrix norms.

Lemma 1. *For symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of order p, we have the following:*

1. $\|\boldsymbol{A}\|_{(2,2)} \leq \|\boldsymbol{A}\|_{(\infty,\infty)} \leq \sqrt{p}\|\boldsymbol{A}\|_{(2,2)}$;
2. $\|\boldsymbol{A}\|_\infty \leq \|\boldsymbol{A}\|_{(2,2)} \leq \|\boldsymbol{A}\|_{(\infty,\infty)} \leq p\|\boldsymbol{A}\|_\infty$;
3. $\|\boldsymbol{A}\|_{(2,2)} \leq \|\boldsymbol{A}\|_2 \leq p\|\boldsymbol{A}\|_\infty$;
4. $\|\boldsymbol{A}\boldsymbol{B}\|_2 \leq \|\boldsymbol{A}\|_{(2,2)}\|\boldsymbol{B}\|_2$, $\|\boldsymbol{A}\boldsymbol{B}\|_2 \leq \|\boldsymbol{A}\|_2\|\boldsymbol{B}\|_{(2,2)}$.

For two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we say that $\boldsymbol{A} \geq \boldsymbol{B}$ (respectively, $\boldsymbol{A} > \boldsymbol{B}$) if $\boldsymbol{A} - \boldsymbol{B}$ is nonnegative definite (respectively, positive definite). Thus $\boldsymbol{A} > \boldsymbol{0}$ for a positive definite matrix

$\boldsymbol{A}$, where $\mathbf{0}$ stands for the zero matrix. $\boldsymbol{A}^{1/2}$ stands for the unique positive definite square root of a positive definite matrix $\boldsymbol{A}$.

Sets are denoted in non-bold uppercase English letters. For a set T, we denote the cardinality, that is, the number of elements in T, by $\#T$. We denote the submatrix of the matrix $\boldsymbol{A}$ induced by the set $T \subset \{1, 2, \ldots, p\}$ by $\boldsymbol{A}_T$, i.e., $\boldsymbol{A}_T = (\!(a_{ij}\colon i, j \in T)\!)$. By $\boldsymbol{A}_T^{-1}$, we mean the inverse $(\boldsymbol{A}_T)^{-1}$ of the submatrix $\boldsymbol{A}_T$. For a $p \times p$ matrix $\boldsymbol{A} = (\!(a_{ij})\!)$, let $(\boldsymbol{A}_T)^0 = (\!(a^*_{ij})\!)$ denote a p-dimensional matrix such that $a^*_{ij} = a_{ij}$ for $(i, j) \in T \times T$, and 0 otherwise. Also we denote the "banded" version of $\boldsymbol{A}$ by $B_k(\boldsymbol{A}) = (\!(a_{ij} \mathbb{1}\{|i-j| \leq k\})\!)$ corresponding to banding parameter k, $k < p$. Finally, we define $\boldsymbol{E}_{(i,j)} = (\!(\mathbb{1}_{\{(i,j),(j,i)\}}(l, m))\!)$, which is a symmetric matrix.

The Hellinger distance between two densities q_1 and q_2 is given by $h(q_1, q_2) = \|\sqrt{q_1} - \sqrt{q_2}\|_2$.

For a subset A of a metric space (S, d), $N(\epsilon, A, d)$ denotes the ϵ-covering number of A with respect to d, that is, the minimum number of d-balls of size ϵ in S needed to cover A.

Chapter 2

Bayesian variable selection in generalized additive partial linear models

2.1 Introduction

Linear models are widely used to analyze the relationship between any response variable with relevant predictors of interest. If the response variable is discrete, a linear model is often inappropriate. Generalized linear models (GLM) (McCullagh and Nelder, 1983) provide useful generalization of linear models which can handle various types of response variables including discrete ones. In a GLM, the relationship between distribution of the response variable and the predictors is expressed in terms of a linear functional form through a link function depending on the distributional assumption of the underlying response. Typically these distributions are assumed to belong to the exponential family, which includes a wide range of distributions including binomial, Poisson and normal distributions. Regression modeling using GLMs can be extended to generalized additive models (GAM), where the linear functional part in regression is substituted by sum of smooth functions of unknown functional form in the underlying pre-

dictors (Hastie and Tibshirani, 1990; Wood, 2006). GAMs are especially useful in situations when the relationship between the response variable and the predictors for a given link function is not linear. GAMs can be made further flexible by allowing a linear component for some predictors which are presumed to have a strictly linear effect, and an additive structure for other predictors. These models, known as generalized additive partial linear models (GAPLM), give a natural extension of additive partial linear models (APLM) for discrete response variables. GAPLMs incorporate more flexibility in the models by allowing both parametric and non-parametric components, and particularly useful in situations where there is a prior knowledge that some predictors can have a linear effect only, for example, binary predictors or dummy indicator variables related to discrete predictors.

Statistical inference for GAPLMs have been well studied in the literature, with procedures including kernel-based backfitting and local scoring (Buja et al., 1989) and marginal integration (Linton and Nielsen, 1995). Methods involving penalized regression splines (Marx and Eilers, 1998; Ruppert et al., 2003; Wood, 2004) are also widely used in this regard, due to computational simplicity and ease of implementation; see 'gam' and 'mgcv' package in R.

An important aspect in this regard is variable selection in these models. Variable selection in generalized linear models have been widely studied in the literature; for example, see Raftery and Richardson (1993); Raftery (1996); Chen et al. (1999); Dellaportas and Forster (1999); Meyer and Laud (2002); Ntzoufras et al. (2003); Wang and George (2007). Variable selection methods for additive models have also been well studied; see Chen (1993); Shively et al. (1999); Shi and Tsai (1999); Gustafson (2000); Wood et al. (2002); Avalos et al. (2003); Lin and Zhang (2006); Belitz and Lang (2008); Meier et al. (2009); Ravikumar et al. (2009); Reich et al. (2009); Huang et al. (2010); Marra and Wood (2011); Curtis et al. (2014). The literature is quite sparse for variable selection in GAPLM. Wang et al. (2011) considered GAPLMs, restricting to variable selection for the parametric part only. Variable selection in

APLMs has been studied much recently. Liu et al. (2011) developed a SCAD-based variable selection procedure using a spline based approximation for the non-parametric components. In contrast to a number of Bayesian methods for variable selection in additive models, to the best of our knowledge, there is no Bayesian variable selection method for GAPLMs available in the literature. Bayesian methods available in variable selection problems provide measures of model uncertainty. However, in the high dimensional setting, Bayesian variable selection methods are computationally expensive, since commonly used MCMC based procedures do not scale well in high dimension. Moreover, even when such a procedure is implemented, estimates of posterior probabilities from MCMC visits to different models can be extremely unreliable when the number of parameters is high. Curtis et al. (2014) developed a Bayesian variable selection method in non-parametric additive regression models by approximating different model posterior probabilities using Laplace approximation, thus avoiding any time-consuming MCMC based procedures. In this chapter we extend the idea to GAPLMs, thus providing a fast Bayesian variable selection technique for the same.

We expand each function in the non-parametric part of the model in a B-spline basis and put a multivariate Laplace prior on the resulting coefficients. For predictors with a linear effect, we put a univariate Laplace prior on each of the coefficients. The posterior mode of can be identified as the group lasso for generalized linear models. We use the group lasso to approximate different model posterior probabilities using Laplace approximation. The approximate probabilities are used to find the model with highest posterior probability and also the median probability model. We compute relevant measures like specificity and sensitivity to evaluate the model selection performance of the Bayesian method.

The chapter is organized as follows. In Section 2.2, we specify the model and discuss the prior distributions on the coefficients. In the next section, we derive the form of the posterior distribution of various models and develop the Laplace approximation along with estimation

methods for selecting the penalty parameter for the group lasso. We also discuss some common families of distributions used often in practice in the context of generalized linear models. In Section 2.4, we discuss APLM as a special case of GAPLM. In Section 2.5, we evaluate the performance of the proposed Bayesian method through simulation studies and also present three real data analyses.

2.2 Model and prior specification

Consider a response variable Y and $p+s$ predictors $\boldsymbol{X} = (X_1, \ldots, X_p)^T$, $\boldsymbol{Z} = (Z_1, \ldots, Z_s)^T$. The response Y follows an exponential family density of the form $\exp\{a(\eta)y + b(\eta) + c(y)\}$, where $a(\eta)$ and $b(\eta)$ are continuously differentiable, with $a(\eta)$ having a non-zero derivative. The mean of the distribution is given by $\mu = -b'(\eta)/a'(\eta) \equiv \psi(\eta)$, and ψ is called the link function. The inverse map, $\eta = \psi^{-1}(\mu)$ which represents the predictor in terms of the mean of the response variable, is called the inverse link function. The true relation between Y and $(\boldsymbol{X}, \boldsymbol{Z})$ is assumed to follow a generalized additive partial linear model (GAPLM) given by

$$\eta = \boldsymbol{X}^T\boldsymbol{\beta} + \sum_{j=1}^{s} f_j(Z_j), \tag{2.1}$$

where $\boldsymbol{X}$ consists of those predictors which have a strictly linear effect, including binary or other categorical covariates, and those with non-linear effects are collected in $\boldsymbol{Z}$. The functions $f_j(\cdot)$, $j = 1, \ldots, s$, are arbitrary smooth functions corresponding to each predictor Z_j, $j = 1, \ldots, s$, such that $\mathrm{E}(f_j(Z_j)) = 0$.

Under certain smoothness assumptions, the functions f_j, $j = 1, \ldots, s$, can be expanded in a convenient basis up to a sufficient number of terms. Specifically, we choose to represent each

of the functions in terms of a B-spline basis as

$$f_j(Z_j) = \sum_{l=1}^{m_j} \alpha_{jl} B_{jl}(Z_j),\ j = 1, \ldots, s, \tag{2.2}$$

where $B_{jl}(Z_j)$ denotes the l^{th} component of the B-spline basis vector for Z_j. The number of terms m_j act as the tuning parameters, and can be selected by using cross-validation. Let $m_0 = \sum_{j=1}^{s} m_j$. Thus, the model in (2.1) can be written as

$$\eta = \boldsymbol{X}^T \boldsymbol{\beta} + \sum_{j=1}^{s} \sum_{l=1}^{m_j} \alpha_{j,l} B_{j,l}(Z_j). \tag{2.3}$$

Consider n independent observations $Y_1, \ldots, Y_n$ with corresponding predictor variables given by $\boldsymbol{X}_i, \boldsymbol{Z}_i,\ 1 \leq i \leq n$ and link functions $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$. The above model can be written in matrix-vector notation as

$$\boldsymbol{\eta} = \mathbb{X}\boldsymbol{\beta} + \mathbb{Z}\boldsymbol{\alpha}, \tag{2.4}$$

where

$$\underset{n \times p}{\mathbb{X}} = \begin{pmatrix} X_{11} & \cdots & X_{1p} \\ X_{21} & \cdots & X_{2p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{pmatrix}, \tag{2.5}$$

and,

$$\underset{n \times m_0}{\mathbb{Z}} = \begin{pmatrix} B_{11}(Z_{11}) & \cdots & B_{1m_1}(Z_{11}) & \cdots & B_{s1}(Z_{1s}) & \cdots & B_{sm_s}(Z_{1s}) \\ B_{11}(Z_{21}) & \cdots & B_{1m_1}(Z_{21}) & \cdots & B_{s1}(Z_{2s}) & \cdots & B_{sm_s}(Z_{2s}) \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ B_{11}(Z_{n1}) & \cdots & B_{1m_1}(Z_{n1}) & \cdots & Z_{s1}(Z_{ns}) & \cdots & B_{sm_s}(Z_{ns}) \end{pmatrix}. \tag{2.6}$$

We denote the $n \times (p + m_0)$ matrix of covariates as

$$\underset{n\times(p+m_0)}{\boldsymbol{\Psi}} = (\mathbb{X} : \mathbb{Z}). \tag{2.7}$$

The vector of co-efficients is $(\underset{p\times 1}{\boldsymbol{\beta}^T}, \underset{m_0\times 1}{\boldsymbol{\alpha}^T})^T$, where

$$\underset{p\times 1}{\boldsymbol{\beta}} = (\beta_1, \ldots, \beta_p)^T, \tag{2.8}$$

and,

$$\underset{m_0\times 1}{\boldsymbol{\alpha}} = (\underset{m_1\times 1}{\boldsymbol{\alpha}_1^T}, \ldots, \underset{m_s\times 1}{\boldsymbol{\alpha}_s^T})^T = (\alpha_{11}, \ldots, \alpha_{1m_1}, \ldots, \alpha_{s1}, \ldots, \alpha_{sm_s})^T. \tag{2.9}$$

We assume that the true underlying model is actually sparse, and hence we are interested in performing variable selection. We define the indicator vector

$$\underset{(p+s)\times 1}{\boldsymbol{\gamma}} = (\gamma_1^X, \ldots, \gamma_p^X, \gamma_1^Z, \ldots, \gamma_s^Z)^T,$$

where

$$\gamma_j^X = \begin{cases} 1, \text{ if } X_j \text{ is in the model, } 1 \le j \le p, \\ 0, \text{ otherwise} \end{cases}$$

$$\gamma_j^Z = \begin{cases} 1, \text{ if } Z_j \text{ is in the model, } 1 \le j \le s, \\ 0, \text{ otherwise.} \end{cases}$$

We denote the vector of co-efficients selected by $\boldsymbol{\gamma}$ as $(\boldsymbol{\beta}_{\boldsymbol{\gamma}}^T, \boldsymbol{\alpha}_{\boldsymbol{\gamma}}^T)^T$ and the corresponding matrix of covariates as $\boldsymbol{\Psi}_{\boldsymbol{\gamma}} = (\mathbb{X}_{\boldsymbol{\gamma}}, \mathbb{Z}_{\boldsymbol{\gamma}})$.

We put priors on the co-efficients $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ separately. The prior on β_j, $1 \le j \le p$ is

degenerate at 0 or has a Laplace density, depending on $\gamma_j^X = 0$ or $\gamma_j^X = 1$ respectively, that is,

$$p(\beta_j \mid \boldsymbol{\gamma}) = (1 - \gamma_j^X)\mathbb{1}_{\{0\}}(\beta_j) + \gamma_j^X \frac{\lambda}{2} \exp\{-\lambda|\beta_j|\}. \tag{2.10}$$

For the co-efficients corresponding to Z_j, given by $\boldsymbol{\alpha}_j$, $1 \leq j \leq s$, we have point-mass at $\mathbf{0}$ corresponding to $\gamma_j^Z = 0$ and a multivariate Laplace density corresponding to $\gamma_j^Z = 1$, so that the prior density of $\boldsymbol{\alpha}_j$ given $\boldsymbol{\gamma}$, denoted by $p(\boldsymbol{\alpha}_j \mid \boldsymbol{\gamma})$ is given by,

$$p(\boldsymbol{\alpha}_j|\boldsymbol{\gamma}) = (1 - \gamma_j^Z)\mathbb{1}_{\{\mathbf{0}\}}(\boldsymbol{\alpha}_j) + \gamma_j^Z \frac{\Gamma(m_j/2)}{2\pi^{m_j/2}\Gamma(m_j)} \lambda_j^{m_j} \exp\{-\lambda_j\|\boldsymbol{\alpha}_j\|\}; \tag{2.11}$$

here λ and λ_j, $j = 1, \ldots, s$, act as the tuning parameters for the different coefficients. The motivation behind the above choice of priors is to get the group lasso solution as the posterior mode corresponding to the likelihood of the response variable. In the R package for the group lasso, the tuning parameter is of the form $\lambda\sqrt{\mathrm{df}}$, where df is the degrees of freedom of the corresponding predictor. Hence we take the tuning parameters to be λ and $\lambda_j = \lambda\sqrt{m_j}$, $j = 1, \ldots, s$, for the linear and the non-linear predictors respectively.

For a binary vector $\boldsymbol{\gamma}$, the number of 1's in $\boldsymbol{\gamma}$ is denoted by $|\boldsymbol{\gamma}|$. The prior on the variable selection indicator $\boldsymbol{\gamma}$ is given by

$$p(\boldsymbol{\gamma}) \propto \det(\mathbb{X}_{\boldsymbol{\gamma}}^T \mathbb{X}_{\boldsymbol{\gamma}}) \det(\boldsymbol{K}_{\boldsymbol{\gamma}}) q^{|\boldsymbol{\gamma}|} (1 - q)^{(p+s)-|\boldsymbol{\gamma}|}, \tag{2.12}$$

where $q \in (0, 1)$ and $\boldsymbol{K}_{\boldsymbol{\gamma}}$ is formed by Kendall's tau coefficients of the variables in $\boldsymbol{\gamma}$ with non-linear effect, that is, $\boldsymbol{K}_{\boldsymbol{\gamma}} = (\!(\tau(Z_j, Z_l) \colon \gamma_j^Z, \gamma_l^Z = 1)\!)$, where

$$\tau(Z_j, Z_l) = \frac{1}{\binom{n}{2}} \sum \sum_{i<i'} \mathrm{sign}(Z_{ij} - Z_{i'j}) \mathrm{sign}(Z_{il} - Z_{i'l})$$

is the Kendall's tau-coefficient between two variables Z_j and Z_l. The factor $\det(\mathbb{X}_{\boldsymbol{\gamma}}^T\mathbb{X}_{\boldsymbol{\gamma}})$ down-weights models with variable combinations having near linear dependence in the linear part, while $\det(\boldsymbol{K}_{\boldsymbol{\gamma}})$ downweights variable combinations with high degree of monotone associations in the non-linear part. The following result shows that the matrix $\boldsymbol{K} = (\!(\tau(Z_j, Z_l))\!)$, and hence all submatrices $\boldsymbol{K}_{\boldsymbol{\gamma}}$, are non-negative definite. Therefore use of the factor $\det(\boldsymbol{K}_{\boldsymbol{\gamma}})$ is justified in the specification of model prior probabilities.

Lemma 2. *The matrix $\boldsymbol{K} = (\!(\tau(Z_j, Z_l))\!)$ is always non-negative definite.*

Proof. By definition of Kendall's tau coefficient, the $(j, l)^{th}$ element of $\boldsymbol{K}$ is given by

$$\begin{aligned}\tau(Z_j, Z_l) &= \frac{\sum_{i=1}^{n}\sum_{i'=i+1}^{n}\text{sign}(Z_{ij} - Z_{i'j})\text{sign}(Z_{il} - Z_{i'l})}{\binom{n}{2}} \\ &= \frac{\sum_{i=1}^{n}\sum_{i'=1}^{n}\text{sign}(Z_{ij} - Z_{i'j})\text{sign}(Z_{il} - Z_{i'l})}{n(n-1)},\end{aligned}$$

$j, l = 1, 2, \ldots, s$, since $\text{sign}(0) = 0$. Therefore we need to show that for any $a_1, \ldots, a_s$,

$$\sum_{j=1}^{s}\sum_{l=1}^{s} a_j a_l \sum_{i=1}^{n}\sum_{i=1}^{n}\text{sign}(Z_{ij} - Z_{i'j})\text{sign}(Z_{il} - Z_{i'l}) \geq 0.$$

The expression is equal to

$$\begin{aligned}&\sum_{i=1}^{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{s} a_j \text{sign}(Z_{ij} - Z_{i'j})\right)\left(\sum_{l=1}^{s} a_l \text{sign}(Z_{il} - Z_{i'l})\right) \\ &= \sum_{i=1}^{n}\sum_{i'=1}^{n}\left(\sum_{j=1}^{s} a_j \text{sign}(Z_{ij} - Z_{i'j})\right)^2 \geq 0,\end{aligned}$$

completing the proof. □

2.3 Posterior computation

The log likelihood of the parameters is given by

$$l(\boldsymbol{\beta}, \boldsymbol{\alpha}; \boldsymbol{Y}) = \sum_{i=1}^{n} \{a(\eta_i)Y_i + b(\eta_i) + c(Y_i)\}. \tag{2.13}$$

Thus, the joint posterior density $p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\alpha}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma} \mid \boldsymbol{Y})$ for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\alpha}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\gamma}$, given $\boldsymbol{Y}$ is proportional to

$$\begin{aligned} & p(\boldsymbol{Y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\alpha}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}) p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\alpha}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}) p(\boldsymbol{\gamma}) \\ = \;& (1-q)^{p+s} \det(\mathbb{X}_{\boldsymbol{\gamma}}^T \mathbb{X}_{\boldsymbol{\gamma}}) \det(\boldsymbol{K}_{\boldsymbol{\gamma}}) \left(\frac{q}{2(1-q)}\right)^{|\boldsymbol{\gamma}|} \left(\prod_{\{j:\boldsymbol{\gamma}_j^Z=1\}} \frac{\Gamma(m_j/2)\lambda^{m_j}}{\pi^{m_j/2}\Gamma(m_j)} \right) \\ & \times \exp\left\{ l(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\alpha}_{\boldsymbol{\gamma}}; \boldsymbol{Y}) - \lambda \sum_{\{j:\gamma_j^X=1\}} |\beta_j| - \lambda_j \sum_{\{j:\gamma_j^Z=1\}} \|\boldsymbol{\alpha}_j\| \right\}. \end{aligned} \tag{2.14}$$

The marginal posterior probability for model $\boldsymbol{\gamma}$ can be obtained by integrating out $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\alpha}_{\boldsymbol{\gamma}}$, that is,

$$p(\boldsymbol{\gamma}|\boldsymbol{Y}) \propto C_1(\boldsymbol{Y}) C_2(\boldsymbol{\gamma}) \int_{\mathbb{R}^{m_{\boldsymbol{\gamma}}}} \exp\{-h(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\alpha}_{\boldsymbol{\gamma}})\} \, d\boldsymbol{\beta}_{\boldsymbol{\gamma}} d\boldsymbol{\alpha}_{\boldsymbol{\gamma}}, \tag{2.15}$$

with

$$\begin{aligned} m_{\boldsymbol{\gamma}} &= |\boldsymbol{\gamma}^X| + \sum_{\{j:\gamma_j^Z=1\}} m_j, \\ C_1(\boldsymbol{Y}) &= (1-q)^{p+s}, \\ C_2(\boldsymbol{\gamma}) &= \det(\mathbb{X}_{\boldsymbol{\gamma}}^T \mathbb{X}_{\boldsymbol{\gamma}}) \det(\boldsymbol{K}_{\boldsymbol{\gamma}}) \left(\frac{q}{2(1-q)}\right)^{|\boldsymbol{\gamma}|} \lambda^{m_{\boldsymbol{\gamma}}} \left(\prod_{\{j:\boldsymbol{\gamma}_j^Z=1\}} \frac{\Gamma(m_j/2)}{\pi^{m_j/2}\Gamma(m_j)} \right), \\ h(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\alpha}_{\boldsymbol{\gamma}}) &= -l(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\alpha}_{\boldsymbol{\gamma}}; \boldsymbol{Y}) + \lambda \sum_{\{j:\gamma_j^X=1\}} |\beta_j| + \lambda_j \sum_{\{j:\gamma_j^Z=1\}} \|\boldsymbol{\alpha}_j\|. \end{aligned} \tag{2.16}$$

The integral in (2.15) can be approximated using the Laplace's approximation. Let $(\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{T*}, \boldsymbol{\alpha}_{\boldsymbol{\gamma}}^{T*})^T$ denote the group lasso solution, that is,

$$(\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{T*}, \boldsymbol{\alpha}_{\boldsymbol{\gamma}}^{T*})^T = \underset{\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\alpha}_{\boldsymbol{\gamma}}}{\operatorname{argmin}}\, h(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\alpha}_{\boldsymbol{\gamma}}). \tag{2.17}$$

Put $\boldsymbol{u} = (\boldsymbol{u}_{\boldsymbol{\beta}}^T, \boldsymbol{u}_{\boldsymbol{\alpha}}^T) = (\boldsymbol{\beta}_{\boldsymbol{\gamma}}^T, \boldsymbol{\alpha}_{\boldsymbol{\gamma}}^T)^T - (\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{T*}, \boldsymbol{\alpha}_{\boldsymbol{\gamma}}^{T*})^T$. Substituting this quantity into (2.15) gives the expression

$$C_1(\boldsymbol{Y})C_2(\boldsymbol{\gamma}) \exp\left\{-h(\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{*}, \boldsymbol{\alpha}_{\boldsymbol{\gamma}}^{*})\right\} \int_{\mathbb{R}^{m_{\boldsymbol{\gamma}}}} \exp\left\{-f(\boldsymbol{u})\right\} d\boldsymbol{u}, \tag{2.18}$$

where,

$$\begin{aligned} f(\boldsymbol{u}) \;=\; & -\{l(\boldsymbol{u}_{\boldsymbol{\beta}} + \boldsymbol{\beta}_{\boldsymbol{\gamma}}^{*}, \boldsymbol{u}_{\boldsymbol{\alpha}} + \boldsymbol{\alpha}_{\boldsymbol{\gamma}}^{*}; \boldsymbol{Y}) - l(\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{*}, \boldsymbol{\alpha}_{\boldsymbol{\gamma}}^{*}; \boldsymbol{Y})\} \\ & + \lambda \sum_{\{j:\gamma_j^X = 1\}} \{|u_{\beta,j} + \beta_j^*| - |\beta_j^*|\} \\ & + \lambda_j \sum_{\{j:\gamma_j^Z = 1\}} \{\|\boldsymbol{u}_{\alpha,j} + \boldsymbol{\alpha}_j^*\| - \|\boldsymbol{\alpha}_j^*\|\}. \end{aligned} \tag{2.19}$$

Clearly $f(\boldsymbol{u})$ is minimized at $\boldsymbol{u} = \mathbf{0}$ by definition. For the Laplace approximation we need to compute the Hessian of the function $f(\boldsymbol{u})$. We have,

$$\begin{aligned} \frac{\partial l(\boldsymbol{u}_{\boldsymbol{\beta}} + \boldsymbol{\beta}_{\boldsymbol{\gamma}}^{*}, \boldsymbol{u}_{\boldsymbol{\alpha}} + \boldsymbol{\alpha}_{\boldsymbol{\gamma}}^{*}; \boldsymbol{Y})}{\partial \boldsymbol{u}} \;&=\; \frac{\partial}{\partial \boldsymbol{u}} \sum_{i=1}^{n} \{a(\eta_i)Y_i + b(\eta_i) + c(Y_i)\} \\ &=\; \sum_{i=1}^{n} \left\{ a'(\eta_i) \frac{\partial \eta_i}{\partial \boldsymbol{u}} Y_i + b'(\eta_i) \frac{\partial \eta_i}{\partial \boldsymbol{u}} \right\} \\ &=\; \sum_{i=1}^{n} \{a'(\eta_i)Y_i + b'(\eta_i)\} \boldsymbol{\Psi}_i, \end{aligned}$$

$\boldsymbol{\Psi}_i$ being the i^{th} row of $\boldsymbol{\Psi}$. Differentiating the above expression again gives,

$$\frac{\partial^2 l(\boldsymbol{u}_{\boldsymbol{\beta}} + \boldsymbol{\beta}^*_{\boldsymbol{\gamma}}, \boldsymbol{u}_{\boldsymbol{\alpha}} + \boldsymbol{\alpha}^*_{\boldsymbol{\gamma}}; \boldsymbol{Y})}{\partial \boldsymbol{u} \partial \boldsymbol{u}^T} = \sum_{i=1}^{n} \{a''(\eta_i) Y_i + b''(\eta_i)\} \boldsymbol{\Psi}_i^T \boldsymbol{\Psi}_i. \tag{2.20}$$

At $\boldsymbol{u} = \mathbf{0}$, we have,

$$\left.\frac{\partial^2 f(\boldsymbol{u})}{\partial \boldsymbol{u} \partial \boldsymbol{u}^T}\right|_{\boldsymbol{u}=\mathbf{0}} = -\sum_{i=1}^{n} \{a''(\eta_i) Y_i + b''(\eta_i)\} \boldsymbol{\Psi}_i^T \boldsymbol{\Psi}_i + \lambda \boldsymbol{A}_{\boldsymbol{\gamma}}, \tag{2.21}$$

where the $m_{\boldsymbol{\gamma}} \times m_{\boldsymbol{\gamma}}$ matrix $\boldsymbol{A}_{\boldsymbol{\gamma}}$ is given by

$$\boldsymbol{A}_{\boldsymbol{\gamma}} = \begin{bmatrix} \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{A}^Z_{\boldsymbol{\gamma}} \end{bmatrix}, \tag{2.22}$$

$\boldsymbol{O}$ stands for a zero matrix of appropriate order, and $\boldsymbol{A}^Z_{\boldsymbol{\gamma}}$ is given by

$$\begin{bmatrix} \boldsymbol{D}_1 & \boldsymbol{O}_{12} & \cdots & \boldsymbol{O}_{1t} \\ \boldsymbol{O}_{21} & \boldsymbol{D}_2 & \cdots & \boldsymbol{O}_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{O}_{t1} & \boldsymbol{O}_{t2} & \cdots & \boldsymbol{D}_t \end{bmatrix}, \tag{2.23}$$

where $\boldsymbol{D}_j = \sqrt{m_j} \left(-\frac{\boldsymbol{\alpha}^*_j \boldsymbol{\alpha}^{*T}_j}{\|\boldsymbol{\alpha}^*_j\|^3} + \frac{\boldsymbol{I}_{m_j - 1}}{\|\boldsymbol{\alpha}^*_j\|} \right)$, $j = 1, \ldots, t (= |\boldsymbol{\gamma}^Z|)$, and $\boldsymbol{O}_{ij}$ is the zero matrix corresponding to the variables in the ith row and jth column selected by $\boldsymbol{\gamma}^Z$.

The above equations can be used to apply Laplace approximation to the quantity in (2.18),

which gives

$$\begin{aligned} p(\boldsymbol{\gamma}|\boldsymbol{Y}) &\propto C_1(\boldsymbol{Y})C_2(\boldsymbol{\gamma})\exp\left\{-h(\boldsymbol{\beta}^*_{\boldsymbol{\gamma}},\boldsymbol{\alpha}^*_{\boldsymbol{\gamma}})\right\} \\ &\quad \times \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp\left\{-f(\boldsymbol{u})\right\}d\boldsymbol{u} \\ &\approx C_1(\boldsymbol{Y})C_2(\boldsymbol{\gamma})\exp\left\{-h(\boldsymbol{\beta}^*_{\boldsymbol{\gamma}},\boldsymbol{\alpha}^*_{\boldsymbol{\gamma}})\right\} \\ &\quad \times \exp\left\{-f(\boldsymbol{0})\right\}(2\pi)^{m_{\boldsymbol{\gamma}}/2}\left|\frac{\partial^2 f(\boldsymbol{0})}{\partial\boldsymbol{u}\partial\boldsymbol{u}^T}\right|^{-1/2}. \end{aligned}$$

The accuracy of the approximation is controlled by the dimension and the posterior convergence rate. Substituting (2.21) in the above quantity, the marginal posterior probability $p(\boldsymbol{\gamma}|\boldsymbol{Y})$ for $\boldsymbol{\gamma}$ is approximately proportional to

$$\begin{aligned} Q(\boldsymbol{\gamma}|\boldsymbol{Y}) &= C_1(\boldsymbol{Y})C_2(\boldsymbol{\gamma})\exp\left\{-h(\boldsymbol{\beta}^*_{\boldsymbol{\gamma}},\boldsymbol{\alpha}^*_{\boldsymbol{\gamma}})\right\} \\ &\quad \times(2\pi)^{m_{\boldsymbol{\gamma}}/2}\left|-\sum_{i=1}^{n}\{a''(\eta_i)Y_i+b''(\eta_i)\}\boldsymbol{\Psi}_i^T\boldsymbol{\Psi}_i+\lambda\boldsymbol{A}_{\boldsymbol{\gamma}}\right|^{-1/2}. \end{aligned} \tag{2.24}$$

The approximation above is valid only for regular models, that is, models indicated by $\boldsymbol{\gamma}$ for which the function $f(\boldsymbol{u})$ is differentiable at $\boldsymbol{u}=\boldsymbol{0}$. This holds only if the group lasso solution for all the coefficients corresponding to the predictors included in the model as indicated by $\boldsymbol{\gamma}$ are non-zero. If the group lasso solution is zero for a subset of variables already in the model, then the corresponding model is termed as a non-regular model, and posterior probability cannot be approximated using the Laplace approximation. This issue of non-regularity also arises in variable selection for linear models and non-parametric additive regression models; see Yuan and Lin (2005) and Curtis et al. (2014). The corresponding regular model for a non-regular model is obtained by excluding the variables for which the group lasso solution is zero in that non-regular model. When q is chosen to be small, the posterior probability of any non-regular model can be shown to be negligible compared to the corresponding regular model.

Therefore it suffices to compute the model posterior probabilities of the regular models only at least, if one wants to obtain the maximum posterior probability model. The approximate posterior probabilities may be re-normalized taking only the regular models into account.

2.3.1 Estimation of λ

The joint density of the response and the regression coefficient vector given the other model parameters is given by

$$\begin{aligned} p(\boldsymbol{Y}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\alpha}_{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma}, \lambda) \quad &= \quad \left(\frac{1}{2}\right)^{|\boldsymbol{\gamma}|} \lambda^{m_{\boldsymbol{\gamma}}} \left(\prod_{\{j:\gamma_j^Z=1\}} \frac{\Gamma(m_j/2)}{\pi^{m_j/2}\Gamma(m_j)} \right) \\ &\quad \times \exp\left\{-h(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\alpha}_{\boldsymbol{\gamma}})\right\}. \end{aligned} \tag{2.25}$$

Integrating out $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\alpha}_{\boldsymbol{\gamma}}$, and using Laplace approximation as before, we get,

$$\begin{aligned} p(\boldsymbol{Y} \mid \boldsymbol{\gamma}, \lambda) \quad &\approx \quad 2^{-(2|\boldsymbol{\gamma}|-m_{\boldsymbol{\gamma}})/2} \pi^{|\boldsymbol{\gamma}^X|} \lambda^{m_{\boldsymbol{\gamma}}} \\ &\quad \times \left(\prod_{\{j:\gamma_j^Z=1\}} \frac{\Gamma(m_j/2)}{\Gamma(m_j)} \right) \exp\left\{-h(\boldsymbol{\beta}^*_{\boldsymbol{\gamma}}, \boldsymbol{\alpha}^*_{\boldsymbol{\gamma}})\right\} \\ &\quad \times \left| -\sum_{i=1}^{n} \{a''(\eta_i) Y_i + b''(\eta_i)\} \boldsymbol{\Psi}_i^T \boldsymbol{\Psi}_i + \lambda \boldsymbol{A}_{\boldsymbol{\gamma}} \right|^{-1/2}. \end{aligned} \tag{2.26}$$

Similar to the penalized maximum likelihood criterion as in Curtis et al. (2014), we can choose λ by penalizing negative 2 times the log-likelihood. Taking negative 2 times the loga-

rithm of (2.26) gives,

$$\begin{aligned} r(\lambda) \quad &= \quad (2|\boldsymbol{\gamma}| - m_{\boldsymbol{\gamma}}) \log 2 - 2|\boldsymbol{\gamma}^X| \log \pi - 2m_{\boldsymbol{\gamma}} \log \lambda \\ &\quad -2 \sum_{\{j:\boldsymbol{\gamma}^Z=1\}} \{\log \Gamma(m_j/2) - \log \Gamma(m_j)\} + 2h(\boldsymbol{\beta}^*_{\boldsymbol{\gamma}}, \boldsymbol{\alpha}^*_{\boldsymbol{\gamma}}) \\ &\quad + \log \left| - \sum_{i=1}^{n} \{a''(\eta_i) Y_i + b''(\eta_i)\} \boldsymbol{\Psi}_i^T \boldsymbol{\Psi}_i + \lambda \boldsymbol{A}_{\boldsymbol{\gamma}} \right|. \end{aligned} \tag{2.27}$$

We estimate λ by minimizing

$$r(\lambda) + m_{\hat{\gamma}_\lambda} \log n. \tag{2.28}$$

2.3.2 Examples of GAPLM

The expression for the approximate posterior probabilities of various models can be derived explicitly for a given exponential family of distributions and the corresponding link function. We discuss some of the densities which are widely used in this context and derive the form of the Hessian required for the Laplace approximation of the posterior probabilities.

1. *Logistic regression*: The mean of the distribution is $\mu = \exp(\eta)/(1 + \exp(\eta))$, and the response is $Y \in \{0, 1\}$. The likelihood is given by $\exp\{\eta y - \log(1 + \exp(\eta))\}$, so that $a(\eta) = \eta, b(\eta) = -\log(1 + \exp(\eta))$. Thus, the expression of the Hessian as in (2.21) becomes,

$$\left. \frac{\partial^2 f(\boldsymbol{u})}{\partial \boldsymbol{u} \partial \boldsymbol{u}^T} \right|_{\boldsymbol{u}=\boldsymbol{0}} = \sum_{i=1}^{n} \left\{ \frac{\exp(\eta_i)}{(1 + \exp(\eta_i))^2} \right\} \boldsymbol{\Psi}_i^T \boldsymbol{\Psi}_i + \lambda \boldsymbol{A}_{\boldsymbol{\gamma}}. \tag{2.29}$$

2. *Poisson regression*: The response is $Y \in \{0, 1, 2, \ldots\}$ and the mean is given by $\mu = \exp(\eta)$. The likelihood is $\exp\{\eta y - \exp(\eta) - \log(y!)\}$. Here $a(\eta) = \eta, b(\eta) = -\exp(\eta)$.

Thus, expression (2.21) gives,

$$\left.\frac{\partial^2 f(\boldsymbol{u})}{\partial \boldsymbol{u}\partial \boldsymbol{u}^T}\right|_{\boldsymbol{u}=\boldsymbol{0}} = \sum_{i=1}^{n} \exp(\eta_i)\boldsymbol{\Psi}_i^T\boldsymbol{\Psi}_i + \lambda \boldsymbol{A}_{\boldsymbol{\gamma}}. \tag{2.30}$$

3. *Probit regression*: The mean of the distribution is given by the distribution function of a standard normal distribution, that is, $\mu = \Phi(\eta)$, and the response is $Y \in \{0, 1\}$. The likelihood is given by

$$\exp\left\{ y \log\left(\frac{\Phi(\eta)}{1-\Phi(\eta)}\right) + \log(1-\Phi(\eta))\right\}.$$

Thus, $a(\eta) = \log(\Phi(\eta)/(1-\Phi(\eta))), b(\eta) = \log(1-\Phi(\eta))$. The expression for (2.21) can be obtained by evaluating the second derivatives of $a(\eta)$ and $b(\eta)$, given by

$$\begin{aligned} a''(\eta) &= \frac{\Phi''(\eta)\Phi(\eta) - (\Phi'(\eta))^2}{(\Phi(\eta))^2} + \frac{\Phi''(\eta)(1-\Phi(\eta)) + (\Phi'(\eta))^2}{(1-\Phi(\eta))^2}, \\ b''(\eta) &= -\frac{\Phi''(\eta)(1-\Phi(\eta)) + (\Phi'(\eta))^2}{(1-\Phi(\eta))^2}. \end{aligned} \tag{2.31}$$

4. *Exponential regression*: For an exponential regression with exponential link, the mean of the distribution is $\mu = \exp(\eta)$, and response $Y \in (0, \infty)$. The likelihood is given by $\exp(-\exp(-\eta)y - \eta)$, so that $a(\eta) = -\exp(-\eta), b(\eta) = -\eta$. Thus, the expression of (2.21) is given by

$$\left.\frac{\partial^2 f(\boldsymbol{u})}{\partial \boldsymbol{u}\partial \boldsymbol{u}^T}\right|_{\boldsymbol{u}=\boldsymbol{0}} = \sum_{i=1}^{n} \{\exp(-\eta_i)Y_i\}\boldsymbol{\Psi}_i^T\boldsymbol{\Psi}_i + \lambda \boldsymbol{A}_{\boldsymbol{\gamma}}. \tag{2.32}$$

5. *Normal linear regression*: For the normal linear regression set-up, we have the special case of additive partial linear models. Assuming variance $\sigma^2 = \tau^{-1}$ to be known, the

mean of the distribution is given by $\mu = \eta$ for response $Y \in \mathbb{R}$. The likelihood is given by $\exp(\tau\eta y - \tau\eta^2/2 - \tau y^2/2 - \log(2\pi\tau)/2)$. Thus $a(\eta) = \tau\eta, b(\eta) = -\tau\eta^2/2$. Thus, the expression in (2.21) becomes

$$\left.\frac{\partial^2 f(\boldsymbol{u})}{\partial \boldsymbol{u} \partial \boldsymbol{u}^T}\right|_{\boldsymbol{u}=\boldsymbol{0}} = \tau \sum_{i=1}^{n} \boldsymbol{\Psi}_i^T \boldsymbol{\Psi}_i + \lambda \boldsymbol{A}_{\boldsymbol{\gamma}}. \tag{2.33}$$

When the variance is unknown, we can modify the prior parameters λ and λ_j in the prior specification of the coefficients in order to obtain the group lasso estimate as the posterior mode. We discuss this case in the next section.

2.4 Additive partial linear models

Substituting the prior parameters λ and λ_j by $\lambda/2\sigma^2$ and $\lambda_j/2\sigma^2$, $j = 1, \ldots, s$, respectively in the prior of the coefficients as in equations (2.10) and (2.11), the expression of the approximate posterior probability of a model $\boldsymbol{\gamma}$ given the data can be obtained as

$$\begin{aligned} Q(\boldsymbol{\gamma}|\boldsymbol{Y}) &= C_1(\boldsymbol{Y}) C_2(\boldsymbol{\gamma}) \exp\left\{-\frac{1}{2\sigma^2} h(\boldsymbol{\beta}_{\boldsymbol{\gamma}}^*, \boldsymbol{\eta}_{\boldsymbol{\gamma}}^*)\right\} \\ &\quad \times (2\pi)^{m_{\boldsymbol{\gamma}}/2} \left|\frac{1}{\sigma^2}(\boldsymbol{\Psi}_{\boldsymbol{\gamma}}^T \boldsymbol{\Psi}_{\boldsymbol{\gamma}} + \frac{\lambda}{2} \boldsymbol{A}_{\boldsymbol{\gamma}})\right|^{-1/2}, \end{aligned} \tag{2.34}$$

where

$$
\begin{aligned}
m_{\boldsymbol{\gamma}} &= |\boldsymbol{\gamma}^X| + \sum_{\{j:\gamma_j^Z=1\}} m_j, \\
C_1(\boldsymbol{Y}) &= (1-q)^{p+s}(2\pi\sigma^2)^{-n/2}, \\
C_2(\boldsymbol{\gamma}) &= \det(\mathbb{X}_{\boldsymbol{\gamma}}^T\mathbb{X}_{\boldsymbol{\gamma}})\det(\boldsymbol{K}_{\boldsymbol{\gamma}}) \left(\frac{q}{2(1-q)}\right)^{|\boldsymbol{\gamma}|} \left(\frac{\lambda}{2\sigma^2}\right)^{m_{\boldsymbol{\gamma}}} \left(\prod_{\{j:\boldsymbol{\gamma}_j^Z=1\}} \frac{\Gamma(m_j/2)}{\pi^{m_j/2}\Gamma(m_j)}\right), \\
h(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\alpha}_{\boldsymbol{\gamma}}) &= \|\boldsymbol{Y} - \mathbb{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}} - \mathbb{Z}_{\boldsymbol{\gamma}}\boldsymbol{\alpha}_{\boldsymbol{\gamma}}\|^2 + \lambda \sum_{\{\gamma_j^X=1:1\le j\le p\}} |\beta_j| + \lambda\sqrt{m_j} \sum_{j:\gamma_j^Z=1} \|\boldsymbol{\alpha}_j\|.
\end{aligned}
\tag{2.35}
$$

2.4.1 Estimation of σ^2 in additive partial linear models

Similar to (2.25) and (2.26), the Laplace approximation to the posterior density of the response given the model indicator $\boldsymbol{\gamma}$ and the parameters λ and σ^2 is given by

$$
\begin{aligned}
p(\boldsymbol{Y} \mid \boldsymbol{\gamma}, \lambda, \sigma^2) \;\approx\; & (2)^{-(n-m_{\boldsymbol{\gamma}})/2}\pi^{|\boldsymbol{\gamma}^X|}\sigma^{-(n+m_{\boldsymbol{\gamma}})}2^{-(m_{\boldsymbol{\gamma}}+|\boldsymbol{\gamma}|)}\lambda^{m_{\boldsymbol{\gamma}}} \left(\prod_{\{j:\boldsymbol{\gamma}_j^Z=1\}} \frac{\Gamma(m_j/2)}{\Gamma(m_j)}\right) \\
& \times \exp\left\{-\frac{1}{2\sigma^2}h(\boldsymbol{\beta}_{\boldsymbol{\gamma}}^*, \boldsymbol{\eta}_{\boldsymbol{\gamma}}^*)\right\} \left|(\boldsymbol{\Psi}_{\boldsymbol{\gamma}}^T\boldsymbol{\Psi}_{\boldsymbol{\gamma}} + \frac{\lambda}{2}\boldsymbol{A}_{\boldsymbol{\gamma}})\right|^{-1/2}.
\end{aligned}
\tag{2.36}
$$

Denoting the model chosen by the group lasso solution for a fixed λ to be $\hat{\boldsymbol{\gamma}}_\lambda$, then maximizing (2.36) with respect to σ^2, an estimate of σ^2 is given by

$$
\hat{\sigma}_\lambda^2 = h(\boldsymbol{\beta}_{\hat{\boldsymbol{\gamma}}_\lambda}^*, \boldsymbol{\alpha}_{\hat{\boldsymbol{\gamma}}_\lambda}^*)/(n + m_{\boldsymbol{\gamma}}). \tag{2.37}
$$

We can choose λ by penalizing negative 2 times the log-likelihood similar to the general case, but now plugging in the estimate of σ^2 from equation (2.37) in equation (2.36).

2.5 Numerical study

2.5.1 Simulation study

We perform two simulation studies for evaluating the performance of our method. We generate 100 data sets with sample sizes $n = 100, 200, 500$ from the data generating processes

1. Logistic regression model

$$\text{logit}\{\text{P}(Y = 1)\} = \log \frac{\text{P}(Y = 1)}{1 - \text{P}(Y = 1)} = \sum_{j=1}^{p} \beta_j X_j + \sum_{j=1}^{s} f_j(Z_j), \tag{2.38}$$

2. Additive partial linear model

$$Y = \sum_{j=1}^{p} \beta_j X_j + \sum_{j=1}^{s} f_j(Z_j) + \varepsilon,\ \varepsilon \sim \text{N}(0, 1), \tag{2.39}$$

where

$$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T = (1, 1.5, 0, \ldots, 0)^T, \tag{2.40}$$

and

$$\begin{aligned} f_1(x) &= 2\sin(\pi x), \\ f_2(x) &= 2/(1 + \exp(-10x)), \\ f_j(x) &= 0, \text{ for all } j = 3, \ldots, s. \end{aligned} \tag{2.41}$$

For each of the two data generating processes, we perform two simulation studies, one in lower dimensions and the other in a high dimensional setting. We specify the total number of $\boldsymbol{X}$ and $\boldsymbol{Z}$ variables to be $p = 10$ and $s = 10$ respectively for the lower dimensional example, and

Table 2.1: Table corresponding to independent predictors, $p = 10,\ s = 10$ in GAPLM (logistic regression). Figures in parentheses represent respective standard errors.

	n	SP	SE	MCC	MCR
HPPM	100	0.820 (0.015)	0.778 (0.020)	0.602 (0.030)	0.093 (0.003)
MPM	100	0.785 (0.017)	0.802 (0.018)	0.594 (0.027)	0.092 (0.003)
GL	100	0.480 (0.018)	0.912 (0.019)	0.417 (0.020)	0.088 (0.003)
BMA	100	–	–	–	0.093 (0.003)
HPPM	200	0.700 (0.018)	0.878 (0.015)	0.577 (0.027)	0.103 (0.002)
MPM	200	0.677 (0.016)	0.920 (0.016)	0.601 (0.022)	0.104 (0.002)
GL	200	0.647 (0.016)	0.938 (0.016)	0.592 (0.020)	0.104 (0.002)
BMA	200	–	–	–	0.106 (0.002)
HPPM	500	0.672 (0.014)	0.968 (0.008)	0.649 (0.014)	0.104 (0.001)
MPM	500	0.621 (0.015)	0.985 (0.005)	0.636 (0.014)	0.104 (0.001)
GL	500	0.600 (0.012)	0.990 (0.004)	0.630 (0.014)	0.104 (0.001)
BMA	500	–	–	–	0.104 (0.001)

$p = 100$, $s = 100$ for the high dimensional example. The $\boldsymbol{X}$ variables are drawn independently from a standard normal distribution and $\boldsymbol{Z}$ consists of independent standard uniform variates. Corresponding to each data sets, we find the approximate posterior probabilities of various models using the Bayesian method. From the approximate computations, we find the model with the highest posterior probability and also the median probability model, the model having those predictors whose inclusion probability is at least 0.5. To assess the variable selection performance of the models selected by the methods, we compute the specificity (SP), sensitivity (SE) and Mathew's correlation coefficient (MCC) averaged across all the replications. The expressions of these performance measures, in terms of the true positives (TP), true negatives

Table 2.2: Table corresponding to independent predictors, $p = 10$, $s = 10$ in GAPLM with misspecification. Figures in parentheses represent respective standard errors.

	n	SP	SE	MCC	MCR
HPPM	100	0.760 (0.014)	0.772 (0.019)	0.530 (0.028)	0.054 (0.002)
MPM	100	0.713 (0.015)	0.822 (0.020)	0.534 (0.026)	0.054 (0.002)
GL	100	0.457 (0.019)	0.932 (0.012)	0.500 (0.019)	0.052 (0.002)
BMA	100	–	–	–	0.057 (0.002)
HPPM	200	0.672 (0.015)	0.852 (0.015)	0.522 (0.025)	0.067 (0.001)
MPM	200	0.635 (0.016)	0.895 (0.015)	0.537 (0.022)	0.067 (0.002)
GL	200	0.620 (0.014)	0.900 (0.015)	0.527 (0.021)	0.067 (0.002)
BMA	200	–	–	–	0.068 (0.002)
HPPM	500	0.658 (0.014)	0.968 (0.008)	0.631 (0.016)	0.068 (0.001)
MPM	500	0.630 (0.015)	0.978 (0.007)	0.620 (0.015)	0.068 (0.001)
GL	500	0.592 (0.013)	0.980 (0.007)	0.589 (0.013)	0.068 (0.001)
BMA	500	–	–	–	0.068 (0.001)

(TN), false positives (FP) and false negatives (FN) in the respective models are given by

$$\begin{aligned} \text{SP} &= \frac{\text{TN}}{\text{TN}+\text{FP}}, \text{SE} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \\ \text{MCC} &= \frac{\text{TN}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}}. \end{aligned}$$

In case of the logistic regression model, we also find the misclassification rate for each method averaged over the replications (denoted by 'MCR'), and for the additive partial linear model, we evaluate the mean prediction error for each method averaged over replications (denoted by 'PE'). The same information are also recorded for the model selected by the group lasso. The simulation results are tabulated in Table 2.1–2.6. We denote the model with highest posterior probability, the median probability model, and the model selected by the group lasso as 'HPPM', 'MPM' and 'GL' respectively. In order to account for model uncertainty, we also perform Bayesian model averaging and find the corresponding misclassification rate or

Table 2.3: Table corresponding to independent predictors, $p = 100$, $s = 100$ in GAPLM (logistic regression). Figures in parentheses represent respective standard errors.

	n	SP	SE	MCC	MCR
HPPM	100	0.946 (0.002)	0.352 (0.020)	0.238 (0.015)	0.120 (0.005)
MPM	100	0.943 (0.002)	0.358 (0.020)	0.234 (0.015)	0.122 (0.006)
GL	100	0.939 (0.002)	0.373 (0.020)	0.236 (0.015)	0.120 (0.005)
BMA	100	–	–	–	0.122 (0.006)
HPPM	200	0.952 (0.002)	0.488 (0.021)	0.356 (0.017)	0.103 (0.003)
MPM	200	0.946 (0.002)	0.500 (0.021)	0.348 (0.017)	0.104 (0.003)
GL	200	0.945 (0.002)	0.512 (0.021)	0.348 (0.016)	0.104 (0.003)
BMA	200	–	–	–	0.105 (0.002)
HPPM	500	0.947 (0.002)	0.610 (0.019)	0.425 (0.016)	0.105 (0.001)
MPM	500	0.942 (0.002)	0.670 (0.020)	0.445 (0.015)	0.105 (0.001)
GL	500	0.940 (0.002)	0.647 (0.020)	0.433 (0.015)	0.105 (0.001)
BMA	500	–	–	–	0.105 (0.001)

prediction error (listed in the row named 'BMA') respectively for the two examples.

For the logistic regression example, in order to evaluate the performance of the methods under a mis-specified model, we run simulations using the same set-up, but now using a probit link to generate the responses using the same linear functional form as above. Corresponding results are tabulated for comparing with that obtained using the true link function.

As we can see from the simulation results, the performance of the Bayesian method is superior compared with the group lasso in terms of the measures used for model assessment. The performance of the methods deteriorates with increasing dimensions as expected, but improves with increasing sample sizes. For the mis-specified model, all methods perform slightly worse compared with the correctly specified situation, but the overall performance of the Bayesian method is better than the group lasso. In the special case of additive partial linear models, the prediction accuracy obtained by Bayesian model averaging is much better compared to the group lasso.

Table 2.4: Table corresponding to independent predictors, $p = 100,\ s = 100$ in GAPLM with misspecification. Figures in parentheses represent respective standard errors.

	n	SP	SE	MCC	MCR
HPPM	100	0.949 (0.002)	0.338 (0.020)	0.232 (0.017)	0.066 (0.003)
MPM	100	0.943 (0.002)	0.357 (0.020)	0.233 (0.017)	0.066 (0.003)
GL	100	0.940 (0.002)	0.362 (0.020)	0.230 (0.017)	0.066 (0.003)
BMA	100	–	–	–	0.067 (0.003)
HPPM	200	0.949 (0.002)	0.440 (0.019)	0.307 (0.017)	0.070 (0.002)
MPM	200	0.943 (0.002)	0.470 (0.023)	0.309 (0.017)	0.070 (0.002)
GL	200	0.941 (0.002)	0.472 (0.023)	0.307 (0.017)	0.070 (0.002)
BMA	200	–	–	–	0.070 (0.002)
HPPM	500	0.937 (0.002)	0.615 (0.022)	0.394 (0.017)	0.072 (0.001)
MPM	500	0.930 (0.002)	0.640 (0.021)	0.397 (0.018)	0.072 (0.001)
GL	500	0.930 (0.002)	0.650 (0.021)	0.392 (0.016)	0.072 (0.001)
BMA	500	–	–	–	0.072 (0.001)

2.5.2 Pima Indian Diabetes study

We apply our Bayesian variable selection method to a data set corresponding to the Pima Indian diabetes study. This data set is originally owned by National Institute of Diabetes and Digestive and Kidney Diseases, and currently available in the UCI Machine Learning Repository. The response variable is a binary variable indicating whether a patient under study tested positive for diabetes, for a total of 768 patients. All the patients under study are females of Pima Indian heritage and are at least 21 years old. Corresponding to each patient, we have the following predictor variables:

1. NumPreg: Number of times pregnant
2. PGC: Plasma glucose concentration at 2 hours in an oral glucose tolerance test
3. DBP: Diastolic blood pressure (mm Hg)

Table 2.5: Table corresponding to independent predictors, $p = 10,\ s = 10$ in APLM. Figures in parentheses represent respective standard errors.

	n	SP	SE	MCC	PE
HPPM	100	0.695 (0.017)	0.995 (0.003)	0.696 (0.015)	1.752 (0.020)
MPM	100	0.685 (0.017)	0.998 (0.002)	0.689 (0.015)	1.752 (0.020)
GL	100	0.568 (0.017)	1.000 (0.000)	0.593 (0.014)	1.750 (0.020)
BMA	100	–	–	–	1.018 (0.044)
HPPM	200	0.637 (0.016)	1.000 (0.000)	0.649 (0.013)	1.401 (0.016)
MPM	200	0.643 (0.015)	1.000 (0.000)	0.654 (0.013)	1.401 (0.016)
GL	200	0.513 (0.017)	1.000 (0.000)	0.548 (0.014)	1.400 (0.016)
BMA	200	–	–	–	0.992 (0.034)
HPPM	500	0.613 (0.012)	1.000 (0.000)	0.627 (0.010)	1.430 (0.007)
MPM	500	0.607 (0.012)	1.000 (0.000)	0.621 (0.010)	1.430 (0.007)
GL	500	0.577 (0.011)	1.000 (0.000)	0.597 (0.009)	1.430 (0.007)
BMA	500	–	–	–	1.304 (0.017)

4. TSFT: Triceps skin fold thickness (mm)

5. Serum: 2-Hour serum insulin (mu U/ml)

6. BMI: Body mass index (weight in kg/$(\text{height in m})^2$)

7. DPF: Diabetes pedigree function

8. Age: Age (years)

These predictors have been believed to be significant risk factors for diabetes among Pimas or other populations. The effect of the predictors 'Age' and 'BMI' are suspected to be non-linear. We carry out or Bayesian variable selection technique to the data set using a GAPLM model with 'Age' and 'BMI' as the variables with additive non-parametric effect and all the remaining predictors having only linear effects. The group lasso selects all the predictors but the Bayesian method selects the linear effects of 'NumPreg', 'PGC', 'DBP' and 'DPF' and

Table 2.6: Table corresponding to independent predictors, $p = 100$, $s = 100$ in APLM. Figures in parentheses represent respective standard errors.

	n	SP	SE	MCC	PE
HPPM	100	0.949 (0.002)	0.617 (0.016)	0.435 (0.014)	6.235 (0.066)
MPM	100	0.944 (0.002)	0.645 (0.016)	0.437 (0.014)	6.205 (0.067)
GL	100	0.941 (0.002)	0.652 (0.017)	0.432 (0.014)	6.190 (0.067)
BMA	100	–	–	–	1.608 (0.122)
HPPM	200	0.941 (0.002)	0.825 (0.016)	0.536 (0.014)	4.210 (0.043)
MPM	200	0.935 (0.002)	0.872 (0.015)	0.552 (0.012)	4.183 (0.043)
GL	200	0.934 (0.002)	0.890 (0.015)	0.543 (0.013)	4.172 (0.043)
BMA	200	–	–	–	1.189 (0.089)
HPPM	500	0.911 (0.008)	0.800 (0.033)	0.440 (0.027)	2.539 (0.034)
MPM	500	0.911 (0.008)	1.000 (0.000)	0.547 (0.018)	2.426 (0.034)
GL	500	0.911 (0.008)	1.000 (0.000)	0.547 (0.018)	2.426 (0.034)
BMA	500	–	–	–	0.737 (0.105)

non-linear effects of 'Age' and 'BMI' as the predictors in the median probability model. Table 2.7 shows the marginal posterior inclusion probabilities of the predictors.

2.5.3 Nutritional epidemiology study

There is an increased risk of developing certain types of cancer including lung, colon, breast and prostrate cancer when the blood plasma concentration of beta-carotene is low, as indicated

Table 2.7: Marginal inclusion probabilities of predictors for Pima Indian Diabetes study

Predictor	Inclusion Probability	Predictor	Inclusion Probability
NumPreg	0.843	Serum	0.004
PGC	1.000	BMI	1.000
DBP	0.999	DPF	0.584
TSFT	0.019	Age	0.826

by some epidemiological studies. It is known that beta-carotene has remarkable anti-oxidant properties and regular dietary intake of fruits and vegetables rich in beta-carotene helps the body's auto-immune system to fight cancer. It is of a lot of interest for clinical practitioners to know how the plasma concentrations of beta-carotene depend on certain regulatory factors like age, gender, regular use of vitamins, dietary intake, smoking status, alcohol consumption, etc. A number of diverse results have been found regarding the relation to these factors; for example, see Nierenberg et al. (1989); Faure et al. (2006).

We use the data-set based on a cross-sectional study provided by Therese Stukel of Dartmouth Hitchcock Medical Center, available at `http://lib.stat.cmu.edu/datasets/Plasma_Retinol`. Details of the data can be found in the above link. The response variables are plasma concentrations of beta-carotene and retinol obtained from 315 patients. Observations are made relating to 12 other factors, namely,

1. AGE: Age (years).
2. SEX: Sex (1=Male, 2=Female).
3. SMOKSTAT: Smoking status (1=Never, 2=Former, 3=Current Smoker).
4. BMI: Body Mass Index (weight/height^2).
5. VITUSE: Vitamin Use (1=Yes, fairly often, 2=Yes, not often, 3=No).
6. CALORIES: Number of calories consumed per day.
7. FAT: Grams of fat consumed per day.
8. FIBER: Grams of fiber consumed per day.
9. ALCOHOL: Number of alcoholic drinks consumed per week.

10. CHOL: Cholesterol consumed (mg per day).

11. BETADIET: Dietary beta-carotene consumed (mcg per day).

12. RETDIET: Dietary retinol consumed (mcg per day).

Liu et al. (2011) used an additive partial linear model for this kind of problem and applied their variable selection method for APLM after doing some primary elicitation of effects which may be presumed to be linear and some effects which do not seem to have a linear effect. Similar to their model, we consider an additive partial linear model with 'AGE' and 'CHOL' having non-linear effects and all other predictors having linear effects only. We perform variable selection using the proposed Bayesian method by computing approximate posterior probabilities of various models.

The median probability model selects the linear effects of 'SMOKESTAT', 'BMI', 'VITUSE', 'FAT', 'FIBER' and the non-linear effects of 'CHOL' as the effective set of variables related to the beta-carotene levels. The median probability model is also the maximum a posteriori model. In comparison, the group lasso selects the linear effects of all the predictors excluding 'SEX', and the non-linear effects of 'CHOL'. Table 2.8 lists the marginal inclusion probabilities of the predictors as obtained from our Bayesian procedure.

Table 2.8: Marginal inclusion probabilities of predictors for Nutritional Epidemiology study

Predictor	Inclusion Probability	Predictor	Inclusion Probability
SEX	0.000	FIBER	0.991
SMOKESTAT	0.813	ALCOHOL	0.185
BMI	0.999	BETADIET	0.018
VITUSE	0.891	RETDIET	0.016
CALORIES	0.022	AGE	0.000
FAT	0.957	CHOL	0.700

2.5.4 Prostate cancer data

We consider the prostate cancer data (Stamey et al., 1989) for analysis using the additive partial linear model framework. The data consists of clinical measures of 97 men who were about to receive a radical prostatectomy (see 'lasso2' package in R, Lokhorst et al., 2013). The response variable is log level of prostate specific antigen (lpsa), corresponding to 8 other predictors, namely,

1. lcavol: log cancer volume.
2. lweight: log prostate weight.
3. age: age.
4. lbph: log benign prostatic hyperplasia amount.
5. svi: seminal vesicle invasion.
6. lcp: log capsular penetration.
7. gleason: Gleason score.
8. pgg45: percentage Gleason scores 4 or 5.

The linear regression model using the lasso selects the predictors 'lcavol', 'lweight' and 'svi' as the important linear predictors (Tibshirani, 1996). In our analysis, we consider an additive partial linear model using 'lcavol' and 'lweight' as the predictors with both linear and non-linear effects, and all remaining predictors having strictly linear effect. The median probability model selects 'lbph', 'svi', 'lcp' 'gleason' among the linear effects and non-linear effects of 'lcavol'. The marginal inclusion probabilities of the various predictors are tabulated below.

Table 2.9: Marginal inclusion probabilities of predictors for Prostate cancer data

Predictor	Inclusion Probability	Predictor	Inclusion Probability
age	0.179	gleason	0.992
lbph	0.828	pgg45	0.000
svi	0.994	lcavol	0.949
lcp	0.681	lweight	0.211

In order to further validate our method, we also conducted an analysis with the above data but now including additional 50 junk variables. The Bayesian method correctly selects a subset of the original predictors discarding the additional predictors. The marginal inclusion probabilities of the original predictors are identical with those in Table 2.9.

Chapter 3

Estimating large precision matrices using graphical models

3.1 Introduction

In this chapter, we consider Bayesian estimation of a precision matrix working with a G-Wishart prior induced by a Gaussian graphical model, which has a Markov property with respect to a decomposable graph G. More specifically, we work with a Gaussian graphical model structure which induces banding in the corresponding precision matrix. Approximate banding structure for precision matrix can arise in certain possibly non-stationary time series framework. Suppose that $\{X_t : t = 1, \ldots, p\}$ is a possibly non-stationary time series with approximately autoregressive-type Markov dependence on neighborhoods. The covariances cannot be estimated based on a single time series due to lack of stationarity. However, if we have replications $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, even when n is much smaller than p, it is still possible to estimate the entire covariance matrix assuming the approximate Markov structure. The graphical model based on the banding structure ensures the decomposability of the graph, along with the

presence of a perfect set of cliques, as explained in Section 3.2. For a G-Wishart prior, we can compute the explicit expression of the normalizing constant of the corresponding marginal distribution of the graph (see Section 3.5). For arbitrary decomposable graphs, the computation of the normalizing constant requires Markov chain Monte-Carlo (MCMC) based methods; see Atay-Kayis and Massam (2005); Carvalho et al. (2007); Carvalho and Scott (2009); Lenkoski and Dobra (2011); Dobra et al. (2011). We obtain posterior convergence rate and convergence rate of the Bayes estimators and the MLE for the graphical model based on banding on the precision matrix. However, we allow the true precision matrix to be outside this class, provided it is well-approximated by banded matrices in an appropriate sense.

The chapter is organized as follows. In the next section, we discuss some preliminaries on graphical models. In Section 3.3, we formulate the estimation problem and then describe the corresponding model assumptions. Section 3.4 deals with the main results related to posterior convergence rates. A method for selecting the banding parameter using the explicit form of the marginal likelihood of a graph is discussed in Section 3.5. In Section 3.6, we compare the performance of the Bayesian estimators with that of the graphical maximum likelihood estimator (MLE) and the banding estimator proposed by Bickel and Levina (2008b). Proof of the main result is shown in Section 3.7. Some auxiliary lemmas and their proofs are included in Section 3.8.

3.2 Preliminaries on graphical models

We have introduced the concept of graphical models in Chapter 1. We now discuss some preliminaries on graphs necessary for this chapter.

An *undirected* graph G consists of a finite non-empty set V of p points, called vertices, and a set of edges $E = \{(i, j) \in V \times V; i < j\}$. For an undirected graph $G = (V, E)$, the

adjacency set of a vertex j in V is defined as $\mathrm{adj}(G, j) = \{k \in V : (j, k) \in E\}$. A *subgraph* $G' = (V', E')$ of $G = (V, E)$ is a graph such that $V' \subseteq V, E' = \{(i, j) \in V' \times V'; i < j\}$. If $V' = V$, $E' \subseteq E$, then G' is called a *spanning subgraph* of G. A graph is called *complete* if all the vertices are adjacent to each other. The *boundary* bd(A) of a subset A of a vertex set V of a graph G is the set of vertices in $V \backslash A$ that are adjacent to vertices in A. The *closure* cl(A) of $A \subseteq V$ is given by $A \cup \mathrm{bd}(A)$.

A *walk* of a graph G is a sequence of vertices $\{j_1, \ldots, j_k\}$ such that $(j_l, j_{l+1}) \in E$, $l = 1, \ldots, k$. The walk is *closed* if $j_1 = j_k$, or *open* otherwise. A walk is a *path* if all the vertices in the sequence are distinct. A graph is called *connected* if every pair of vertices are joined by a path. A maximal connected subgraph of a graph G is called a *connected component* or simply a *component* of G.

If G_1, G_2, G_3 are subgraphs of G, then G_3 is said to *separate* G_1 and G_2 if every path from $j \in G_1$ to $k \in G_2$ contains a vertex in G_3. A *clique* is a maximal complete subgraph. A graph G *decomposes* in to disjoint subgraphs G_1, G_2, G_3 if

1. $G_1 \cup G_2 \cup G_3 = G$,
2. G_3 is complete,
3. G_3 separates G_1 and G_2.

The decomposition of a graph is *proper* if neither G_1 or G_2 is empty. A sequence of subgraphs that cannot be further decomposed are the *prime components* of a graph. A graph is said to be *decomposable* if every prime component is complete. Hence for a decomposable graph, the prime components are the cliques of the graph.

Let $B_1, \ldots, B_k$ be a sequence of subsets of the vertex set V of a graph G. Let $H_j = B_1 \cup \ldots \cup B_j$, $j \geq 1$ and $R_j = B_j \backslash H_{j-1}$, $S_j = B_j \cup H_{j-1}$, $j \geq 2$. The sequence is said to be *perfect* if,

1. for all $i > 1$, there exists $j < i$ such that $S_i \subseteq B_j$,

2. S_j is complete for all $j \geq 2$.

The sets H_j, R_j and S_j are termed as histories, residuals and separators of the sequence respectively. A *perfect numbering* of the vertices V of G is a numbering $\alpha_1, \ldots, \alpha_k$ such that

$$B_j = \mathrm{cl}(\alpha_j) \cap \{\alpha_1, \ldots, \alpha_j\},\ j \geq 1 \tag{3.1}$$

is a perfect sequence of sets. Note that the construction of B_j implies that all B_js are complete, and also they are the cliques of the graph G. We have the following characterization for decomposable graphs.

Proposition 2. *The following are equivalent for an undirected graph G:*

1. *the vertices of G admit a perfect numbering;*

2. *the cliques of G can be numbered to form a perfect sequence;*

3. *G is decomposable.*

We are mostly interested in the equivalence of (2) and (3) in the proposition above, so that a perfect sequence of cliques can be identified once we are confident that the graph is decomposable. For a graphical model, the vertices in V are the indices of the components of a p-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_p)^T$. For a Gaussian random variable $\boldsymbol{X}$ with precision matrix $\boldsymbol{\Omega} = (\!(\omega_{ij})\!)$, the absence of an edge between X_i and X_j implies conditional independence of these two variables given the rest, which is equivalent to $\omega_{ij} = 0$. Figure 3.1 illustrates the connection between a banded precision matrix and the corresponding graphical model. Following the notation in Letac and Massam (2007), we restrict the canonical parameter $\boldsymbol{\Omega}$ in $\mathcal{P}_G$, where $\mathcal{P}_G$ is the cone of positive definite symmetric matrices of order p having

zero entry corresponding to each missing edge in E. Denoting the linear space of symmetric matrices of order p by $\mathcal{M}$, let $\mathcal{M}_p^+ \subset \mathcal{M}$ be the cone of positive definite matrices. The linear space of symmetric incomplete matrices $\boldsymbol{A} = (\!(a_{ij})\!)$ with missing entries a_{ij}, $(i,j) \notin E$, will be denoted by $\mathcal{I}_G$. The parameter space of the Gaussian graphical model can be described by the set of incomplete matrices $\boldsymbol{\Sigma} = \kappa(\boldsymbol{\Omega}^{-1})$, $\boldsymbol{\Omega} \in \mathcal{P}_G$, where $\kappa\colon \mathcal{M} \to \mathcal{I}_G$ is the projection of $\mathcal{M}$ into $\mathcal{I}_G$; see Letac and Massam (2007).

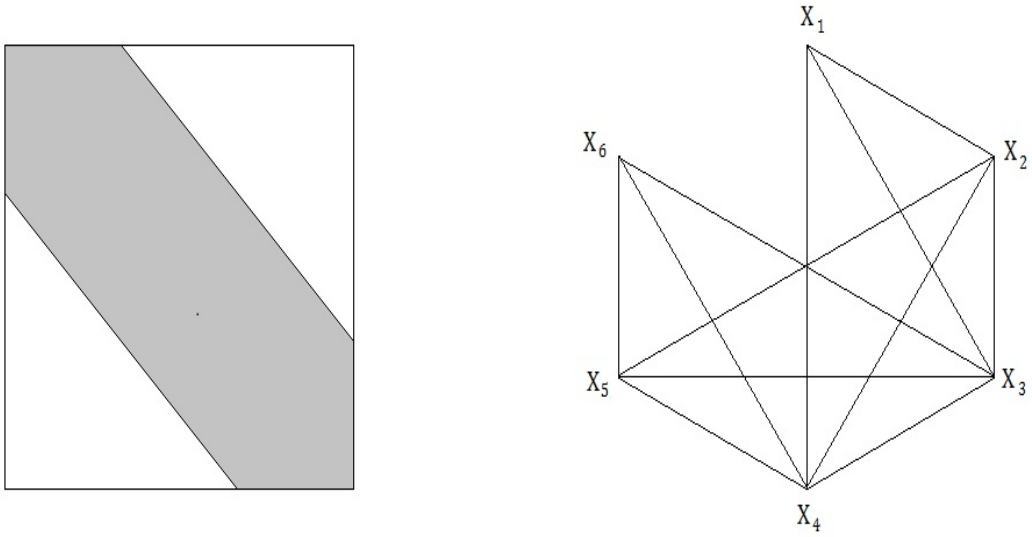


Figure 3.1: [Left] Structure of a banded precision matrix with shaded non-zero entries. [Right] The graphical model corresponding to a banded precision matrix of dimension 6 and banding parameter 3.

For a decomposable graph G with a perfect order of the cliques $\{C_1, \ldots, C_r\}$ and the precision matrix $\boldsymbol{\Omega}$ is given to lie in $\mathcal{P}_G$, the incomplete matrix $\boldsymbol{\Sigma}$ is defined in terms of the submatrices corresponding to the cliques, that is, for each $i = 1, \ldots, r$, $\boldsymbol{\Sigma}_{C_i}$ is positive definite. Thus we have the parameter space for the decomposable Gaussian graphical models restricted

to the two cones

$$\mathcal{P}_G = \{\boldsymbol{A} = (\!(a_{ij})\!) \in \mathcal{M}_p^+ \colon\ a_{ij} = 0, (i,j) \notin E\}, \tag{3.2}$$

$$\mathcal{Q}_G = \{\boldsymbol{B} \in \mathcal{I}_G \colon\ \boldsymbol{B}_{C_i} > \mathbf{0}, i = 1, \ldots, r\}, \tag{3.3}$$

respectively for $\boldsymbol{\Omega}$ and $\boldsymbol{\Sigma}$.

3.3 Model assumption and prior specification

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be independent and identically distributed (i.i.d.) random p-vectors with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. Write $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T$, and assume that the $\boldsymbol{X}_i$, $i = 1, \ldots, n$, are multivariate Gaussian. Consistent estimators for the covariance matrix were obtained in Bickel and Levina (2008b) by banding the sample covariance matrix, assuming a certain sparsity structure on the true covariance. Our aim is to obtain consistency of the graphical MLE and and Bayes estimates of the precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ under the condition $n^{-1} \log p \to 0$ where $\boldsymbol{\Omega}$ ranges over some fairly natural families. For a given positive sequence $\gamma(k) \downarrow 0$, we consider the class of positive definite symmetric matrices $\boldsymbol{\Omega} = (\!(\omega_{ij})\!)$ as

$$\begin{aligned} \mathcal{U}(\varepsilon_0, \gamma) = \Bigg\{ & \boldsymbol{\Omega} \colon \max_j \sum_i \{|\omega_{ij}| \colon |i-j| > k\} \le \gamma(k) \text{ for all } k > 0, \\ & 0 < \varepsilon_0 \le \min \operatorname{eig}_j(\boldsymbol{\Omega}) \le \max \operatorname{eig}_j(\boldsymbol{\Omega}) \le \varepsilon_0^{-1} < \infty \Bigg\}. \end{aligned} \tag{3.4}$$

We also define another class of positive definite symmetric matrices as

$$\mathcal{V}(K,\gamma) = \left\{ \boldsymbol{\Omega} \colon \max_j \sum_i \{|\omega_{ij}| \colon |i-j| > k\} \le \gamma(k) \text{ for all } k > 0, \right. \\ \left. \max\left\{ \|\boldsymbol{\Omega}^{-1}\|_{(\infty,\infty)}, \|\boldsymbol{\Omega}\|_{(\infty,\infty)} \right\} \le K \right\}. \tag{3.5}$$

These two classes are closely related, as shown by the following lemma.

Lemma 3. *For every ε_0, there exist $K_1 \le K_2$ such that*

$$\mathcal{V}(K_1,\gamma) \subset \mathcal{U}(\varepsilon_0,\gamma) \subset \mathcal{V}(K_2,\gamma). \tag{3.6}$$

Proof. We rewrite the class of matrices defined in (3.4) as

$$\mathcal{U}(\varepsilon_0,\gamma) = \left\{ \boldsymbol{\Omega} \colon \max_j \sum_i \{|\omega_{ij}| \colon |i-j| > k\} \le \gamma(k) \text{ for all } k > 0, \right. \\ \left. \max\left\{ \|\boldsymbol{\Omega}^{-1}\|_{(2,2)}, \|\boldsymbol{\Omega}\|_{(2,2)} \right\} \le \varepsilon_0^{-1} \right\}. \tag{3.7}$$

Now, $\max\left\{ \|\boldsymbol{\Omega}^{-1}\|_{(\infty,\infty)}, \|\boldsymbol{\Omega}\|_{(\infty,\infty)} \right\} \le K_1$ implies $\max\left\{ \|\boldsymbol{\Omega}^{-1}\|_{(2,2)}, \|\boldsymbol{\Omega}\|_{(2,2)} \right\} \le \varepsilon_0^{-1}$ for $K_1 = \varepsilon_0^{-1}$, using Lemma 1. Thus $\mathcal{V}(K_1,\gamma) \subset \mathcal{U}(\varepsilon_0,\gamma)$.

To see the other way, note that, for any fixed k_0,

$$\begin{aligned} \|\boldsymbol{\Omega}\|_{(\infty,\infty)} &\le \|\boldsymbol{\Omega} - B_{k_0}(\boldsymbol{\Omega})\|_{(\infty,\infty)} + \|B_{k_0}(\boldsymbol{\Omega})\|_{(\infty,\infty)} \\ &\le \gamma(k_0) + (2k_0+1)\|\boldsymbol{\Omega}\|_{\infty} \\ &\le \gamma(k_0) + (2k_0+1)\|\boldsymbol{\Omega}\|_{(2,2)} \\ &\le \gamma(k_0) + (2k_0+1)\varepsilon_0^{-1}. \end{aligned} \tag{3.8}$$

Choosing $K_2 = \gamma(k_0) + (2k_0 + 1)\varepsilon_0^{-1}$ gives $\mathcal{U}(\varepsilon_0, \gamma) \subset \mathcal{V}(K_2, \gamma)$. □

The sequence $\gamma(k)$ which bounds $\|\mathbf{\Omega} - B_k(\mathbf{\Omega})\|_{(\infty,\infty)}$ has been kept flexible so as to include a number of matrix classes.

1. Exact banding: $\gamma(k) = 0$ for all $k \geq k_0$, which means that the true precision matrix is banded, with banding parameter k_0. For instance, any autoregressive process has such a form of precision matrix.

2. Exponential decay: $\gamma(k) = e^{-ck}$. For instance, any moving average process has such a form of precision matrix.

3. Polynomial decay: $\gamma(k) = \gamma k^{-\alpha}$, $\alpha > 0$. This class of matrices has been considered in Bickel and Levina (2008b) and also in Cai and Yuan (2012).

We shall work with these two general classes $\mathcal{U}(\varepsilon_0, \gamma)$ and $\mathcal{V}(K, \gamma)$ for estimating $\mathbf{\Omega}$. A banding structure in the precision matrix can be induced by a Gaussian graphical model model. Since $\omega_{ij} = 0$ implies that the components X_i and X_j of $\boldsymbol{X}$ are conditionally independent given the others, we can thus define a Gaussian graphical model $G = (V, E)$, where $V = \{1, \ldots, p\}$ indexing the p components $X_1, \ldots, X_p$, and E is the corresponding edge set defined by $E = \{(i, j) : |i - j| \leq k\}$, where k is the size of the band. This describes a parameter space for precision matrices consisting of k-banded matrices, and can be used for the maximum likelihood or the Bayesian approach, where for the latter, a prior distribution on these matrices must be specified. Clearly, G is an undirected, decomposable graphical model for which a perfect order of cliques exist, given by $\mathcal{C} = \{C_1, \ldots, C_{p-k}\}$, $C_j = \{j, \ldots, j+k\}$, $j = 1, \ldots, p-k$. The corresponding separators are given by $\mathcal{S} = \{S_2, \ldots, S_{p-k}\}$, $S_j = \{j, \ldots, j + k - 1\}$, $j = 2, \ldots, p - k$. The choice of the perfect set of cliques is not unique, but the estimator for the precision matrix $\mathbf{\Omega}$ under all choices of the order remains the same.

The W_{P_G}-Wishart distribution $W_{P_G}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{D})$ has three sets of parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{D}$, where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are suitable functions defined on the cliques and separators of the graph respectively, and $\boldsymbol{D}$ is a scaling matrix. The W_{P_G}-family, as a prior distribution for $\boldsymbol{\Omega}$, is conjugate — if the prior distribution on $\frac{1}{2}\boldsymbol{\Omega}$ is $W_{P_G}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{D})$, then the posterior distribution of $\frac{1}{2}\boldsymbol{\Omega}$ given the sample covariance $\boldsymbol{S} = n^{-1}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^T$ is given by $W_{P_G}(\boldsymbol{\alpha} - \frac{n}{2}\mathbf{1}, \boldsymbol{\beta} - \frac{n}{2}\mathbf{1}, \boldsymbol{D} + \kappa(n\boldsymbol{S}))$. The G-Wishart distribution $W_G(\delta, \boldsymbol{D})$ is a special case of the W_{P_G}-Wishart family where

$$\begin{aligned} \alpha_i &= -\frac{\delta + \#C_i - 1}{2},\ i = 1, \ldots, r, \\ \beta_i &= -\frac{\delta + \#S_i - 1}{2},\ i = 2, \ldots, r. \end{aligned} \tag{3.9}$$

We shall specifically work with a G-Wishart prior for $\boldsymbol{\Omega}$. In our case, $\#C_i = k + 1$ for all $i = 1, \ldots, p - k$, and $\#S_j = k$ for all $j = 2, \ldots, p - k$. Thus

$$\begin{aligned} \alpha_i &= -\frac{\delta + k}{2},\ i = 1, \ldots, p - k, \\ \beta_j &= -\frac{\delta + k - 1}{2},\ j = 2, \ldots, p - k. \end{aligned} \tag{3.10}$$

The posterior mean of $\boldsymbol{\Omega}$, given $\boldsymbol{S}$ is

$$\begin{aligned} \mathrm{E}(\boldsymbol{\Omega}|\boldsymbol{S}) = -2 \Bigg[&\sum_{j=1}^{p-k} (\alpha_j - \frac{n}{2}) \left((\boldsymbol{D} + \kappa(n\boldsymbol{S}))_{C_j}^{-1} \right)^0 \\ &- \sum_{j=2}^{p-k} (\beta_j - \frac{n}{2}) \left((\boldsymbol{D} + \kappa(n\boldsymbol{S}))_{S_j}^{-1} \right)^0 \Bigg]. \end{aligned} \tag{3.11}$$

Taking $\boldsymbol{D} = \boldsymbol{I}_p$, the p dimensional indicator matrix, and plugging in the values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, we

get the posterior mean with respect to the G-Wishart prior $W_G(\delta, \boldsymbol{I}_p)$ as,

$$\begin{aligned} \mathrm{E}(\boldsymbol{\Omega}|\boldsymbol{S}) = \frac{\delta+k+n}{n} \Bigg[\sum_{j=1}^{p-k} \left((n^{-1}\boldsymbol{I}_{k+1} + \boldsymbol{S}_{C_j})^{-1} \right)^0 \\ - \sum_{j=2}^{p-k} \left((n^{-1}\boldsymbol{I}_{k} + \boldsymbol{S}_{S_j})^{-1} \right)^0 \Bigg] + n^{-1} \sum_{j=2}^{r} \left((n^{-1}\boldsymbol{I}_{k} + \boldsymbol{S}_{S_j})^{-1} \right)^0 . \end{aligned} \tag{3.12}$$

For a sample of size n from a p-dimensional Gaussian distribution with mean $\mathbf{0}$ and precision matrix $\boldsymbol{\Omega}$, we consider the following two loss functions:

$$\begin{aligned} \text{Stein's loss: } L_1(\widehat{\boldsymbol{\Omega}}, \boldsymbol{\Omega}) = \frac{1}{2}\mathrm{tr}(\widehat{\boldsymbol{\Omega}}\boldsymbol{\Omega}^{-1}) - \log|\widehat{\boldsymbol{\Omega}}\boldsymbol{\Omega}^{-1}| - p, \\ \text{Squared-error loss: } L_2(\widehat{\boldsymbol{\Omega}}, \boldsymbol{\Omega}) = \mathrm{tr}(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega})^2, \end{aligned} \tag{3.13}$$

corresponding to an arbitrary estimator $\widehat{\boldsymbol{\Omega}}$ of $\boldsymbol{\Omega}$. The Bayes estimators corresponding to the above two loss functions was derived by Rajaratnam et al. (2008). Under the G-Wishart prior $W_G(\delta, \boldsymbol{I}_p)$, the Bayes estimators corresponding to Stein's loss function is given by,

$$\widehat{\boldsymbol{\Omega}}^{\mathrm{B}}_{L_1} = \frac{\delta+n-2}{n} \left[\sum_{j=1}^{p-k} \left((n^{-1}\boldsymbol{I}_{k+1} + \boldsymbol{S}_{C_j})^{-1} \right)^0 - \sum_{j=2}^{p-k} \left((n^{-1}\boldsymbol{I}_{k} + \boldsymbol{S}_{S_j})^{-1} \right)^0 \right]. \tag{3.14}$$

For the squared-error loss function, the corresponding Bayes estimator is clearly the posterior mean of $\boldsymbol{\Omega}$ as in (3.12). We denote this estimator by $\widehat{\boldsymbol{\Omega}}^{\mathrm{B}}_{L_2}$. Some other loss functions for estimation of $\boldsymbol{\Omega}$ have also been considered in the literature; see Yang and Berger (1994).

The graphical MLE for $\boldsymbol{\Omega}$ under the graphical model with banding parameter k is given by (see Lauritzen, 1996),

$$\widehat{\boldsymbol{\Omega}}^{\mathrm{M}} = \sum_{j=1}^{p-k} (\boldsymbol{S}^{-1}_{C_j})^0 - \sum_{j=2}^{p-k} (\boldsymbol{S}^{-1}_{S_j})^0. \tag{3.15}$$

3.4 Main results

In this section, we determine the convergence rate of the posterior distribution of the precision matrix. The following theorem describes the behavior of the entire posterior distribution.

Theorem 1. *Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be random samples from a p-dimensional Gaussian distribution with mean zero and precision matrix $\boldsymbol{\Omega}_0 \in \mathcal{U}(\varepsilon_0, \gamma)$ for some $\varepsilon_0 > 0$ and $\gamma(\cdot)$. Suppose that $\boldsymbol{\Omega}$ is given the G-Wishart prior $W_G(\delta, \boldsymbol{I}_p)$, where the graph G has banding of order k. Then posterior distribution of the precision matrix $\boldsymbol{\Omega}$ satisfies*

$$\mathrm{E}_0 \left[\mathrm{P} \left\{ \|\boldsymbol{\Omega} - \boldsymbol{\Omega}_0\|_{(\infty,\infty)} > M\epsilon_{n,k} | \boldsymbol{X} \right\} \right] \to 0 \tag{3.16}$$

for $\epsilon_{n,k} = \max\left\{k^{5/2}(n^{-1} \log\, p)^{1/2}, \gamma(k)\right\}$ and a sufficiently large constant $M > 0$.

In particular, the posterior distribution is consistent in the L_∞-operator norm if $k \to \infty$ such that $k^5 n^{-1} \log p \to 0$.

An important step towards the proof of the above result is to find the convergence rate of the graphical MLE, which is also of independent interest. For high-dimensional situations, even when the sample covariance matrix is singular, the graphical MLE will be positive definite if the number of elements in the cliques of the corresponding graphical model is less than the sample size.

Convergence results for banded empirical covariance (or precision) matrix or estimators based on thresholding approaches are typically given in terms of the L_2-operator norm in the literature. We however use the stronger L_∞-operator norm (or equivalently, L_1-operator norm), so the implication of a convergence rate in our theorems is stronger.

Proposition 3. *Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be random samples from a p-dimensional Gaussian distribution with mean zero and precision matrix $\boldsymbol{\Omega}_0 \in \mathcal{U}(\varepsilon_0, \gamma)$ for some $\varepsilon_0 > 0$ and $\gamma(\cdot)$. Then*

the graphical MLE $\widehat{\mathbf{\Omega}}^{\mathrm{M}}$ of $\mathbf{\Omega}$, corresponding to the Gaussian graphical model with banding parameter k, has convergence rate given by

$$\|\widehat{\mathbf{\Omega}}^{\mathrm{M}} - \mathbf{\Omega}_0\|_{(\infty,\infty)} = O_P\left(\max\left\{k^{5/2}(n^{-1}\,\log\,p)^{1/2}, \gamma(k)\right\}\right). \tag{3.17}$$

In particular, $\widehat{\mathbf{\Omega}}^{\mathrm{M}}$ is consistent in the L_∞-operator norm if $k \to \infty$ such that $k^5 n^{-1} \log p \to 0$.

Proof. The L_∞-operator norm of the difference between the graphical MLE $\widehat{\mathbf{\Omega}}^{\mathrm{M}}$ and the true precision matrix $\mathbf{\Omega}_0$ can be written as

$$\|\widehat{\mathbf{\Omega}}^{\mathrm{M}} - \mathbf{\Omega}_0\|_{(\infty,\infty)} \leq \|\widehat{\mathbf{\Omega}}^{\mathrm{M}} - B_k(\mathbf{\Omega}_0)\|_{(\infty,\infty)} + \|\mathbf{\Omega}_0 - B_k(\mathbf{\Omega}_0)\|_{(\infty,\infty)}. \tag{3.18}$$

As shown in Lauritzen (1996), in a graphical model,

$$\sum_{j=1}^{p-k} \left(\mathbf{\Omega}_{C_j}\right)^0 - \sum_{j=2}^{p-k} \left(\mathbf{\Omega}_{S_j}\right)^0 = \sum_{j=1}^{p-k} \left(\mathbf{\Sigma}_{C_j}^{-1}\right)^0 - \sum_{j=2}^{p-k} \left(\mathbf{\Sigma}_{S_j}^{-1}\right)^0.$$

Hence the first term can be written as

$$\begin{aligned}
&\left\|\sum_{j=1}^{p-k} \left(\boldsymbol{S}_{C_j}^{-1}\right)^0 - \sum_{j=2}^{p-k} \left(\boldsymbol{S}_{S_j}^{-1}\right)^0 - \sum_{j=1}^{p-k} \left(\mathbf{\Omega}_{C_j}\right)^0 + \sum_{j=2}^{p-k} \left(\mathbf{\Omega}_{S_j}\right)^0 \right\|_{(\infty,\infty)} \\
&\quad \leq \left\|\sum_{j=1}^{p-k} \left(\left(\boldsymbol{S}_{C_j}^{-1}\right)^0 - \left(\mathbf{\Sigma}_{C_j}^{-1}\right)^0\right)\right\|_{(\infty,\infty)} + \left\|\sum_{j=2}^{p-k} \left(\left(\boldsymbol{S}_{S_j}^{-1}\right)^0 - \left(\mathbf{\Sigma}_{S_j}^{-1}\right)^0\right)\right\|_{(\infty,\infty)}.
\end{aligned}$$

Let us first bound the first term. Using the fact that there are only $(2k+1)$ terms in above

expressions inside the norms which have a given row non-zero, it follows that

$$\begin{aligned}
&\left\| \sum_{j=1}^{p-k} \left\{ \left(\boldsymbol{S}_{C_j}^{-1} \right)^0 - \left(\boldsymbol{\Sigma}_{C_j}^{-1} \right)^0 \right\} \right\|_{(\infty,\infty)} \\
&\quad = \max_l \sum_{l'} \left| \left[\sum_{j=1}^{p-k} \left\{ \left(\boldsymbol{S}_{C_j}^{-1} \right)^0 - \left(\boldsymbol{\Sigma}_{C_j}^{-1} \right)^0 \right\} \right]_{(l,l')} \right| \\
&\quad \leq \max_l \sum_{j=1}^{p-k} \sum_{l'} \left| \left[\left(\boldsymbol{S}_{C_j}^{-1} \right)^0 - \left(\boldsymbol{\Sigma}_{C_j}^{-1} \right)^0 \right]_{(l,l')} \right| \\
&\quad \leq (2k+1) \max_j \max_l \sum_{l'} \left| \left[(\boldsymbol{S}_{C_j}^{-1} - \boldsymbol{\Sigma}_{C_j}^{-1} \right]_{(l,l')} \right| \\
&\quad = (2k+1) \max_j \left\| \boldsymbol{S}_{C_j}^{-1} - \boldsymbol{\Sigma}_{C_j}^{-1} \right\|_{(\infty,\infty)} \\
&\quad \lesssim k^{3/2} \max_j \left\| \boldsymbol{S}_{C_j}^{-1} - \boldsymbol{\Sigma}_{C_j}^{-1} \right\|_{(2,2)}, \qquad (3.19)
\end{aligned}$$

where the subscript (l, l') on the matrices above stand for their respective (l, l')th entries. Using the multiplicative inequality $\|\boldsymbol{AB}\| \leq \|\boldsymbol{A}\|\|\boldsymbol{B}\|$ of operator norms, we have

$$\begin{aligned}
&\max_j \left\| \boldsymbol{S}_{C_j}^{-1} - \boldsymbol{\Sigma}_{C_j}^{-1} \right\|_{(2,2)} \\
&\quad = \max_j \left\| \boldsymbol{\Sigma}_{C_j}^{-1} (\boldsymbol{\Sigma}_{C_j} - \boldsymbol{S}_{C_j}) \boldsymbol{S}_{C_j}^{-1} \right\|_{(2,2)} \qquad (3.20) \\
&\quad \leq \max_j \left\{ \left\| \boldsymbol{\Sigma}_{C_j}^{-1} \right\|_{(2,2)} \left\| \boldsymbol{\Sigma}_{C_j} - \boldsymbol{S}_{C_j} \right\|_{(2,2)} \left\| \boldsymbol{S}_{C_j}^{-1} \right\|_{(2,2)} \right\}.
\end{aligned}$$

By the assumption on the class of matrices, $\left\| \boldsymbol{\Sigma}_{C_j}^{-1} \right\|_{(2,2)}$ is bounded by K_2. From Lemma 7,

$$\begin{aligned}
\mathrm{P}\left(\max_j \left\| \boldsymbol{S}_{C_j}^{-1} \right\|_{(2,2)} \geq M_1 \right) &\leq p \max_j \mathrm{P}\left(\left\| \boldsymbol{S}_{C_j}^{-1} \right\|_{(2,2)} \geq M_1 \right) \\
&\leq M_1' p k^2 \exp[-m_1 n k^{-2}]
\end{aligned}$$

for some constant $M_1, M_1', m_1 > 0$, while from Lemma 6,

$$\mathrm{P}\left(\max_j \left\|\boldsymbol{\Sigma}_{C_j} - \boldsymbol{S}_{C_j}\right\|_{(2,2)} \geq t\right) \leq M_2 p k^2 \exp[-m_2 n k^{-2} t^2]$$

for $|t| < m_2'$ for some constants $M_2, m_2, m_2' > 0$.

We choose $t = Ak(n^{-1}\log\, p)^{1/2}$ for some sufficiently large A to get the bound

$$\left\|\sum_{j=1}^{p-k}\left(\left(\boldsymbol{S}_{C_j}^{-1}\right)^0 - \left(\boldsymbol{\Sigma}_{C_j}^{-1}\right)^0\right)\right\|_{(\infty,\infty)} = O_P\left(k^{5/2}(n^{-1}\log p)^{1/2}\right). \tag{3.21}$$

By a similar argument, we can establish

$$\left\|\sum_{j=2}^{p-k}\left(\left(\boldsymbol{S}_{S_j}^{-1}\right)^0 - \left(\boldsymbol{\Sigma}_{S_j}^{-1}\right)^0\right)\right\|_{(\infty,\infty)} = O_P\left(k^{5/2}(n^{-1}\log p)^{1/2}\right). \tag{3.22}$$

Therefore, in view of the assumption $\|\boldsymbol{\Omega}_0 - B_k(\boldsymbol{\Omega}_0)\|_{(\infty,\infty)} \leq \gamma(k)$, we obtain the result. □

The proof uses the explicit form of the graphical MLE and proceed by bounding the mean squared error in the L_∞-operator norm. However, as the graphical MLE involves $(k+1)(p - k/2)$ many terms, a naive approach will lead to a factor p in the estimate, which will not be able to establish a convergence rate in the truly high dimensional situations $p \gg n$. We overcome this obstacle by looking more carefully at the structure of the graphical MLE, and note that for any row i, the number of terms in (3.15) which have non-zero ith row is only at most $(2k+1) \ll p$. This along with the description of L_∞-operator norm in terms of row sums give rise to a much smaller factor than p.

Now we treat the Bayes estimators. Consider the G-Wishart prior $W_G(\delta, \boldsymbol{I}_p)$ for $\boldsymbol{\Omega}$, where the graph G has banding of order k and δ is a positive integer. The following result bounds the difference between $\widehat{\boldsymbol{\Omega}}^{\mathrm{M}}$ and the estimators $\widehat{\boldsymbol{\Omega}}^{\mathrm{B}}_{L_1}$ and $\widehat{\boldsymbol{\Omega}}^{\mathrm{B}}_{L_2}$.

Lemma 4. *Assume the conditions of Proposition 3 and suppose that $\boldsymbol{\Omega}$ is given the G-Wishart prior $W_G(\delta, \boldsymbol{I}_p)$, where the graph G has banding of order k. Then $\|\widehat{\boldsymbol{\Omega}}^{\mathrm{B}}_{L_1} - \widehat{\boldsymbol{\Omega}}^{\mathrm{M}}\|_{(\infty,\infty)} = O_P(k^2/n)$, $\|\widehat{\boldsymbol{\Omega}}^{\mathrm{B}}_{L_2} - \widehat{\boldsymbol{\Omega}}^{\mathrm{M}}\|_{(\infty,\infty)} = O_P(k^{5/2}/n)$.*

Proof. We shall first prove the result for $\widehat{\boldsymbol{\Omega}}^{\mathrm{B}}_{L_2}$. The L_∞-operator norm of $\widehat{\boldsymbol{\Omega}}^{\mathrm{B}}_{L_2} - \widehat{\boldsymbol{\Omega}}^{\mathrm{M}}$ can be bounded by

$$\frac{1}{n}\left\|\sum_{j=2}^{p-k}\left((n^{-1}\boldsymbol{I}_k + \boldsymbol{S}_{S_j})^{-1}\right)^0\right\|_{(\infty,\infty)} \tag{3.23}$$

$$+\frac{\delta+k+n}{n}\left\|\sum_{j=1}^{p-k}\left((n^{-1}\boldsymbol{I}_{k+1} + \boldsymbol{S}_{C_j})^{-1}\right)^0 - \sum_{j=1}^{p-k}(\boldsymbol{S}_{C_j}^{-1})^0\right\|_{(\infty,\infty)} \tag{3.24}$$

$$+\frac{\delta+k+n}{n}\left\|\sum_{j=2}^{p-k}\left((n^{-1}\boldsymbol{I}_k + \boldsymbol{S}_{S_j})^{-1}\right)^0 - \sum_{j=2}^{p-k}(\boldsymbol{S}_{S_j}^{-1})^0\right\|_{(\infty,\infty)} \tag{3.25}$$

$$+\left|\frac{\delta+k+n}{n} - 1\right|\left\|\sum_{j=1}^{p-k}\left((\boldsymbol{S}_{C_j})^{-1}\right)^0 - \sum_{j=2}^{p-k}\left((\boldsymbol{S}_{S_j})^{-1}\right)^0\right\|_{(\infty,\infty)}. \tag{3.26}$$

Now, (3.23) above is

$$\begin{aligned}
&\frac{1}{n}\max_l\sum_{l'}\left|\left[\sum_{j=2}^{p-k}\left((n^{-1}\boldsymbol{I}_k + \boldsymbol{S}_{S_j})^{-1}\right)^0\right]_{(l,l')}\right| \\
&\quad\le \frac{1}{n}\max_l\sum_{j=2}^{p-k}\sum_{l'}\left|\left[\left((n^{-1}\boldsymbol{I}_k + \boldsymbol{S}_{S_j})^{-1}\right)^0\right]_{(l,l')}\right| \\
&\quad\le \frac{2k+1}{n}\max_j\max_l\sum_{l'}\left|\left[(n^{-1}\boldsymbol{I}_k + \boldsymbol{S}_{S_j})^{-1}\right]_{(l,l')}\right| \\
&\quad= \frac{2k+1}{n}\max_j\left\|(n^{-1}\boldsymbol{I}_k + \boldsymbol{S}_{S_j})^{-1}\right\|_{(\infty,\infty)},
\end{aligned}$$

which is bounded by a multiple of

$$\frac{k^{3/2}}{n} \max_j \left\| (n^{-1}\boldsymbol{I}_k + \boldsymbol{S}_{S_j})^{-1} \right\|_{(2,2)} \leq \frac{k^{3/2}}{n} \max_j \left\| \boldsymbol{S}_{S_j}^{-1} \right\|_{(2,2)}. \tag{3.27}$$

In view of Lemma 7, we have that for some $M_3, M_3', m_3 > 0$,

$$\mathrm{P}\left(\max_j \left\| \boldsymbol{S}_{S_j}^{-1} \right\|_{(2,2)} \geq M_3 \right) \leq M_3' p k^2 \exp[-m_3 n k^{-2}],$$

which converges to zero if $k^2(\log p)/n \to 0$. This leads to the estimate

$$n^{-1} \left\| \sum_{j=2}^{p-k} \left((n^{-1}\boldsymbol{I}_k + \boldsymbol{S}_{S_j})^{-1} \right)^0 \right\|_{(\infty,\infty)} = O_P\left(k^{3/2}/n \right). \tag{3.28}$$

For (3.24), we observe that

$$\begin{aligned}
&\left\| \sum_{j=1}^{p-k} \left((n^{-1}\boldsymbol{I}_{k+1} + \boldsymbol{S}_{C_j})^{-1} \right)^0 - \sum_{j=1}^{p-k} (\boldsymbol{S}_{C_j}^{-1})^0 \right\|_{(\infty,\infty)} \\
&\quad \leq (2k+1) \max_j \left\| (n^{-1}\boldsymbol{I}_{k+1} + \boldsymbol{S}_{C_j})^{-1} - \boldsymbol{S}_{C_j}^{-1} \right\|_{(\infty,\infty)} \\
&\quad \lesssim k^{3/2} \max_j \left\| (n^{-1}\boldsymbol{I}_{k+1} + \boldsymbol{S}_{C_j})^{-1} - \boldsymbol{S}_{C_j}^{-1} \right\|_{(2,2)}
\end{aligned}$$

and that

$$\begin{aligned}
&\left\| (n^{-1}\boldsymbol{I}_{k+1} + \boldsymbol{S}_{C_j})^{-1} - \boldsymbol{S}_{C_j}^{-1} \right\|_{(2,2)} \\
&\quad \leq \left\| (n^{-1}\boldsymbol{I}_{k+1} + \boldsymbol{S}_{C_j})^{-1} \right\|_{(2,2)} \left\| n^{-1}\boldsymbol{I}_{k+1} \right\|_{(2,2)} \left\| \boldsymbol{S}_{C_j}^{-1} \right\|_{(2,2)} \\
&\quad \leq n^{-1} \left\| \boldsymbol{S}_{C_j}^{-1} \right\|_{(2,2)}^2.
\end{aligned}$$

Now under $k^2(\log p)/n \to 0$, an application of Lemma 7 leads to the bound $O_P(k^{3/2}/n)$ for

(3.24).

A similar argument gives rise to the same $O_P(k^{3/2}/n)$ bound for (3.25).

Finally to consider (3.26). As argued in bounding (3.23), we have that

$$\begin{aligned}&\left\|\sum_{j=1}^{p-k}\left((\boldsymbol{S}_{C_j})^{-1}\right)^0-\sum_{j=2}^{p-k}\left((\boldsymbol{S}_{S_j})^{-1}\right)^0\right\|_{(\infty,\infty)}\\&\quad\le k^{1/2}(2k+1)\left[\max_j\left\|\boldsymbol{S}_{C_j}^{-1}\right\|_{(2,2)}+\max_j\left\|\boldsymbol{S}_{S_j}^{-1}\right\|_{(2,2)}\right]=O_P(k^{3/2}),\end{aligned}$$

under the assumption $k^2(\log p)/n \to 0$ by another application of Lemma 7. Since $n^{-1}(\delta+k+n)-1=O(k/n)$, it follows that (3.26) is $O_P(k^{5/2}/n)$, which is the weakest estimate among all terms in the bound for $\|\widehat{\boldsymbol{\Omega}}^B_{L_2}-\widehat{\boldsymbol{\Omega}}^M\|$. The result thus follows. The proof of the result for $\widehat{\boldsymbol{\Omega}}^{\mathrm{B}}_{L_1}$ follows similarly. □

Proposition 3 and Lemma 4 together lead to the following result for the convergence rate of the Bayes estimators under the G in the L_∞-operator norm.

Proposition 4. *In the setting of Lemma 4, for $\widehat{\boldsymbol{\Omega}}^{\mathrm{B}}$ either $\widehat{\boldsymbol{\Omega}}^{\mathrm{B}}_{L_1}$ or $\widehat{\boldsymbol{\Omega}}^{\mathrm{B}}_{L_2}$, we have*

$$\|\widehat{\boldsymbol{\Omega}}^{\mathrm{B}}-\boldsymbol{\Omega}_0\|_{(\infty,\infty)}=O_P\left(\max\left\{k^{5/2}(n^{-1}\log p)^{1/2},\gamma(k)\right\}\right). \tag{3.29}$$

In particular, the Bayes estimators $\widehat{\boldsymbol{\Omega}}^{\mathrm{B}}_{L_1}$ and $\widehat{\boldsymbol{\Omega}}^{\mathrm{B}}_{L_2}$ are consistent in the L_∞-operator norm if $k\to\infty$ such that $k^5 n^{-1}\log p\to 0$.

Proof. The proof directly follows from Theorem 3 and Lemma 4 using the triangle inequality. □

Remarks on the convergence rates. Observe that the convergence rates of the graphical MLE, the Bayes estimators and the posterior distribution obtained above are the same. The

obtained rates can be optimized by choosing k appropriately as in a bias-variance trade-off. The fastest possible rates obtained from the theorems may be summarized for the different decay rates of $\gamma(k)$ as follows: If the true precision matrix is banded with banding parameter k_0, then the optimal rate of convergence $n^{-1/2}(\log p)^{1/2}$ is obtained by choosing any fixed $k \geq k_0$. When $\gamma(k)$ decays exponentially, the rate of convergence $n^{-1/2}(\log p)^{1/2}(\log n)^2$ can be obtained by choosing k_n approximately proportional to $\log n$ with some sufficiently large constant of proportionality. If $\gamma(k)$ decays polynomially with index α as in Bickel and Levina (2008b), we get the consistency rate of $(n^{-1} \log p)^{\alpha/(2\alpha+5)}$ corresponding to $k_n \asymp (n^{-1} \log p)^{-1/(2\alpha+5)}$.

It is to be noted that we have not assumed that the true structure of the precision matrix arises from a graphical model. The graphical model is a convenient tool to generate useful estimators through the maximum likelihood and Bayesian approach, but the graphical model itself may be a misspecified model. Further, it can be inspected from the proof of the theorems that the Gaussianity assumption on true distribution of the observations is not essential, although the graphical model assumes Gaussianity to generate estimators. The Gaussianity assumption is used to control certain probabilities by applying the probability inequality Lemma A.3 of Bickel and Levina (2008b). However, it was also observed by Bickel and Levina (2008b) that one only requires bounds on the moment generating function of X_i^2, $i = 1, \ldots, p$. In particular, any thinner tailed distribution, such as one with a bounded support, will allow the arguments to go through.

3.4.1 Estimation using a reference prior

A reference prior for the covariance matrix $\boldsymbol{\Sigma}$, obtained in Rajaratnam et al. (2008), can also be used to induce a prior on $\boldsymbol{\Omega}$. This corresponds to an improper $W_{P_G}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{0})$ distribution for

$\frac{1}{2}\boldsymbol{\Omega}$ with

$$\begin{aligned} &\alpha_i = 0, \qquad i = 1, \ldots, r, \\ &\beta_2 = \frac{1}{2}(c_1 + c_2) - s_2,\ \beta_j = \frac{1}{2}(c_j - s_j), \qquad j = 2, \ldots, r. \end{aligned} \tag{3.30}$$

By Corollary 4.1 in Rajaratnam et al. (2008), the posterior mean $\widehat{\boldsymbol{\Omega}}^{\mathrm{R}}$ of the precision matrix is given by

$$\begin{aligned} \widehat{\boldsymbol{\Omega}}^{\mathrm{R}} = &\sum_{j=1}^{r} (\boldsymbol{S}_{C_j}^{-1})^0 - \{1 - n^{-1}(c_1 + c_2 - 2s_2)\}(\boldsymbol{S}_{S_2}^{-1})^0 \\ &- \sum_{j=3}^{r} \{1 - n^{-1}(c_j - s_j)\}(\boldsymbol{S}_{S_j}^{-1})^0. \end{aligned} \tag{3.31}$$

In our scenario, the Bayes estimator under the reference prior is given by the expression

$$\widehat{\boldsymbol{\Omega}}^{R} = \mathrm{E}(\boldsymbol{\Omega}|\boldsymbol{S}) = \sum_{j=1}^{p-k} (\boldsymbol{S}_{C_j}^{-1})^0 - (1 - n^{-1})(\boldsymbol{S}_{S_2}^{-1})^0 - (1 - n^{-1}) \sum_{j=3}^{p-k} (\boldsymbol{S}_{S_j}^{-1})^0.$$

Therefore

$$\begin{aligned} \|\widehat{\boldsymbol{\Omega}}^{\mathrm{R}} - \widehat{\boldsymbol{\Omega}}^{\mathrm{M}}\|_{(\infty,\infty)} &= \left\| n^{-1}(\boldsymbol{S}_{S_2}^{-1})^0 + n^{-1} \sum_{j=3}^{p-k} (\boldsymbol{S}_{S_j}^{-1})^0 \right\|_{(\infty,\infty)} \\ &\leq n^{-1} \left\| \sum_{j=2}^{p-k} (\boldsymbol{S}_{S_j}^{-1})^0 \right\|_{(\infty,\infty)} + n^{-1} \left\| (\boldsymbol{S}_{S_2}^{-1})^0 \right\|_{(\infty,\infty)}. \end{aligned}$$

Proceeding as in the proof of Lemma 4, we get that using the reference prior, the L_∞-operator norm of the difference between the Bayes estimator $\widehat{\boldsymbol{\Omega}}^{\mathrm{R}}$ and the graphical MLE $\widehat{\boldsymbol{\Omega}}^{\mathrm{M}}$ satisfies

$$\|\widehat{\boldsymbol{\Omega}}^{\mathrm{R}} - \widehat{\boldsymbol{\Omega}}^{\mathrm{M}}\|_{(\infty,\infty)} = O_P\left(k^2/n\right). \tag{3.32}$$

3.5 Estimation of banding parameter

In this section, we propose a method of selecting the banding parameter k of the graphical model using the marginal posterior probabilities of the graph induced by banding k, $k = 1, 2, \ldots$. For the G-Wishart prior $W_G(\delta, \boldsymbol{D})$ for $\boldsymbol{\Omega}$, the density is given by

$$p(\boldsymbol{\Omega}|G) = (I_G(\delta, \boldsymbol{D}))^{-1}(\det(\boldsymbol{\Omega}))^{(\delta-2)/2} \exp\left[-\frac{1}{2}\text{tr}(\boldsymbol{D}\boldsymbol{\Omega})\right], \tag{3.33}$$

where $\boldsymbol{D}$ is a symmetric positive definite matrix and

$$I_G(\delta, \boldsymbol{D}) = \int_{\boldsymbol{\Omega}\in\mathcal{P}_G} (\det(\boldsymbol{\Omega}))^{(\delta-2)/2} \exp\left[-\frac{1}{2}\text{tr}(\boldsymbol{D}\boldsymbol{\Omega})\right] d\boldsymbol{\Omega}, \tag{3.34}$$

is the normalizing constant, which is finite for $\delta > 2$. The posterior is given by $W_G(\delta + n, \boldsymbol{D} + n\boldsymbol{S})$. Thus we can get the marginal likelihood for G as

$$p(\boldsymbol{X}|G) = (2\pi)^{-np/2} \frac{I_G(\delta + n, \boldsymbol{D} + n\boldsymbol{S})}{I_G(\delta, \boldsymbol{D})}. \tag{3.35}$$

For a complete graph G, Muirhead (2005) showed that

$$I_G(\delta, \boldsymbol{D}) = \frac{2^{(\delta+p-1)p/2}\Gamma_p\left(\frac{\delta+p-1}{2}\right)}{(\det(\boldsymbol{D}))^{\frac{\delta+p-1}{2}}}, \tag{3.36}$$

where for $a > (p-1)/2$, $\Gamma_p(a) = \pi^{p(p-1)/4}\prod_{i=0}^{p-1}\Gamma(a - \frac{i}{2})$ and $\Gamma(\cdot)$ denotes the gamma function. Roverato (2000) showed that for a decomposable graph G,

$$I_G(\delta, \boldsymbol{D}) = \frac{\prod_{j=1}^{r} I_{C_j}(\delta, \boldsymbol{D}_{C_j})}{\prod_{j=2}^{r} I_{S_j}(\delta, \boldsymbol{D}_{S_j})}, \tag{3.37}$$

where $\{C_1, \ldots, C_r\}$ and $\{S_2, \ldots, S_r\}$ denote the set of cliques and separators respectively

corresponding to G.

In our case, the fitted model which is fit has a banded structure of the precision matrix. We denote the graphical model induced by banding parameter k by G^k. Let ρ_k be a prior on the graph with banding parameter k. Then the corresponding posterior probability of G^k is given by

$$p(G^k \mid \boldsymbol{X}) = \frac{J_{G^k}(\delta, n, \boldsymbol{D}, n\boldsymbol{S})\rho_k}{\sum_{k'} J_{G^{k'}}(\delta, n, \boldsymbol{D}, n\boldsymbol{S})\rho_{k'}}, \tag{3.38}$$

where

$$J_{G^k}(\delta, n, \boldsymbol{D}, n\boldsymbol{S}) = \frac{I_{G^k}(\delta + n, \boldsymbol{D} + n\boldsymbol{S})}{I_{G^k}(\delta, \boldsymbol{D})}. \tag{3.39}$$

Let the cliques and separators be respectively denoted by $C_j^k = \{j, \ldots, j+k\}$, $j = 1, \ldots, p-k$, and $S_j^k = \{j, \ldots, j+k-1\}$, $j = 2, \ldots, p-k$. Note that the sub-graphs corresponding to the cliques and separators are complete, with respective dimensions $k+1$ and k, and $r = p-k$. Therefore (3.36) and (3.37) together lead to

$$I_{G^k}(\delta, \boldsymbol{D}) = \frac{\prod_{j=1}^{p-k} \left\{2^{(\delta+k)(k+1)/2}\Gamma_{k+1}\left(\frac{\delta+k}{2}\right)/(\det(\boldsymbol{D}_{C_j}))^{(\delta+k)/2}\right\}}{\prod_{j=2}^{p-k} \left\{2^{(\delta+k-1)k/2}\Gamma_{k}\left(\frac{\delta+k-1}{2}\right)/(\det(\boldsymbol{D}_{S_j}))^{(\delta+k-1)/2}\right\}}. \tag{3.40}$$

Now, with the choice $\boldsymbol{D} = \boldsymbol{I}_p$ used in the prior $W_G(\delta, \boldsymbol{I})$, (3.39) and (3.40) give

$$\begin{aligned} J_{G^k}(\delta, n, \boldsymbol{I}_p, n\boldsymbol{S}) &= 2^{np/2}\left(\prod_{i=0}^{k}\frac{\Gamma\left(\frac{\delta+n+i}{2}\right)}{\Gamma\left(\frac{\delta+i}{2}\right)}\right)\left(\frac{\Gamma\left(\frac{\delta+n+k}{2}\right)}{\Gamma\left(\frac{\delta+k}{2}\right)}\right)^{p-k-1} \\ &\quad \times \frac{\prod_{j=2}^{p-k}(\det((\boldsymbol{I}_p + n\boldsymbol{S})_{S_j^k}))^{(\delta+k+n-1)/2}}{\prod_{j=1}^{p-k}(\det((\boldsymbol{I}_p + n\boldsymbol{S})_{C_j^k}))^{(\delta+n+k)/2}}. \end{aligned} \tag{3.41}$$

Substituting this expression in (3.38), we get an explicit expression for the posterior distribution of G^k.

A natural method of selecting k is to consider the posterior mode. In the next section, we

investigate the performance of the posterior mode of G^k through a simulation study.

3.6 Numerical results

We check the performance of the Bayes estimators of the precision matrix and compare with the graphical MLE and the banded estimator as proposed in Bickel and Levina (2008b).

Data is simulated from $\mathrm{N}_p(0, \mathbf{\Sigma})$, assuming specific structures of the covariance $\mathbf{\Sigma}$ or the precision $\mathbf{\Omega}$. For all simulations, we compute the L_∞- operator norm, L_2- operator norm, L_2 norm and L_∞ norm of the difference between the estimate and the true parameter for sample sizes $n = 100, 200, 500$ and $p = 50, 100, 200, 500$, representing cases like $p < n$, $p \sim n$, $p > n$ and $p \gg n$. We simulate 100 replications in each cases. Some of the simulation models are the same as those in Bickel and Levina (2008b).

Example 1 (Autoregressive process: AR(1) covariance structure)**.** Let the true covariance matrix have entries given by

$$\sigma_{ij} = \rho^{|i-j|},\ 1 \leq i, j \leq p, \tag{3.42}$$

with $\rho = 0.3$ in our simulation experiment. The precision matrix is banded in this case, with banding parameter 1.

Example 2 (Autoregressive process: AR(4) covariance structure)**.** The elements of true precision matrix are given by

$$\begin{aligned} \omega_{ij} =& \mathbb{1}(|i-j| = 0) + 0.4\ \mathbb{1}(|i-j| = 1) + 0.2\ \mathbb{1}(|i-j| = 2) \\ &+ 0.2\ \mathbb{1}(|i-j| = 3) + 0.1\ \mathbb{1}(|i-j| = 4). \end{aligned} \tag{3.43}$$

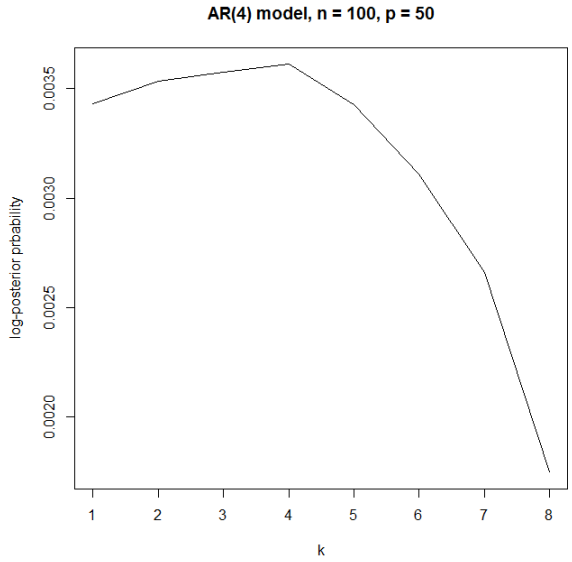
This is the precision matrix corresponding to an AR(4) process.

Example 3 (Long range dependence)**.** We consider a fractional Gaussian Noise process, that is, the increment process of fractional Brownian motion. The elements of the true covariance matrix are given by

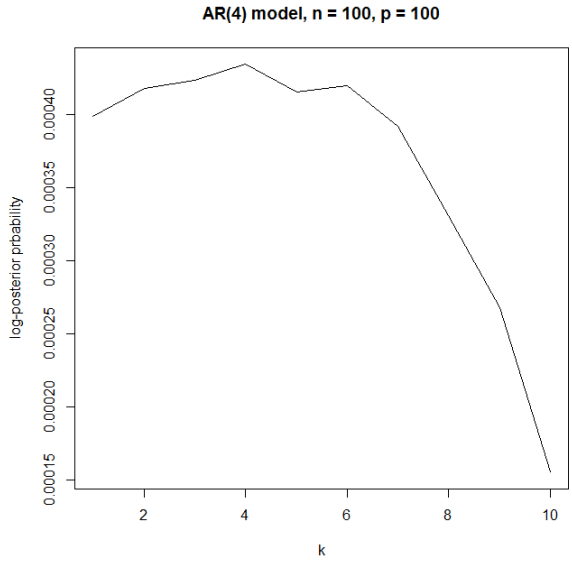
$$\sigma_{ij} = \frac{1}{2}[||i-j|+1|^{2H} - 2|i-j|^{2H} + ||i-j|-1|^{2H}],\ 1 \leq i,j \leq p, \tag{3.44}$$

where $H \in [0.5, 1]$ is the Hurst parameter. We take $H = 0.7$ in the simulation example. This precision matrix does not fall in the polynomial smoothness class used in the theorems. We include this example in the simulation study to check how the proposed method is performing when the assumptions of the theorems are not met.

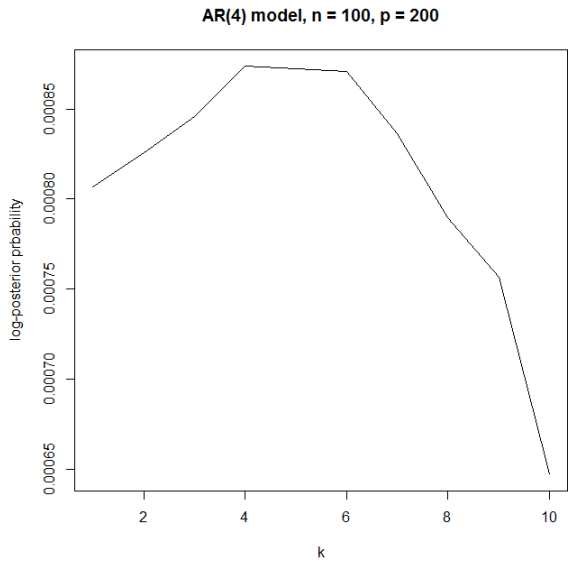
Tables 3.1-3.3 show the simulation results for the different scenarios and compares the performance of the Bayes estimators with the graphical MLE and the banded estimator obtained in Bickel and Levina (2008b) based on a modified Cholesky decomposition. The banding parameter k is chosen using the method discussed in Section 3.5. Figure 3.2 shows the log-posterior probabilities of the graphs corresponding to banding parameter k where prior probability of G^k is taken to be $\rho_k \propto \exp(-k^4)$.



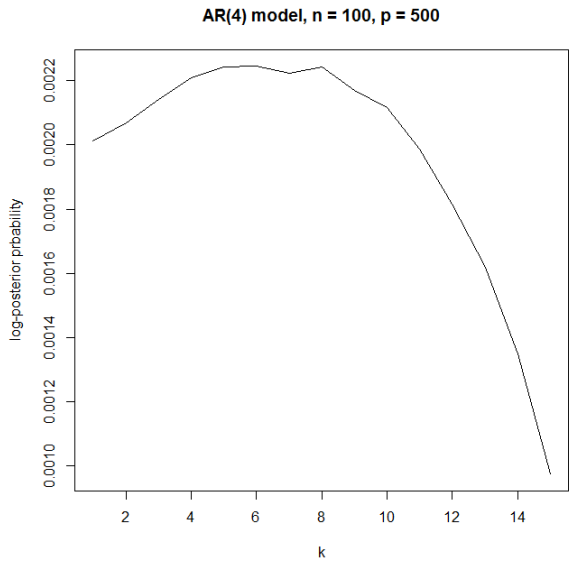
(a) AR(4) model, $n = 100, p = 50$



(b) AR(4) model, $n = 100, p = 100$



(c) AR(4) model, $n = 100, p = 200$



(d) AR(4) model, $n = 100, p = 500$

Figure 3.2: Figures showing log-posterior probabilities of graphs corresponding to different banding parameters k. The graphs are trimmed for larger values of k as the log-posterior probabilities decay further.

Table 3.1: Simulation results for AR(1) model based on 100 replications; figures in parentheses indicate standard errors

		$n = 100$				$n = 200$				$n = 500$			
p	Norm	MLE	$\mathbf{\Omega}^{\mathrm{B}}_{L_2}$	$\mathbf{\Omega}^{\mathrm{B}}_{L_1}$	Cholesky	MLE	$\mathbf{\Omega}^{\mathrm{B}}_{L_2}$	$\mathbf{\Omega}^{\mathrm{B}}_{L_1}$	Cholesky	MLE	$\mathbf{\Omega}^{\mathrm{B}}_{L_2}$	$\mathbf{\Omega}^{\mathrm{B}}_{L_1}$	Cholesky
	$L_{\infty,\infty}$	1.252	1.295	1.175	1.249	0.799	0.820	0.773	0.797	0.477	0.485	0.470	0.477
		(0.029)	(0.029)	(0.027)	(0.029)	(0.018)	(0.018)	(0.017)	(0.018)	(0.009)	(0.009)	(0.008)	(0.008)
	$L_{2,2}$	1.003	1.044	0.940	0.999	0.644	0.663	0.623	0.642	0.374	0.381	0.368	0.374
		(0.029)	(0.023)	(0.021)	(0.023)	(0.016)	(0.016)	(0.015)	(0.016)	(0.007)	(0.007)	(0.007)	(0.007)
50	L_2	2.374	2.454	2.275	2.366	1.609	1.643	1.575	1.607	0.976	0.986	0.968	0.975
		(0.026)	(0.027)	(0.023)	(0.026)	(0.017)	(0.017)	(0.016)	(0.016)	(0.008)	(0.008)	(0.008)	(0.008)
	L_∞	0.729	0.767	0.688	0.726	0.450	0.468	0.437	0.450	0.272	0.278	0.268	0.272
		(0.018)	(0.018)	(0.017)	(0.018)	(0.011)	(0.011)	(0.010)	(0.011)	(0.005)	(0.005)	(0.005)	(0.005)
	$L_{\infty,\infty}$	1.378	1.420	1.295	1.374	0.889	0.912	0.861	0.889	0.525	0.534	0.516	0.525
		(0.029)	(0.029)	(0.027)	(0.029)	(0.018)	(0.018)	(0.017)	(0.018)	(0.009)	(0.009)	(0.009)	(0.009)
	$L_{2,2}$	1.112	1.152	1.042	1.107	0.712	0.734	0.687	0.711	0.408	0.416	0.401	0.408
		(0.022)	(0.022)	(0.021)	(0.023)	(0.015)	(0.015)	(0.015)	(0.015)	(0.007)	(0.007)	(0.007)	(0.007)
100	L_2	3.365	3.482	3.223	3.354	2.264	2.310	2.217	2.263	1.383	1.397	1.371	1.382
		(0.027)	(0.030)	(0.024)	(0.027)	(0.016)	(0.017)	(0.015)	(0.016)	(0.010)	(0.010)	(0.009)	(0.010)
	L_∞	0.814	0.852	0.768	0.808	0.500	0.518	0.484	0.499	0.298	0.305	0.293	0.298
		(0.018)	(0.018)	(0.017)	(0.018)	(0.010)	(0.011)	(0.010)	(0.010)	(0.005)	(0.005)	(0.005)	(0.005)

Table 3.1: (continued)

		$n = 100$				$n = 200$				$n = 500$			
p	Norm	MLE	$\mathbf{\Omega}^{\mathrm{B}}_{L_2}$	$\mathbf{\Omega}^{\mathrm{B}}_{L_1}$	Cholesky	MLE	$\mathbf{\Omega}^{\mathrm{B}}_{L_2}$	$\mathbf{\Omega}^{\mathrm{B}}_{L_1}$	Cholesky	MLE	$\mathbf{\Omega}^{\mathrm{B}}_{L_2}$	$\mathbf{\Omega}^{\mathrm{B}}_{L_1}$	Cholesky
	$L_{\infty,\infty}$	1.558	1.602	1.463	1.557	1.002	1.027	0.967	1.001	0.582	0.593	0.572	0.582
		(0.028)	(0.028)	(0.027)	(0.029)	(0.019)	(0.019)	(0.019)	(0.019)	(0.010)	(0.010)	(0.010)	(0.010)
	$L_{2,2}$	1.237	1.276	1.160	1.233	0.791	0.814	0.763	0.790	0.453	0.463	0.445	0.453
		(0.022)	(0.021)	(0.021)	(0.022)	(0.015)	(0.015)	(0.015)	(0.015)	(0.007)	(0.007)	(0.007)	(0.007)
200	L_2	4.750	4.915	4.548	4.738	3.211	3.277	3.143	3.209	1.971	1.987	1.955	1.970
		(0.024)	(0.026)	(0.022)	(0.025)	(0.017)	(0.018)	(0.017)	(0.017)	(0.010)	(0.010)	(0.010)	(0.010)
	L_∞	0.923	0.961	0.872	0.918	0.564	0.584	0.545	0.564	0.327	0.334	0.321	0.326
		(0.018)	(0.018)	(0.017)	(0.018)	(0.011)	(0.011)	(0.011)	(0.011)	(0.006)	(0.006)	(0.006)	(0.006)
	$L_{\infty,\infty}$	1.765	1.805	1.657	1.763	1.109	1.134	1.069	1.108	0.642	0.653	0.631	0.643
		(0.028)	(0.028)	(0.027)	(0.028)	(0.017)	(0.017)	(0.017)	(0.017)	(0.010)	(0.010)	(0.010)	(0.010)
	$L_{2,2}$	1.407	1.443	1.321	1.406	0.887	0.909	0.856	0.886	0.504	0.514	0.495	0.504
		(0.022)	(0.022)	(0.021)	(0.022)	(0.014)	(0.013)	(0.013)	(0.014)	(0.007)	(0.007)	(0.007)	(0.007)
500	L_2	7.527	7.783	7.209	7.505	5.079	5.177	4.975	5.074	3.133	3.160	3.107	3.131
		(0.029)	(0.030)	(0.026)	(0.029)	(0.015)	(0.016)	(0.015)	(0.015)	(0.010)	(0.010)	(0.009)	(0.010)
	L_∞	1.030	1.066	0.974	1.028	0.626	0.646	0.606	0.625	0.359	0.367	0.354	0.359
		(0.017)	(0.017)	(0.016)	(0.017)	(0.010)	(0.010)	(0.010)	(0.010)	(0.006)	(0.006)	(0.006)	(0.006)

Table 3.2: Simulation results for AR(4) model based on 100 replications; figures in parentheses indicate standard errors

		$n = 100$				$n = 200$				$n = 500$			
p	Norm	MLE	$\mathbf{\Omega}^{\mathrm{B}}_{L_2}$	$\mathbf{\Omega}^{\mathrm{B}}_{L_1}$	Cholesky	MLE	$\mathbf{\Omega}^{\mathrm{B}}_{L_2}$	$\mathbf{\Omega}^{\mathrm{B}}_{L_1}$	Cholesky	MLE	$\mathbf{\Omega}^{\mathrm{B}}_{L_2}$	$\mathbf{\Omega}^{\mathrm{B}}_{L_1}$	Cholesky
	$L_{\infty,\infty}$	1.836	2.066	1.758	1.821	1.078	1.177	1.053	1.076	0.642	0.673	0.636	0.641
		(0.040)	(0.041)	(0.038)	(0.038)	(0.020)	(0.021)	(0.019)	(0.020)	(0.011)	(0.011)	(0.011)	(0.011)
	$L_{2,2}$	1.158	1.340	1.101	1.149	0.672	0.754	0.654	0.672	0.399	0.426	0.394	0.399
		(0.027)	(0.028)	(0.025)	(0.025)	(0.014)	(0.015)	(0.014)	(0.014)	(0.008)	(0.009)	(0.008)	(0.008)
50	L_2	2.539	2.951	2.463	2.526	1.635	1.789	1.612	1.631	0.988	1.030	0.983	0.987
		(0.027)	(0.030)	(0.025)	(0.026)	(0.015)	(0.018)	(0.014)	(0.015)	(0.008)	(0.009)	(0.008)	(0.008)
	L_∞	0.574	0.692	0.554	0.572	0.326	0.378	0.320	0.325	0.180	0.196	0.179	0.180
		(0.014)	(0.015)	(0.014)	(0.014)	(0.007)	(0.007)	(0.007)	(0.007)	(0.003)	(0.004)	(0.003)	(0.004)
	$L_{\infty,\infty}$	1.993	2.231	1.907	1.973	1.210	1.315	1.182	1.209	0.702	0.738	0.694	0.702
		(0.037)	(0.038)	(0.035)	(0.036)	(0.018)	(0.019)	(0.017)	(0.018)	(0.010)	(0.010)	(0.010)	(0.010)
	$L_{2,2}$	1.263	1.451	1.200	1.255	0.761	0.849	0.738	0.760	0.440	0.471	0.434	0.440
		(0.025)	(0.025)	(0.023)	(0.024)	(0.012)	(0.013)	(0.012)	(0.012)	(0.008)	(0.008)	(0.007)	(0.008)
100	L_2	3.626	4.220	3.518	3.607	2.337	2.561	2.303	2.331	1.408	1.466	1.400	1.407
		(0.028)	(0.032)	(0.026)	(0.028)	(0.014)	(0.016)	(0.013)	(0.014)	(0.008)	(0.009)	(0.008)	(0.008)
	L_∞	0.626	0.749	0.605	0.625	0.357	0.411	0.351	0.356	0.196	0.215	0.194	0.196
		(0.015)	(0.015)	(0.014)	(0.014)	(0.006)	(0.006)	(0.006)	(0.006)	(0.003)	(0.003)	(0.003)	(0.003)

Table 3.2: (continued)

		$n = 100$				$n = 200$				$n = 500$			
p	Norm	MLE	$\mathbf{\Omega}^{\mathrm{B}}_{L_2}$	$\mathbf{\Omega}^{\mathrm{B}}_{L_1}$	Cholesky	MLE	$\mathbf{\Omega}^{\mathrm{B}}_{L_2}$	$\mathbf{\Omega}^{\mathrm{B}}_{L_1}$	Cholesky	MLE	$\mathbf{\Omega}^{\mathrm{B}}_{L_2}$	$\mathbf{\Omega}^{\mathrm{B}}_{L_1}$	Cholesky
	$L_{\infty,\infty}$	2.165	2.413	2.069	2.145	1.324	1.435	1.292	1.319	0.763	0.802	0.754	0.762
		(0.034)	(0.035)	(0.032)	(0.033)	(0.018)	(0.019)	(0.018)	(0.018)	(0.011)	(0.011)	(0.011)	(0.011)
	$L_{2,2}$	1.376	1.569	1.307	1.363	0.841	0.932	0.816	0.838	0.479	0.512	0.472	0.478
		(0.022)	(0.022)	(0.021)	(0.021)	(0.013)	(0.013)	(0.012)	(0.013)	(0.008)	(0.008)	(0.008)	(0.008)
200	L_2	5.145	5.988	4.992	5.116	3.332	3.652	3.283	3.324	1.995	2.079	1.984	1.994
		(0.028)	(0.032)	(0.026)	(0.028)	(0.015)	(0.017)	(0.014)	(0.015)	(0.007)	(0.008)	(0.007)	(0.007)
	L_∞	0.695	0.821	0.671	0.689	0.393	0.449	0.386	0.393	0.215	0.235	0.213	0.215
		(0.013)	(0.014)	(0.013)	(0.013)	(0.006)	(0.006)	(0.006)	(0.006)	(0.003)	(0.004)	(0.003)	(0.003)
	$L_{\infty,\infty}$	2.476	2.732	2.362	2.447	1.482	1.599	1.444	1.480	0.833	0.875	0.823	0.830
		(0.035)	(0.036)	(0.034)	(0.034)	(0.018)	(0.018)	(0.018)	(0.019)	(0.010)	(0.011)	(0.010)	(0.010)
	$L_{2,2}$	1.579	1.778	1.498	1.562	0.946	1.039	0.918	0.945	0.526	0.561	0.518	0.524
		(0.023)	(0.024)	(0.022)	(0.023)	(0.014)	(0.014)	(0.014)	(0.014)	(0.007)	(0.007)	(0.006)	(0.006)
500	L_2	8.205	9.553	7.957	8.159	5.287	5.793	5.210	5.273	3.161	3.296	3.143	3.158
		(0.026)	(0.030)	(0.024)	(0.026)	(0.016)	(0.018)	(0.015)	(0.016)	(0.007)	(0.008)	(0.006)	(0.007)
	L_∞	0.787	0.916	0.759	0.779	0.434	0.491	0.426	0.434	0.244	0.265	0.242	0.244
		(0.013)	(0.013)	(0.012)	(0.013)	(0.006)	(0.006)	(0.006)	(0.006)	(0.004)	(0.004)	(0.004)	(0.004)

Table 3.3: Simulation results for FGN model based on 100 replications; figures in parentheses indicate standard errors

		$n = 100$				$n = 200$				$n = 500$			
p	Norm	MLE	$\mathbf{\Omega}^{\mathrm{B}}_{L_2}$	$\mathbf{\Omega}^{\mathrm{B}}_{L_1}$	Cholesky	MLE	$\mathbf{\Omega}^{\mathrm{B}}_{L_2}$	$\mathbf{\Omega}^{\mathrm{B}}_{L_1}$	Cholesky	MLE	$\mathbf{\Omega}^{\mathrm{B}}_{L_2}$	$\mathbf{\Omega}^{\mathrm{B}}_{L_1}$	Cholesky
	$L_{\infty,\infty}$	1.530	1.588	1.493	1.527	1.184	1.213	1.170	1.182	0.969	0.981	0.965	0.969
		(0.025)	(0.025)	(0.024)	(0.024)	(0.015)	(0.015)	(0.014)	(0.015)	(0.008)	(0.008)	(0.008)	(0.008)
	$L_{2,2}$	0.849	0.902	0.820	0.846	0.587	0.614	0.577	0.586	0.412	0.422	0.409	0.412
		(0.019)	(0.019)	(0.018)	(0.019)	(0.012)	(0.012)	(0.011)	(0.011)	(0.005)	(0.005)	(0.005)	(0.005)
50	L_2	2.169	2.271	2.116	2.164	1.666	1.706	1.646	1.665	1.329	1.340	1.323	1.329
		(0.020)	(0.022)	(0.020)	(0.020)	(0.012)	(0.013)	(0.012)	(0.012)	(0.006)	(0.006)	(0.006)	(0.006)
	L_∞	0.570	0.615	0.552	0.569	0.360	0.376	0.355	0.359	0.221	0.226	0.219	0.221
		(0.015)	(0.015)	(0.014)	(0.015)	(0.008)	(0.008)	(0.008)	(0.008)	(0.003)	(0.004)	(0.003)	(0.003)
	$L_{\infty,\infty}$	1.687	1.745	1.647	1.682	1.316	1.345	1.301	1.315	1.058	1.069	1.053	1.058
		(0.024)	(0.024)	(0.023)	(0.024)	(0.014)	(0.014)	(0.014)	(0.014)	(0.007)	(0.007)	(0.007)	(0.007)
	$L_{2,2}$	0.939	0.992	0.907	0.937	0.645	0.672	0.633	0.645	0.441	0.451	0.438	0.441
		(0.018)	(0.018)	(0.017)	(0.018)	(0.011)	(0.011)	(0.011)	(0.011)	(0.005)	(0.005)	(0.005)	(0.005)
100	L_2	3.076	3.222	3.000	3.068	2.351	2.406	2.322	2.351	1.886	1.902	1.876	1.885
		(0.021)	(0.023)	(0.020)	(0.022)	(0.013)	(0.014)	(0.012)	(0.013)	(0.006)	(0.007)	(0.006)	(0.006)
	L_∞	0.634	0.680	0.614	0.632	0.395	0.413	0.388	0.394	0.238	0.243	0.236	0.238
		(0.015)	(0.016)	(0.015)	(0.015)	(0.008)	(0.008)	(0.007)	(0.008)	(0.003)	(0.003)	(0.003)	(0.003)

Table 3.3: (continued)

		$n = 100$				$n = 200$				$n = 500$			
p	Norm	MLE	$\mathbf{\Omega}^{\mathrm{B}}_{L_2}$	$\mathbf{\Omega}^{\mathrm{B}}_{L_1}$	Cholesky	MLE	$\mathbf{\Omega}^{\mathrm{B}}_{L_2}$	$\mathbf{\Omega}^{\mathrm{B}}_{L_1}$	Cholesky	MLE	$\mathbf{\Omega}^{\mathrm{B}}_{L_2}$	$\mathbf{\Omega}^{\mathrm{B}}_{L_1}$	Cholesky
	$L_{\infty,\infty}$	1.855	1.914	1.811	1.857	1.446	1.475	1.430	1.446	1.134	1.145	1.129	1.134
		(0.022)	(0.022)	(0.021)	(0.022)	(0.014)	(0.014)	(0.014)	(0.014)	(0.008)	(0.008)	(0.008)	(0.008)
	$L_{2,2}$	1.024	1.078	0.989	1.025	0.705	0.733	0.693	0.705	0.471	0.481	0.468	0.471
		(0.016)	(0.016)	(0.016)	(0.016)	(0.011)	(0.011)	(0.011)	(0.011)	(0.005)	(0.005)	(0.005)	(0.005)
200	L_2	4.348	4.555	4.239	4.340	3.335	3.414	3.294	3.334	2.659	2.680	2.645	2.659
		(0.020)	(0.021)	(0.019)	(0.020)	(0.013)	(0.014)	(0.013)	(0.013)	(0.006)	(0.006)	(0.006)	(0.006)
	L_∞	0.712	0.760	0.689	0.711	0.439	0.460	0.431	0.439	0.257	0.262	0.255	0.256
		(0.015)	(0.015)	(0.014)	(0.015)	(0.008)	(0.008)	(0.008)	(0.008)	(0.003)	(0.004)	(0.003)	(0.004)
	$L_{\infty,\infty}$	2.059	2.116	2.008	2.056	1.565	1.595	1.548	1.566	1.219	1.231	1.214	1.219
		(0.022)	(0.022)	(0.022)	(0.022)	(0.012)	(0.012)	(0.012)	(0.013)	(0.007)	(0.007)	(0.007)	(0.007)
	$L_{2,2}$	1.156	1.208	1.116	1.151	0.773	0.800	0.759	0.773	0.507	0.517	0.504	0.507
		(0.017)	(0.016)	(0.016)	(0.017)	(0.010)	(0.010)	(0.010)	(0.010)	(0.004)	(0.005)	(0.004)	(0.005)
500	L_2	6.865	7.190	6.694	6.850	5.266	5.386	5.201	5.263	4.215	4.249	4.194	4.215
		(0.022)	(0.023)	(0.021)	(0.022)	(0.012)	(0.013)	(0.012)	(0.012)	(0.007)	(0.007)	(0.007)	(0.007)
	L_∞	0.802	0.851	0.776	0.801	0.482	0.505	0.474	0.482	0.282	0.288	0.280	0.282
		(0.014)	(0.014)	(0.014)	(0.014)	(0.007)	(0.007)	(0.007)	(0.007)	(0.004)	(0.004)	(0.004)	(0.004)

3.7 Proof of Theorem 1

Proof. The posterior distribution of the precision matrix $\boldsymbol{\Omega}$ given the data $\boldsymbol{X}$ is a G-Wishart distribution $W_G(\delta + n, \boldsymbol{I}_p + n\boldsymbol{S})$. We can write $\boldsymbol{\Omega}$ as

$$\boldsymbol{\Omega} = \sum_{j=1}^{p-k} \left(\boldsymbol{\Omega}_{C_j}\right)^0 - \sum_{j=2}^{p-k} \left(\boldsymbol{\Omega}_{S_j}\right)^0 = \sum_{j=1}^{p-k} \left(\boldsymbol{\Sigma}_{C_j}^{-1}\right)^0 - \sum_{j=2}^{p-k} \left(\boldsymbol{\Sigma}_{S_j}^{-1}\right)^0. \tag{3.45}$$

The submatrix $\boldsymbol{\Sigma}_{C_j}$ for any clique C_j has a inverse Wishart distribution with parameters $\delta + n$ and scale matrix $(\boldsymbol{I}_p + n\boldsymbol{S})_{C_j}$, $j = 1, \ldots, p-k$. Thus, $W_{C_j} = \boldsymbol{\Sigma}_{C_j}^{-1}$ has a Wishart distribution induced by the corresponding inverse Wishart distribution. In particular, if $i \in C_j$, then $\tau_{in}^{-1} w_{ii}$ has chi-square distribution with $(\delta + n)$ degrees of freedom, where τ_{in} is the (i,i)th entry of $((\boldsymbol{I} + \boldsymbol{S}_{C_j})^{-1})^0$. Fix a clique $C = C_j$ and define $\boldsymbol{T}_n = \mathrm{diag}(w_{ii}\colon i \in C)$. For $i, j \in C$, let $w_{ij}^* = w_{ij}/\sqrt{\tau_{in}\tau_{jn}}$ and $\boldsymbol{W}_C^* = (\!(w_{ij}^*\colon i, j \in C)\!)$. Then $\boldsymbol{W}_C^*$ given $\boldsymbol{X}$ has a Wishart distribution with parameters $\delta + n$ and scale matrix $\boldsymbol{T}_n^{-1/2}(\boldsymbol{I}_{k+1} + n\boldsymbol{S}_C)\boldsymbol{T}_n^{-1/2}$.

We first note that $\max_i \tau_{in} = O_P(n^{-1})$. To see this, observe that $(\boldsymbol{I}_k + n\boldsymbol{S}_C)^{-1} \leq n^{-1}\boldsymbol{S}_C^{-1}$, so that

$$\max_i |\tau_{in}| \leq \frac{1}{n} \|\boldsymbol{S}_C^{-1}\|_{(2,2)} = O_P(n^{-1})$$

in view of Lemma 7. On the other hand, from Lemma 6, it follows that $\max_C \|\boldsymbol{S}_C\|_{(2,2)} = O_P(1)$, so with probability tending to one, $\boldsymbol{S}_C \leq L\boldsymbol{I}_C$, and hence $(\boldsymbol{I} + n\boldsymbol{S})_C^{-1} \geq (1 + nL)^{-1}\boldsymbol{I}_C$ simultaneously for all cliques, for some constant $L > 0$. Hence $\max_i \tau_{in}^{-1} = O_P(n)$. Consequently, with probability tending to one, the maximum eigenvalue of $\boldsymbol{T}_n^{-1/2}(\boldsymbol{I}_{k+1} + n\boldsymbol{S}_C)\boldsymbol{T}_n^{-1/2}$ is bounded by a constant depending only on ϵ_0, simultaneously for all cliques. Hence applying Lemma A.3 of Bickel and Levina (2008b), it follows that for all i, j,

$$\mathrm{P}\left[|w_{ij} - \mathrm{E}(w_{ij}|\boldsymbol{X})| \geq t\right] \leq M_4 \exp[-m_4(\delta + n)t^2], \quad |t| < m_4', \tag{3.46}$$

for some constants $M_4, m_4, m_4' > 0$ depending on ϵ_0 only.

Now, as a G-Wishart prior gives rise to a k-banded structure, as arguing in the bounding of (3.23) and using (3.45), we have that, for some $M_5, m_5, m_5' > 0$, and all $|t| < m_5'$,

$$\mathrm{P}\left\{\|\boldsymbol{\Omega} - \widehat{\boldsymbol{\Omega}}^{\mathrm{B}}_{L_2}\|_{(\infty,\infty)} \geq k^2 t | \boldsymbol{X}\right\} \leq M_5 p k^2 \exp[-m_5 n t^2]. \tag{3.47}$$

The reduction in the number of terms in the rows from p to $(2k+1)$ arises due to the fact that the G-Wishart posterior preserves the banded structure of the precision matrix. Choosing $t = A(n^{-1}\log p)^{-1/2}$, with A sufficiently large, we get

$$\mathrm{E}_0[\mathrm{P}\{\|\boldsymbol{\Omega} - \widehat{\boldsymbol{\Omega}}^{\mathrm{B}}_{L_2}\|_{(\infty,\infty)} \geq A k^2 (n^{-1}\log p)^{-1/2} | \boldsymbol{X}\}] \to 0. \tag{3.48}$$

Therefore, using Proposition 4,

$$\begin{aligned}
&\mathrm{E}_0\left[\mathrm{P}\left\{\|\boldsymbol{\Omega} - \boldsymbol{\Omega}_0\|_{(\infty,\infty)} > 2\epsilon_n | \boldsymbol{X}\right\}\right] \\
&\quad \leq \mathrm{P}_0\left\{\|\widehat{\boldsymbol{\Omega}}^{\mathrm{B}}_{L_2} - \boldsymbol{\Omega}_0\|_{(\infty,\infty)} > \epsilon_n | \boldsymbol{X}\right\} + \mathrm{E}_0\left[\mathrm{P}\left\{\|\boldsymbol{\Omega} - \widehat{\boldsymbol{\Omega}}^{\mathrm{B}}_{L_2}\|_{(\infty,\infty)} > \epsilon_n | \boldsymbol{X}\right\}\right],
\end{aligned}$$

which converges to zero if $\epsilon_n = \max\{A k^{5/2}(n^{-1}\log p)^{-1/2}, \gamma(k)\}$. □

3.8 Proofs of auxiliary results

In this section we state and prove some lemmas necessary for proving the main results in this chapter. The lemmas are also of some general interest. The next lemma is a version of Lemma A.3 from Bickel and Levina (2008b).

Lemma 5 (Lemma A.3 of Bickel and Levina (2008b))**.** *Let $\boldsymbol{Z}_i$, $i = 1, \ldots, n$, be i.i.d. k-dimensional random vectors distributed as $\mathrm{N}_k(\mathbf{0}, \boldsymbol{D})$ and $\|\boldsymbol{D}\|_{(2,2)} \leq \varepsilon_0^{-1} < \infty$. Then if*

$\boldsymbol{D} = (\!(d_{jk})\!)$,

$$\mathrm{P}\left[\left|\sum_{i=1}^{n}(Z_{ij}Z_{ik} - d_{jk})\right| \geq nt\right] \leq C_1 \exp(-C_2 nt^2)\ for |t| \leq \delta, \tag{3.49}$$

where C_1, C_2 and δ depend on ε_0 only.

Lemma 6. *Let $\boldsymbol{Z}_i$, $i = 1, \ldots, n$, be i.i.d. k-dimensional random vectors distributed as $\mathrm{N}_k(\boldsymbol{0}, \boldsymbol{D})$ and $\max\left\{\|\boldsymbol{D}^{-1}\|_{(r,r)}, \|\boldsymbol{D}\|_{(r,r)}\right\} \leq K$ for $r \in \{2, \infty\}$. Then for the sample variance $\boldsymbol{S}_n = n^{-1}\sum_{i=1}^{n}\boldsymbol{Z}_i\boldsymbol{Z}_i^T$, we have*

$$\mathrm{P}\left[\|\boldsymbol{S}_n - \boldsymbol{D}\|_{(r,r)} \geq t\right] \leq Mk^2 \exp(-mnk^{-2}t^2), \quad |t| \leq m', \tag{3.50}$$

where $M, m, m' > 0$ depend on K only.

In particular, if $k^2(\log k)/n \to 0$, then $\|\boldsymbol{S}_n\|_{(r,r)} = O_P(1)$.

Proof. The proof directly follows from 5 and noting from Lemma 1 that $\|\boldsymbol{S}_n - \boldsymbol{D}\|_{(r,r)} \leq k\|\boldsymbol{S}_n - \boldsymbol{D}\|_{\infty}$. □

Lemma 7. *Let $\boldsymbol{Z}_i$, $i = 1, \ldots, n$, be i.i.d. k-dimensional random vectors distributed as $\mathrm{N}_k(\boldsymbol{0}, \boldsymbol{D})$ and $\max\left\{\|\boldsymbol{D}^{-1}\|_{(r,r)}, \|\boldsymbol{D}\|_{(r,r)}\right\} \leq K$ for $r \in \{2, \infty\}$. Then for the sample variance $\boldsymbol{S}_n = n^{-1}\sum_{i=1}^{n}\boldsymbol{Z}_i\boldsymbol{Z}_i^T$, we have*

$$\mathrm{P}\left[\|\boldsymbol{S}_n^{-1}\|_{(r,r)} \geq M\right] \leq M'k^2 \exp(-mnk^{-2}C'^2), \tag{3.51}$$

where $M > K$ and $M', m > 0$ depend on M and K only.

Proof. Note that,

$$\begin{aligned} \|\boldsymbol{S}_n^{-1}\|_{(r,r)} &\leq \|\boldsymbol{D}^{-1}\|_{(r,r)} + \|\boldsymbol{S}_n^{-1} - \boldsymbol{D}^{-1}\|_{(r,r)} \\ &= \|\boldsymbol{D}^{-1}\|_{(r,r)} + \|\boldsymbol{D}^{-1}\|_{(r,r)}\|\boldsymbol{S}_n - \boldsymbol{D}\|_{(r,r)}\|\boldsymbol{S}_n^{-1}\|_{(r,r)} \\ &\leq K(1 + \|\boldsymbol{S}_n - \boldsymbol{D}\|_{(r,r)}\|\boldsymbol{S}_n^{-1}\|_{(r,r)}). \end{aligned} \tag{3.52}$$

This implies that

$$\|\boldsymbol{S}_n^{-1}\|_{(r,r)} \leq \frac{K}{1 - \|\boldsymbol{S}_n - \boldsymbol{D}\|_{(r,r)}K}.$$

Thus, using Lemma 6, we obtain

$$\begin{aligned} \mathrm{P}\left[\|\boldsymbol{S}_n^{-1}\|_{(r,r)} \geq M\right] &\leq \mathrm{P}\left[\frac{K}{1 - \|\boldsymbol{S}_n - \boldsymbol{D}\|_{(r,r)}K} \geq M\right] \\ &\leq \mathrm{P}\left[\|\boldsymbol{S}_n - \boldsymbol{D}\|_{(r,r)} \geq K^{-1} - M^{-1}\right] \\ &\leq M'k^2 \exp(-mnk^{-2}). \end{aligned} \tag{3.53}$$

□

Chapter 4

Bayesian estimation of a sparse precision matrix

4.1 Introduction

In this chapter, we consider estimation of a large precision matrix under sparsity assumptions on the true precision matrix. As introduced in Chapter 1, there are several regularization based methods for large matrices, which includes banding, thresholding, tapering or penalization based methods. The primary goal of these methods is to impose a sparsity structure in the matrix. Most of them 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.

We consider the Bayesian graphical lasso prior developed in Wang (2012) to estimate a

sparse precision matrix. 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. We shall work with the Bayesian graphical lasso prior with additional point mass at zero for the off-diagonal elements in the precision matrix.

We derive posterior convergence rates for the Bayesian graphical lasso prior in terms of the Frobenius norm under appropriate sparsity conditions. 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 (2005); Curtis et al. (2014). 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 to be non-regular graphical models following the terminology used by Yuan and Lin (2005) 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 chapter 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 4.2, 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. (2000). In Section 4.3, 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 small under appropriate conditions. A simulation study is performed in the Section 4.4 followed by a real data example in Section 4.5. Proofs of main results and additional lemmas are included in Section 4.6.

4.2 Model, prior and posterior concentration

Consider n independent random samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from $\mathrm{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is nonsingular and the precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ is sparse. The problem is to estimate $\boldsymbol{\Omega}$ and to learn the underlying graphical structure. We denote the natural unbiased estimator of $\boldsymbol{\Sigma}$ by $\widehat{\boldsymbol{\Sigma}} = n^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^T$.

The graphical lasso produces sparse solutions for the precision matrix, in similar lines to that of the lasso in case of linear regression. The graphical lasso estimator minimizes two times the penalized negative log-likelihood

$$- \log \det(\boldsymbol{\Omega}) + \mathrm{tr}(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Omega}) + \rho \|\boldsymbol{\Omega}\|_1, \tag{4.1}$$

over the class of positive definite matrices, denoted by $\mathcal{M}^+$, and $\rho \geq 0$ acts as the penalty parameter. Rothman et al. (2008) derived frequentist convergence rates of the penalized estimator under some sparsity assumptions on the true precision matrix. More specifically, consider the following class of positive definite matrices of order p:

$$\begin{aligned} \mathcal{U}(\varepsilon_0, s) = \{\boldsymbol{\Omega} \colon & \#\{(i,j) \colon \omega_{ij} \neq 0, 1 \leq i < j \leq p\} \leq s, \\ & 0 < \varepsilon_0 \leq \mathrm{eig}_1(\boldsymbol{\Omega}) \leq \mathrm{eig}_p(\boldsymbol{\Omega}) \leq \varepsilon_0^{-1} < \infty \big\} . \end{aligned} \tag{4.2}$$

Though Rothman et al. (2008) considered penalizing only the off-diagonal elements of $\boldsymbol{\Omega}$, some

modification of the proof of their result leads to the same convergence rate for the graphical lasso estimator, obtained by additionally penalizing the diagonal elements. Let us denote $\mathbf{\Omega}^*$ as the graphical lasso estimator based on a sample of size n from a p-dimensional Gaussian distribution with precision matrix $\mathbf{\Omega}_0 \in \mathcal{U}(\varepsilon_0, s)$, where $\mathcal{U}(\varepsilon_0, s)$ is given by (4.2). Then, it follows from Theorem 1 in Rothman et al. (2008) that the rate of convergence of $\mathbf{\Omega}^*$ is $n^{-1/2}(p+s)^{1/2}\log p$. By the triangle inequality,

$$\|\mathbf{\Omega}^*\|_{(2,2)} \leq \|\mathbf{\Omega}_0\|_{(2,2)} + \|\mathbf{\Omega}^* - \mathbf{\Omega}_0\|_{(2,2)}.$$

Also, the triangle inequality and sub-multiplicative property for matrix operator norms gives,

$$\begin{aligned} \|\mathbf{\Omega}^{*-1}\|_{(2,2)} &\leq \|\mathbf{\Omega}_0^{-1}\|_{(2,2)} + \|\mathbf{\Omega}^{*-1} - \mathbf{\Omega}_0^{-1}\|_{(2,2)} \\ &\leq \|\mathbf{\Omega}_0^{-1}\|_{(2,2)} + \|\mathbf{\Omega}_0^{-1}\|_{(2,2)}\|\mathbf{\Omega}^* - \mathbf{\Omega}_0\|_{(2,2)}\|\mathbf{\Omega}^{*-1}\|_{(2,2)}. \end{aligned}$$

Thus, we get,

$$\|\mathbf{\Omega}^{*-1}\|_{(2,2)} \leq \frac{\|\mathbf{\Omega}_0^{-1}\|_{(2,2)}}{1 - \|\mathbf{\Omega}_0^{-1}\|_{(2,2)}\|\mathbf{\Omega}^* - \mathbf{\Omega}_0\|_{(2,2)}}.$$

Now, we have, $\|\mathbf{\Omega}_0\|_{(2,2)} \leq \varepsilon_0^{-1}$ by assumption, and it follows from Theorem 1 in Rothman et al. (2008) that $\|\mathbf{\Omega}^* - \mathbf{\Omega}_0\|_2 = o_P(1)$ as $n \to \infty$. Noting that $\|\mathbf{\Omega}^* - \mathbf{\Omega}_0\|_{(2,2)} \leq \|\mathbf{\Omega}^* - \mathbf{\Omega}_0\|_2$, we get,

$$\|\mathbf{\Omega}^*\|_{(2,2)} = O_P(1),\ \|\mathbf{\Omega}^{*-1}\|_{(2,2)} = O_P(1). \tag{4.3}$$

In the Bayesian context, Wang (2012) introduced the graphical lasso prior, which uses exponential distributions on diagonal elements and Laplace density $\lambda e^{-\lambda|x|}/2$ on off-diagonal elements, all independently of each other, and finally imposes a positive definiteness constraint. The graphical lasso prior has a drawback that it puts absolutely continuous priors on the ele-

ments of the precision matrix, and hence the posterior probabilities of the event $\{\omega_{ij} = 0\}$ is always exactly zero.

Wang (2012) also mentioned an extension of the graphical lasso by putting an additional level of prior on the underlying graphical model structure using point mass priors on the events corresponding to the absence of an edge in the edge-set E, although did not develop the method. We put point-mass prior on the events $\{\omega_{ij} = 0\}$ to make posterior inference about the sparse structure of the underlying graphical model. Define $\mathbf{\Gamma} = (\gamma_{ij}: 1 \leq i < j \leq p)$ to be a $\binom{p}{2}$ vector of edge-inclusion indicator, that is,

$$\gamma_{ij} = \mathbb{1}\{(i,j) \in E\},\ 1 \leq i < j \leq p. \tag{4.4}$$

Similar to the Bayesian graphical lasso prior, given the underlying graphical structure, we put a Laplace prior on the non-zero off-diagonal elements of the precision matrix and for the diagonal elements we have a exponential prior, overall maintaining the positive definiteness of the parameter. Then the joint prior density on $\mathbf{\Omega}$ is given by,

$$p(\mathbf{\Omega}|\mathbf{\Gamma}) \propto \prod_{\gamma_{ij}=1} \{\exp(-\lambda|\omega_{ij}|)\} \prod_{i=1}^{p} \{\exp(-\lambda\omega_{ii}/2)\}\, \mathbb{1}_{\mathcal{M}^+}(\mathbf{\Omega}). \tag{4.5}$$

We propose two different priors on the graphical structure indicator $\mathbf{\Gamma}$. The edge indicators $\gamma_{ij}, 1 \leq i < j \leq p$ are considered to be independent and identically distributed (i.i.d) Bernoulli(q) random variables, and conditioned to the restriction that the model size given by $\sum_{1\leq i<j\leq p} \gamma_{ij}$ does not exceed $\bar{R}$. For some $a_1, a_2 > 0$, the prior distribution on $\bar{R}$ is assumed to satisfy

$$\mathrm{P}(\bar{R} > a_1 m) \leq e^{-a_2 m \log m}. \tag{4.6}$$

This prior is similar to that used by Castillo and van der Vaart (2012), which chooses the model

size first according to a distribution with a similar tail decay and then subsets are selected randomly with equal probability. We can also specify the individual priors on γ_{ij} the same as above, but now truncating the model size to some fixed $\bar{r}$, where $\bar{r}$ is chosen so as to satisfy the metric entropy condition required for posterior convergence.

Thus, in the first situation, the prior on the graphical structure indicator $\mathbf{\Gamma}$, given $\bar{R}$, is given by,

$$p(\mathbf{\Gamma} \mid \bar{R}) \propto q^{\#\mathbf{\Gamma}}(1-q)^{\binom{p}{2}-\#\mathbf{\Gamma}} \mathbb{1}(\#\mathbf{\Gamma} \leq \bar{R}), \tag{4.7}$$

leading to

$$p(\mathbf{\Gamma}) \propto q^{\#\mathbf{\Gamma}}(1-q)^{\binom{p}{2}-\#\mathbf{\Gamma}} \mathrm{P}(\bar{R} \geq \#\mathbf{\Gamma}). \tag{4.8}$$

In the second case, the prior on $\mathbf{\Gamma}$ is simply given by

$$p(\mathbf{\Gamma}) \propto q^{\#\mathbf{\Gamma}}(1-q)^{\binom{p}{2}-\#\mathbf{\Gamma}} \mathbb{1}(\#\mathbf{\Gamma} \leq \bar{r}). \tag{4.9}$$

Smaller values of q prefer graphical models with fewer number of edges, hence inducing more sparsity in the precision matrix.

Due to the positive definiteness constraint on the parameter $\mathbf{\Omega}$, the normalizing constant corresponding posterior distribution of the graphical model becomes intractable and hence was not explored in Wang (2012). One possible solution is to employ a reversible jump Markov chain Monte Carlo (RJMCMC) algorithm, which jumps from models of varying dimensions to evaluate the posterior probabilities. As there are as many as $2^{\binom{p}{2}}$ possible models, the posterior model probabilities estimated by RJMCMC visits are extremely unreliable. We consider a radically different approach to posterior computation based on Laplace approximations, elaborated in the next section.

Under the above prior specifications, the joint posterior distribution of $\mathbf{\Omega}$ and $\mathbf{\Gamma}$ given the

data $\boldsymbol{X}^{(n)} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ is given by

$$\begin{aligned} p\{\boldsymbol{\Omega}, \boldsymbol{\Gamma} | \boldsymbol{X}^{(n)}\} &\propto p\{\boldsymbol{X}^{(n)} | \boldsymbol{\Omega}, \boldsymbol{\Gamma}\} p(\boldsymbol{\Omega} | \boldsymbol{\Gamma}) p(\boldsymbol{\Gamma}) \\ &= (2\pi)^{np/2} \{\det(\boldsymbol{\Omega})\}^{n/2} \exp\left\{-n \operatorname{tr}(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Omega})/2\right\} \\ &\quad \times \prod_{\gamma_{ij}=1} \{\lambda \exp(-\lambda|\omega_{ij}|)/2\} \prod_{i=1}^{p} \{\lambda \exp\left(-\lambda\omega_{ii}/2\right)/2\} \\ &\quad \times p(\boldsymbol{\Gamma}) \mathbb{1}_{\mathcal{M}^+}(\boldsymbol{\Omega}). \end{aligned} \tag{4.10}$$

Thus,

$$p\{\boldsymbol{\Omega}, \boldsymbol{\Gamma} | \boldsymbol{X}^{(n)}\} \propto C_{\boldsymbol{\Gamma}} Q\{\boldsymbol{\Omega}, \boldsymbol{\Gamma} | \boldsymbol{X}^{(n)}\}, \tag{4.11}$$

where

$$\begin{aligned} C_{\boldsymbol{\Gamma}} &= (2\pi)^{np/2} q^{\#\boldsymbol{\Gamma}} (1-q)^{\binom{p}{2} - \#\boldsymbol{\Gamma}} (\lambda/2)^{p+\#\boldsymbol{\Gamma}} \beta(\boldsymbol{\Gamma}), \\ \beta(\boldsymbol{\Gamma}) &= \begin{cases} \mathrm{P}(\bar{R} \geq \#\boldsymbol{\Gamma}), \text{ for prior as in (4.8)}, \\ \mathbb{1}(\#\boldsymbol{\Gamma} \leq \bar{r}), \text{ for prior as in (4.9)}, \end{cases} \\ Q\{\boldsymbol{\Omega}, \boldsymbol{\Gamma} | \boldsymbol{X}^{(n)}\} &= \{\det(\boldsymbol{\Omega})\}^{n/2} \exp\left\{-n \operatorname{tr}(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Omega})/2\right\} \prod_{\gamma_{ij}=1} \{\exp(-\lambda|\omega_{ij}|)\} \\ &\quad \times \prod_{i=1}^{p} \{\exp\left(-\lambda\omega_{ii}/2\right)\} \mathbb{1}_{\mathcal{M}^+}(\boldsymbol{\Omega}). \end{aligned} \tag{4.12}$$

The following result gives posterior convergence rate as $n \to \infty$. We assume that the true model is sparse, as given by the class of positive definite matrices in (4.2).

Theorem 2. *Let $\boldsymbol{X}^{(n)} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ be a random sample from a p-dimensional Gaussian distribution with mean $\mathbf{0}$ and precision matrix $\boldsymbol{\Omega}_0 \in \mathcal{U}(\varepsilon_0, s)$ for some $0 < \varepsilon_0 < \infty$ and $0 \leq s \leq p(p-1)/2$. Also assume that the prior distributions $p(\boldsymbol{\Omega} \mid \boldsymbol{\Gamma})$ and $p(\boldsymbol{\Gamma})$ as in (4.5)*

and (4.8) or (4.9) with $q < 1/2$. Then the posterior distribution of $\boldsymbol{\Omega}$ satisfies

$$\mathrm{E}_0 \left[\mathrm{P} \left\{ \|\boldsymbol{\Omega} - \boldsymbol{\Omega}_0\|_2 > M\epsilon_n \mid \boldsymbol{X}^{(n)} \right\} \right] \to 0, \tag{4.13}$$

for $\epsilon_n = n^{-1/2}(p+s)^{1/2}(\log p)^{1/2}$ and a sufficiently large constant $M > 0$.

The proof uses the general theory of posterior convergence of Ghosal et al. (2000) and will be given in the appendix. The above posterior convergence rate matches exactly with the frequentist convergence rate of the penalized estimator obtained in Rothman et al. (2008).

Note that, Theorem 2 gives $\|\boldsymbol{\Omega} - \boldsymbol{\Omega}_0\|_2 = O(\epsilon_n)$ with posterior probability tending to one in probability and from Rothman et al. (2008) it follows that, $\|\boldsymbol{\Omega}^* - \boldsymbol{\Omega}_0\|_2 = O_P(\epsilon_n)$, where $\boldsymbol{\Omega}^*$ is the graphical lasso estimate. Hence, by the triangle inequality, $\|\boldsymbol{\Omega} - \boldsymbol{\Omega}^*\|_2 = O(\epsilon_n)$ with posterior probability tending to one in probability. This gives,

$$\frac{\int_{\|\boldsymbol{\Omega}-\boldsymbol{\Omega}^*\|_2 \le \epsilon_n} \exp\{-n\, h(\boldsymbol{\Omega})/2\} \prod_{(i,j)\in\mathcal{V}_{\boldsymbol{\Gamma}}} d\omega_{ij}}{\int_{\boldsymbol{\Omega}\in\mathcal{M}^+} \exp\{-n\, h(\boldsymbol{\Omega})/2\} \prod_{(i,j)\in\mathcal{V}_{\boldsymbol{\Gamma}}} d\omega_{ij}} \to 1. \tag{4.14}$$

4.3 Posterior Computation

The marginal posterior density of the graphical structure indicator $\boldsymbol{\Gamma}$ can be obtained by integrating out elements of the precision matrix in the joint posterior density in (4.10), to get

$$p\{\boldsymbol{\Gamma}|\boldsymbol{X}^{(n)}\} \propto C_{\boldsymbol{\Gamma}} \int_{\boldsymbol{\Omega}\in\mathcal{M}^+} \exp\{-n\, h(\boldsymbol{\Omega})/2\} \prod_{(i,j)\in\mathcal{V}_{\boldsymbol{\Gamma}}} d\omega_{ij}, \tag{4.15}$$

where

$$h(\boldsymbol{\Omega}) = -\log\det(\boldsymbol{\Omega}) + \mathrm{tr}(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Omega}) + \frac{2\lambda}{n} \sum_{\gamma_{ij}=1} |\omega_{ij}| + \frac{\lambda}{n} \sum_{i=1}^{p} \omega_{ii}. \tag{4.16}$$

Note that $h(\boldsymbol{\Omega})$ is minimized at $\boldsymbol{\Omega} = \boldsymbol{\Omega}^*$, the graphical lasso estimate corresponding to the penalty parameter $\rho = \lambda/n$. The marginal posterior of $\boldsymbol{\Gamma}$ is, however, intractable. We give an approximate method for the posterior probability computations of various models using Laplace approximation. The Laplace approximation requires expanding the integrand in (4.15) around the maximum, which in this case, coincides with the graphical lasso solution.

4.3.1 Approximating model posterior probabilities

Define $\boldsymbol{\Delta} = \boldsymbol{\Omega} - \boldsymbol{\Omega}^* = (\!(u_{ij})\!)$, where $\boldsymbol{\Omega}^*$ is the graphical lasso solution corresponding to the underlying graphical model structure and penalty parameter λ/n. Then,

$$\begin{aligned} p\{\boldsymbol{\Gamma}|\boldsymbol{X}\} \propto C_{\boldsymbol{\Gamma}} \exp\{-n\, h(\boldsymbol{\Omega}^*)/2\} \left\{\det(\boldsymbol{\Omega}^*)\right\}^{-n/2} \\ \times \int_{\boldsymbol{\Delta}+\boldsymbol{\Omega}^* \in \mathcal{M}^+} \exp\{-n\, g(\boldsymbol{\Delta})/2\} \prod_{(i,j)\in\mathcal{V}_{\boldsymbol{\Gamma}}} du_{ij}, \end{aligned} \tag{4.17}$$

where $g(\boldsymbol{\Delta})$ is

$$-\log\det(\boldsymbol{\Delta}+\boldsymbol{\Omega}^*) + \operatorname{tr}(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Delta}) + \frac{2\lambda}{n} \sum_{\gamma_{ij}=1} (|u_{ij}+\omega^*_{ij}| - |\omega^*_{ij}|) + \frac{\lambda}{n}\sum_{i=1}^{p} u_{ii}. \tag{4.18}$$

Clearly $g(\boldsymbol{\Delta})$ is minimized at $\boldsymbol{\Delta} = \mathbf{0}$ by the definition of $\boldsymbol{\Omega}^*$, so the first derivative of $g(\boldsymbol{\Delta})$ vanishes at $\mathbf{0}$, provided that it is differentiable at $\mathbf{0}$. Define the matrix $\boldsymbol{H}_{\boldsymbol{B}} = [h_{\boldsymbol{B}}\{(i,j),(l,m)\}]$, where

$$h_{\boldsymbol{B}}\{(i,j),(l,m)\} = \operatorname{tr}\left\{\boldsymbol{B}^{-1}\boldsymbol{E}_{(i,j)}\boldsymbol{B}^{-1}\boldsymbol{E}_{(l,m)}\right\}. \tag{4.19}$$

Using standard matrix calculus (for example, see Section 15.9 of Harville, 2008), we can find that the Hessian of $g(\boldsymbol{\Delta})$ is the $\#\mathcal{V}_{\boldsymbol{\Gamma}} \times \#\mathcal{V}_{\boldsymbol{\Gamma}}$ matrix $\boldsymbol{H}_{\boldsymbol{\Delta}+\boldsymbol{\Omega}^*}$, whose $\{(i,j),(l,m)\}$th entry for

$(i,j),(l,m) \in \mathcal{V}_{\mathbf{\Gamma}}$ is given by

$$\frac{\partial^2 g(\mathbf{\Delta})}{\partial u_{ij} \partial u_{lm}} = \operatorname{tr}\left\{(\mathbf{\Delta}+\mathbf{\Omega}^*)^{-1}\boldsymbol{E}_{(i,j)}(\mathbf{\Delta}+\mathbf{\Omega}^*)^{-1}\boldsymbol{E}_{(l,m)}\right\}. \tag{4.20}$$

Thus the Laplace approximation $p^*\{\mathbf{\Gamma} \mid \boldsymbol{X}^{(n)}\}$ to the posterior probability $p\{\mathbf{\Gamma} \mid \boldsymbol{X}^{(n)}\}$ is given by

$$\begin{aligned} p^*\{\mathbf{\Gamma}|\boldsymbol{X}^{(n)}\} &\propto C_{\mathbf{\Gamma}} \exp\{-n\, h(\mathbf{\Omega}^*)/2\} \left\{\det(\mathbf{\Omega}^*)\right\}^{-n/2} \exp\{-n\, g(\mathbf{0})/2\} \\ &\quad \times (2\pi)^{\#\mathcal{V}_{\mathbf{\Gamma}}/2}(n/2)^{-\#\mathcal{V}_{\mathbf{\Gamma}}/2}\left[\det\left\{\frac{\partial^2 g(\mathbf{\Delta})}{\partial \mathbf{\Delta}\partial \mathbf{\Delta}^T}\bigg|_{\mathbf{0}}\right\}\right]^{-1/2} \\ &= C_{\mathbf{\Gamma}} \exp\{-n\, h(\mathbf{\Omega}^*)/2\}(\pi/n)^{\#\mathcal{V}_{\mathbf{\Gamma}}/2}\{\det(\boldsymbol{H}_{\mathbf{\Omega}^*})\}^{-1/2}. \end{aligned} \tag{4.21}$$

The approximation in (4.21) is meaningful only if all the graphical lasso estimates of the off-diagonal elements corresponding to the graph generated by $\mathbf{\Gamma}$ are non-zero; otherwise the derivative of $g(\mathbf{\Delta})$ does not exist. A similar situation arises in the context of regression models; see Yuan and Lin (2005) and Curtis et al. (2014). In the next section, we show that such "non-regular models" can essentially be ignored for the purpose of posterior probability evaluation.

4.3.2 Ignorability of non-regular models

As discussed in the previous section, the objective function of the graphical lasso problem is not differentiable if the graphical lasso solution is zero for at least one pair $(i,j) \in E$. These models are referred to as non-regular models. This essentially means that given a fixed graphical structure index $\mathbf{\Gamma}$, the graphical lasso solution is $\omega_{ij}^* = 0$ for at least one $\gamma_{ij} = 1$. Let us assume, for notational simplicity, that the first t elements of $\mathbf{\Gamma}$ are 1 and the rest are 0. Also, among those t 1's, the last r of them have corresponding graphical lasso solution equal to zero. For such a non-regular model, we argue that the submodel $\mathbf{\Gamma}'$, with first $(t-r)$ 1's and rest 0's,

provides the same graphical lasso solution for the non-zero elements as the bigger model $\mathbf{\Gamma}$. This means that for (i, j) such that $\gamma_{ij} = \gamma'_{ij} = 1$, the graphical lasso solution corresponding to $\mathbf{\Gamma}$, given by $\omega^*_{\mathbf{\Gamma},ij}$ is identical with that corresponding to $\mathbf{\Gamma}'$, given by $\omega^*_{\mathbf{\Gamma}',ij}$. We refer to such a submodel $\mathbf{\Gamma}'$ as the regular submodel of the non-regular model $\mathbf{\Gamma}$.

Lemma 8. *For a submodel $\mathbf{\Gamma}'$ of $\mathbf{\Gamma}$ as defined above, the graphical lasso solution corresponding to the two models are identical.*

Proof. The graphical lasso solution for the model $\mathbf{\Gamma}$, given by $\mathbf{\Omega}^*_{\mathbf{\Gamma}} = (\!(\omega^*_{\mathbf{\Gamma},ij})\!)$ satisfies, by the Karush-Kuhn-Tucker (KKT) condition; see, for example, Boyd and Vandenberghe (2004), Witten et al. (2011);

$$\mathbf{\Omega}^{*-1}_{\mathbf{\Gamma}} - \widehat{\mathbf{\Sigma}} - \lambda \boldsymbol{G} = 0, \tag{4.22}$$

where $\boldsymbol{G}$ is a matrix with elements

$$\boldsymbol{G}_{ij} = \begin{cases} \omega^*_{\mathbf{\Gamma},ij}/|\omega^*_{\mathbf{\Gamma},ij}| & \text{if } \omega^*_{\mathbf{\Gamma},ij} \neq 0 \\ g_{ij} \in [-1, 1] & \text{if } \omega^*_{\mathbf{\Gamma},ij} = 0. \end{cases} \tag{4.23}$$

For a non-regular model $\mathbf{\Gamma}$, consider the non-zero elements $\omega^*_{\mathbf{\Gamma},ij}$ of the graphical lasso solution. Corresponding to the submodel Γ' of Γ, we construct a matrix $\mathbf{\Omega}^*_{\mathbf{\Gamma}'} = (\!(\omega_{\mathbf{\Gamma}',ij})\!)$ such that,

$$\omega^*_{\mathbf{\Gamma}',ij} = \begin{cases} \omega^*_{\mathbf{\Gamma},ij} & \text{if } \omega^*_{\mathbf{\Gamma},ij} \neq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{4.24}$$

Then, $\mathbf{\Omega}^*_{\mathbf{\Gamma}'}$ satisfies the KKT condition corresponding to the model $\mathbf{\Gamma}'$, and hence $\mathbf{\Omega}^*_{\mathbf{\Gamma}'}$ is a graphical lasso solution for $\mathbf{\Gamma}'$. But the construction of the above solution gives that $\mathbf{\Omega}^*_{\mathbf{\Gamma}} = \mathbf{\Omega}^*_{\mathbf{\Gamma}'}$. This completes the proof. □

We denote the graphical lasso solution in the non-regular model $\mathbf{\Gamma}$ and the corresponding regular submodel $\mathbf{\Gamma}'$ by $\mathbf{\Omega}^*$. For notational convenience, let us denote the precision matrix $\mathbf{\Omega}$ corresponding to the structure indicator $\mathbf{\Gamma}$ by $\mathbf{\Omega}_{\mathbf{\Gamma}}$, and corresponding matrix $\mathbf{\Delta}_{\mathbf{\Gamma}}$ is defined by $\mathbf{\Omega}_{\mathbf{\Gamma}} - \mathbf{\Omega}^* = (\!(u_{\mathbf{\Gamma},ij})\!)$. The ratio of the posterior model probabilities of the two model is given by,

$$\frac{p\{\mathbf{\Gamma}|\boldsymbol{X}^{(n)}\}}{p\{\mathbf{\Gamma}'|\boldsymbol{X}^{(n)}\}} = \frac{C_{\mathbf{\Gamma}} \int_{\mathbf{\Delta}_{\mathbf{\Gamma}}+\mathbf{\Omega}^* \in \mathcal{M}^+} \exp\{-n\, h(\mathbf{\Delta}_{\mathbf{\Gamma}})/2\} \prod_{(i,j)\in\mathcal{V}_{\mathbf{\Gamma}}} du_{\mathbf{\Gamma},ij}}{C_{\mathbf{\Gamma}'} \int_{\mathbf{\Delta}_{\mathbf{\Gamma}'}+\mathbf{\Omega}^* \in \mathcal{M}^+} \exp\{-n\, h(\mathbf{\Delta}_{\mathbf{\Gamma}'})/2\} \prod_{(i,j)\in\mathcal{V}_{\mathbf{\Gamma}'}} du_{\mathbf{\Gamma}',ij}}. \tag{4.25}$$

The following result shows the ignorability of the non-regular models.

Theorem 3. *Consider the prior on $\mathbf{\Gamma}$ as given in (4.8) or (4.9) with $q < 1/2$. The posterior probability of a non-regular model $\mathbf{\Gamma}$, as defined above, is always less than that of its regular submodel $\mathbf{\Gamma}'$.*

Proof. Using (4.14), we have,

$$\frac{p\{\mathbf{\Gamma}|\boldsymbol{X}^{(n)}\}}{p\{\mathbf{\Gamma}'|\boldsymbol{X}^{(n)}\}} = \frac{C_{\mathbf{\Gamma}} \int_{\|\mathbf{\Delta}_{\mathbf{\Gamma}}\|_2 \le \epsilon_n} \exp\{-n\, h(\mathbf{\Delta}_{\mathbf{\Gamma}})/2\} \prod_{(i,j)\in\mathcal{V}_{\mathbf{\Gamma}}} du_{\mathbf{\Gamma},ij} + o(1)}{C_{\mathbf{\Gamma}'} \int_{\|\mathbf{\Delta}_{\mathbf{\Gamma}'}\|_2 \le \epsilon_n} \exp\{-n\, h(\mathbf{\Delta}_{\mathbf{\Gamma}'})/2\} \prod_{(i,j)\in\mathcal{V}_{\mathbf{\Gamma}'}} du_{\mathbf{\Gamma}',ij} + o(1)}.$$

Now, note that for (i,j) such that $\gamma_{ij} = \gamma'_{ij} = 1$, we have,

$$\{u_{\mathbf{\Gamma},ij} \colon \|\mathbf{\Delta}_{\mathbf{\Gamma}}\|_2 \le \epsilon_n\} \subset \{u_{\mathbf{\Gamma}',ij} \colon \|\mathbf{\Delta}_{\mathbf{\Gamma}'}\|_2 \le \epsilon_n\}.$$

Hence, using Lemma 11, we get

$$\begin{aligned}
\frac{p\{\boldsymbol{\Gamma}|\boldsymbol{X}^{(n)}\}}{p\{\boldsymbol{\Gamma}'|\boldsymbol{X}^{(n)}\}} &\leq \frac{C_{\boldsymbol{\Gamma}}}{C_{\boldsymbol{\Gamma}'}} \int_{\|\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}\|_2 \leq \epsilon_n} \exp\left(-\frac{n}{2}\frac{2\lambda}{n} \sum_{\gamma_{ij}=1,\gamma'_{ij}=0} |u_{\boldsymbol{\Gamma},ij}|\right) \prod_{(i,j)\in\mathcal{V}_{\boldsymbol{\Gamma}}\cap\mathcal{V}^c_{\boldsymbol{\Gamma}'}} du_{\boldsymbol{\Gamma},ij} \\
&\leq \frac{C_{\boldsymbol{\Gamma}}}{C_{\boldsymbol{\Gamma}'}} \int \exp\left(-\lambda \sum_{\gamma_{ij}=1,\gamma'_{ij}=0} |u_{\boldsymbol{\Gamma},ij}|\right) \prod_{(i,j)\in\mathcal{V}_{\boldsymbol{\Gamma}}\cap\mathcal{V}^c_{\boldsymbol{\Gamma}'}} du_{\boldsymbol{\Gamma},ij} \\
&= \frac{C_{\boldsymbol{\Gamma}}}{C_{\boldsymbol{\Gamma}'}} \left(\frac{2}{\lambda}\right)^{\#\boldsymbol{\Gamma}-\#\boldsymbol{\Gamma}'} \\
&= \left(\frac{q}{1-q}\right)^r \frac{\beta(\boldsymbol{\Gamma})}{\beta(\boldsymbol{\Gamma}')} \\
&\leq \left(\frac{q}{1-q}\right)^r. \qquad (4.26)
\end{aligned}$$

The last inequality follows from the fact that if the prior as in (4.8) is used, then $\mathrm{P}(\bar{R} \geq \#\boldsymbol{\Gamma}) \leq \mathrm{P}(\bar{R} \geq \#\boldsymbol{\Gamma}')$ since $\#\boldsymbol{\Gamma} > \#\boldsymbol{\Gamma}'$. For the other prior as in (4.9), the inequality follows trivially as it involves the ratio of two indicator variables only.

For $q < 1/2$, the above ratio is less than 1. This completes the proof. □

The above result is particularly important in the sense that we can focus on the regular models only, ignoring the non-regular ones especially if q is chosen to be small. While approximating the posterior probabilities of the regular models, we re-normalize the values considering the regular models only.

4.3.3 Error in Laplace approximation

The approximation in the posterior probability of the graphical model is based on a Taylor series expansion of the function $h(\boldsymbol{\Omega})$ around the graphical lasso solution $\boldsymbol{\Omega}^*$. Let $\boldsymbol{\Delta} = \boldsymbol{\Omega} - \boldsymbol{\Omega}^*$, and $\mathrm{vec}(\boldsymbol{\Delta})$ denote the vectorized version of $\boldsymbol{\Delta}$, but excluding entries corresponding to the missing edges in the underlying graphical model. Thus $\mathrm{vec}(\boldsymbol{\Delta})$ is a vector of dimension $\#\mathcal{V}_{\boldsymbol{\Gamma}}$

corresponding to the graphical structure indicator $\mathbf{\Gamma}$. If the graphical model is s-sparse, that is, there are s edges present in the graph, then $\#\mathcal{V}_{\mathbf{\Gamma}} = p + s$. The following result gives the bound on the remainder term of the Taylor series expansion under the above assumptions.

Lemma 9. *Consider a graphical model with p variables such that the graph is s-sparse. Then, with probability tending to 1, the remainder term in the expansion of the function $h(\mathbf{\Omega})$ as defined in (4.16), around the graphical lasso solution $\mathbf{\Omega}^*$, is bounded by*

$$\frac{1}{2}(p+s)\|\mathbf{\Delta}\|_2^2 \left(C_1\|\mathbf{\Delta}\|_2 + C_2\|\mathbf{\Delta}\|_2^2\right),$$

where $\mathbf{\Delta} = \mathbf{\Omega} - \mathbf{\Omega}^*$.

This result can be used to find a bound for the error in Laplace approximation of the posterior probabilities of the graphical model structures. The following result gives the condition for which the error in approximation is asymptotically negligible.

Theorem 4. *The error in Laplace approximation of the posterior probability of a graphical model structure is asymptotically negligible if $(p+s)^2\epsilon_n \to 0$, where ϵ_n is the posterior convergence rate, that is, the error in the Laplace approximation tends to zero if $n^{-1/2}(p+s)^{5/2}(\log p)^{1/2} \to 0$.*

The proof of the above result depends on Lemma 9 involving the bound on the remainder term in the Taylor series expansion of $h(\mathbf{\Omega})$. We give proofs of the above results in Section 4.6.

4.4 Simulation results

We perform a simulation study to assess the performance of the Bayesian method for graphical structure learning. We use 4 different models for our simulations, and we specify these models

in terms of the elements of the covariance matrix $\mathbf{\Sigma} = (\!(\sigma_{ij})\!)$ or the precision matrix $\mathbf{\Omega} = (\!(\omega_{ij})\!)$, as follows:

1. Model 1: AR(1) model, $\sigma_{ij} = 0.7^{|i-j|}$.

2. Model 2: AR(2) model, $\omega_{ii} = 1, \omega_{i,i-1} = \omega_{i-1,i} = 0.5, \omega_{i,i-2} = \omega_{i-2,i} = 0.25$.

3. Model 3: Star model, where every node is connected to the first node, and $\omega_{ii} = 1, \omega_{1,i} = \omega_{i,1} = 0.1$, and $\omega_{ij} = 0$ otherwise.

4. Model 4: Circle model, $\omega_{ii} = 2, \omega_{i,i-1} = \omega_{i,i-1} = 1, \omega_{1,p} = \omega_{p,1} = 0.9$.

Corresponding to each model, we generate samples of size $n = 100, 200$ and dimension $p = 30, 50, 100$. The penalty parameter for the graphical lasso algorithm is chosen to be 0.5 and the value of q appearing in the prior of the graphical structure indicator to be 0.4. We run 100 replications for each of the models and find the median probability model for each replication. To assess the performance of the median probability model (denoted by 'MPP'), we compute the specificity, sensitivity and Matthews Correlation Coefficient (MCC) averaged across the replications as defined below and also compute the same for the graphical lasso (denoted by 'GL'). The results are presented in Table 4.1.

$$\begin{aligned} \text{SP} &= \frac{\text{TN}}{\text{TN}+\text{FP}}, \text{SE} = \frac{\text{TP}}{\text{TP}+\text{FN}} \\ \text{MCC} &= \frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}}, \end{aligned} \tag{4.27}$$

where TP, TN, FP and FN respectively denote the true positives, true negatives, false positives and false negatives in the selected model, which in our case is the median probability model.

Table 4.1: Simulation results for different structures of precision matrices

		$n = 100$						$n = 200$					
		MPP			GL			MPP			GL		
Model	p	SP	SE	MCC	SP	SE	MCC	SP	SE	MCC	SP	SE	MCC
	30	0.977	0.941	0.831	0.961	0.983	0.784	0.986	0.996	0.907	0.969	1.000	0.823
		(0.003)	(0.019)	(0.015)	(0.003)	(0.010)	(0.013)	(0.002)	(0.003)	(0.014)	(0.002)	(0.000)	(0.013)
AR(1)	50	0.987	0.953	0.841	0.977	0.986	0.785	0.991	0.992	0.903	0.980	1.000	0.823
		(0.002)	(0.013)	(0.010)	(0.001)	(0.004)	(0.010)	(0.001)	(0.004)	(0.008)	(0.001)	(0.000)	(0.006)
	100	0.992	0.967	0.837	0.989	0.991	0.804	0.994	0.995	0.890	0.991	0.999	0.827
		(0.001)	(0.008)	(0.007)	(0.001)	(0.003)	(0.006)	(0.001)	(0.002)	(0.008)	(0.001)	(0.001)	(0.006)
	30	0.975	0.470	0.546	0.964	0.535	0.558	0.987	0.495	0.617	0.982	0.517	0.610
		(0.003)	(0.014)	(0.013)	(0.002)	(0.013)	(0.012)	(0.002)	(0.008)	(0.008)	(0.002)	(0.009)	(0.007)
AR(2)	50	0.983	0.462	0.541	0.971	0.508	0.522	0.993	0.489	0.629	0.987	0.534	0.622
		(0.001)	(0.013)	(0.011)	(0.002)	(0.010)	(0.009)	(0.001)	(0.005)	(0.007)	(0.001)	(0.001)	(0.006)
	100	0.989	0.470	0.537	0.980	0.531	0.514	0.995	0.484	0.624	0.993	0.529	0.624
		(0.001)	(0.006)	(0.006)	(0.001)	(0.007)	(0.007)	(0.001)	(0.006)	(0.004)	(0.001)	(0.009)	(0.005)

Table 4.1: (continued)

		$n = 100$						$n = 200$					
		MPP			GL			MPP			GL		
Model	p	SP	SE	MCC	SP	SE	MCC	SP	SE	MCC	SP	SE	MCC
	30	0.947	0.289	0.228	0.937	0.310	0.224	0.995	0.210	0.378	0.993	0.252	0.402
		(0.004)	(0.038)	(0.036)	(0.003)	(0.043)	(0.036)	(0.001)	(0.032)	(0.041)	(0.001)	(0.036)	(0.038)
Star	50	0.945	0.492	0.332	0.934	0.514	0.317	0.993	0.475	0.585	0.990	0.514	0.577
		(0.003)	(0.034)	(0.025)	(0.003)	(0.035)	(0.023)	(0.000)	(0.034)	(0.024)	(0.001)	(0.032)	(0.022)
	100	0.939	1.000	0.485	0.927	1.000	0.452	0.988	1.000	0.792	0.984	1.000	0.748
		(0.002)	(0.000)	(0.007)	(0.002)	(0.000)	(0.005)	(0.000)	(0.000)	(0.008)	(0.001)	(0.000)	(0.007)
	30	0.733	1.000	0.399	0.694	1.000	0.369	0.719	1.000	0.388	0.674	1.000	0.354
		(0.004)	(0.000)	(0.003)	(0.006)	(0.000)	(0.004)	(0.005)	(0.000)	(0.004)	(0.004)	(0.000)	(0.003)
Circle	50	0.831	1.000	0.409	0.822	1.000	0.398	0.833	1.000	0.411	0.814	1.000	0.390
		(0.003)	(0.000)	(0.003)	(0.002)	(0.000)	(0.003)	(0.002)	(0.000)	(0.003)	(0.002)	(0.000)	(0.002)
	100	0.891	1.000	0.378	0.894	1.000	0.383	0.903	1.000	0.399	0.902	1.000	0.397
		(0.001)	(0.000)	(0.002)	(0.001)	(0.000)	(0.002)	(0.008)	(0.000)	(0.002)	(0.001)	(0.000)	(0.002)

4.5 Illustration with real data

In this section we illustrate the Bayesian graphical structure learning method with the stock price data from Yahoo! Finance. Description of the data set can be found in Liu et al. (2009) and available in the `huge` package on `CRAN` (Zhao et al., 2012) as `stockdata`. The data set consists of closing prices of stocks that were consistently included in the S&P 500 index in the time period January 1, 2003 to January 1, 2008 for a total of 1258 days. The stocks are also categorized into 10 Global Industry Classification Standard (GICS) sectors, namely, "Health Care", "Materials", "Industrials", "Consumer Staples", "Consumer Discretionary", "Utilities", "Information Technology", "Financials", "Energy", "Telecommunication Services".

Denoting Y_{tj} as the closing stock price for the jth stock on day t, we construct the 1257×452 data matrix $\boldsymbol{S}$ with entries $s_{tj} = \log(Y_{(t+1)j}/Y_{tj})$, $t = 1, \ldots, 1257$, $j = 1, \ldots, 452$. For analysis, we construct the data matrix $\boldsymbol{X}$ by standardizing $\boldsymbol{S}$, so that each stock has mean zero and standard deviation one. We find the median probability model as selected by the Bayesian graphical structure learning method. The corresponding graphical structure is displayed in Figure 4.1. The vertices of the graph are colored corresponding to the different GICS sectors. We find that stocks from the same sectors tend to be related with other members from that category, and generally not related across different sectors, though there are some connections, which may be due to some other possible latent factors affecting all of them. The grouping of the stocks corresponding to their sectors is expected, implying that the stock prices for a particular sector are conditionally independent of those of other sectors.

We also individually study data pertaining to some of the specific sectors to have a closer look at the strength of the groupings where perturbations due to latent factors is least expected. For this, we consider the sectors "Utilities" and "Information Technology". The graphical structure is displayed in Figure 4.2. The stock prices for the two sectors clearly separate as

desired.

4.6 Proofs and additional results

We first give a proof of the result on posterior convergence rate of the precision matrix.

Proof of Theorem 2. In order to establish the rates of convergence of the posterior distribution, we first need to check the prior concentration rate, that is,

$$\Pi\left\{B(p_{\boldsymbol{\Omega}_0}, \epsilon_n)\right\} := \Pi\left\{p\colon K(p_{\boldsymbol{\Omega}_0}, p_{\boldsymbol{\Omega}}) \leq \epsilon_n^2, V(p_{\boldsymbol{\Omega}_0}, p_{\boldsymbol{\Omega}}) \leq \epsilon_n^2\right\} \geq \exp(-bn\epsilon_n^2), \tag{4.28}$$

where $K(p_{\boldsymbol{\Omega}_0}, p_{\boldsymbol{\Omega}}) = \int p_{\boldsymbol{\Omega}_0} \log\left(p_{\boldsymbol{\Omega}_0}/p_{\boldsymbol{\Omega}}\right), V(p_{\boldsymbol{\Omega}_0}, p_{\boldsymbol{\Omega}}) = \int p_{\boldsymbol{\Omega}_0} \left\{\log\left(p_{\boldsymbol{\Omega}_0}/p_{\boldsymbol{\Omega}}\right)\right\}^2$. Note that, for $\boldsymbol{Z} \sim \mathrm{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ and a $p \times p$ symmetric matrix $\boldsymbol{A}$, we have,

$$\mathrm{E}(\boldsymbol{Z}^T\boldsymbol{A}\boldsymbol{Z}) = \mathrm{tr}(\boldsymbol{A}\boldsymbol{\Sigma}), \quad \mathrm{Var}(\boldsymbol{Z}^T\boldsymbol{A}\boldsymbol{Z}) = 2\,\mathrm{tr}(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma}). \tag{4.29}$$

We use the above result to find the expressions for $K(p_{\boldsymbol{\Omega}_0}, p_{\boldsymbol{\Omega}})$ and $V(p_{\boldsymbol{\Omega}_0}, p_{\boldsymbol{\Omega}})$. Denoting the eigenvalues of the matrix $\boldsymbol{\Omega}_0^{-1/2}\boldsymbol{\Omega}\boldsymbol{\Omega}_0^{-1/2}$ by d_i, $i = 1, \ldots, p$, using $\mathrm{tr}(\boldsymbol{A}\boldsymbol{B}) = \mathrm{tr}(\boldsymbol{B}\boldsymbol{A})$ and (4.29), we get,

$$\begin{aligned} K(p_{\boldsymbol{\Omega}_0}, p_{\boldsymbol{\Omega}}) &= \frac{1}{2}(\log\det\boldsymbol{\Omega}_0 - \log\det\boldsymbol{\Omega}) - \frac{1}{2}\mathrm{tr}(\boldsymbol{I}_p - \boldsymbol{\Omega}\boldsymbol{\Omega}_0^{-1}) \\ &= \frac{1}{2}(\log\det\boldsymbol{\Omega}_0 - \log\det\boldsymbol{\Omega}) - \frac{1}{2}\mathrm{tr}(\boldsymbol{I}_p - \boldsymbol{\Omega}_0^{-1/2}\boldsymbol{\Omega}\boldsymbol{\Omega}_0^{-1/2}) \\ &= -\frac{1}{2}\sum_{i=1}^{p}\log d_i - \frac{1}{2}\sum_{i=1}^{p}(1 - d_i). \end{aligned} \tag{4.30}$$

Now, $K(p_{\boldsymbol{\Omega}_0}, p_{\boldsymbol{\Omega}}) \geq h^2(p_{\boldsymbol{\Omega}_0}, p_{\boldsymbol{\Omega}})$, and from an argument in Lemma 10 it implies that if $K(p_{\boldsymbol{\Omega}_0}, p_{\boldsymbol{\Omega}}) \leq \epsilon_n^2$, then $\max_i |d_i - 1| < 1$. Hence we can expand $\log d_i$ in the powers of

$(1 - d_i)$ to get

$$K(p_{\boldsymbol{\Omega}_0}, p_{\boldsymbol{\Omega}}) \sim \frac{1}{4}\sum_{i=1}^{p}(1 - d_i)^2. \tag{4.31}$$

Also,

$$\begin{aligned} V(p_{\boldsymbol{\Omega}_0}, p_{\boldsymbol{\Omega}}) &= \frac{1}{2}\mathrm{tr}(\boldsymbol{I}_p - 2\boldsymbol{\Omega}\boldsymbol{\Omega}_0^{-1} + \boldsymbol{\Omega}\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Omega}\boldsymbol{\Omega}_0^{-1}) \\ &= \mathrm{tr}(\boldsymbol{I}_p - 2\boldsymbol{\Omega}_0^{-1/2}\boldsymbol{\Omega}\boldsymbol{\Omega}_0^{-1/2} + \boldsymbol{\Omega}_0^{-1/2}\boldsymbol{\Omega}\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Omega}\boldsymbol{\Omega}_0^{-1/2}) \\ &= \frac{1}{2}\mathrm{tr}(\boldsymbol{I}_p - \boldsymbol{\Omega}_0^{-1/2}\boldsymbol{\Omega}\boldsymbol{\Omega}_0^{-1/2})^2 \\ &= \frac{1}{2}\sum_{i=1}^{p}(1 - d_i)^2. \end{aligned}$$

Thus,

$$\Pi\left\{p\colon K(p_{\boldsymbol{\Omega}_0}, p_{\boldsymbol{\Omega}}) \leq \epsilon_n^2, V(p_{\boldsymbol{\Omega}_0}, p_{\boldsymbol{\Omega}}) \leq \epsilon_n^2\right\} \geq \Pi\left\{\sum_{i=1}^{p}(1 - d_i)^2 \leq 8\epsilon_n^2\right\}. \tag{4.32}$$

Now, using the assumptions on the true precision matrix $\boldsymbol{\Omega}_0$ and the matrix norm relations given by Lemma 1, we have,

$$\begin{aligned} \sum_{i=1}^{p}(1 - d_i)^2 &= \|\boldsymbol{I}_p - \boldsymbol{\Omega}_0^{-1/2}\boldsymbol{\Omega}\boldsymbol{\Omega}_0^{-1/2}\|_2^2 \\ &= \|\boldsymbol{\Omega}_0^{-1/2}(\boldsymbol{\Omega}_0 - \boldsymbol{\Omega})\boldsymbol{\Omega}_0^{-1/2}\|_2^2 \\ &\leq \|\boldsymbol{\Omega}_0^{-1}\|_{(2,2)}^2\|\boldsymbol{\Omega}_0 - \boldsymbol{\Omega}\|_2^2 \\ &\leq \varepsilon_0^{-2}\|\boldsymbol{\Omega}_0 - \boldsymbol{\Omega}\|_2^2. \end{aligned} \tag{4.33}$$

Hence, equations (4.32) and (4.33), along with Lemma 1 give,

$$\begin{aligned}\Pi\left\{p\colon K(p_{\mathbf{\Omega}_0}, p_{\mathbf{\Omega}}) \leq \epsilon_n^2, V(p_{\mathbf{\Omega}_0}, p_{\mathbf{\Omega}}) \leq \epsilon_n^2\right\} &\geq \Pi\left\{\|\mathbf{\Omega}_0 - \mathbf{\Omega}\|_2^2 \leq c\epsilon_n^2\right\} \\ &\geq \Pi\left(\|\mathbf{\Omega}_0 - \mathbf{\Omega}\|_\infty \leq c'\epsilon_n/p\right).\end{aligned}$$

The components of $\mathbf{\Omega}$ are not independently distributed, but a truncation applies because of the positive definite restriction. However, as the true $\mathbf{\Omega}_0$ lies in the set of positive definite matrices which is open, the truncation can only increase concentration in a small ball centered at the truth, so we can pretend componentwise independence for the purpose of lower bounding the above prior probability. This gives

$$\Pi\left(\|\mathbf{\Omega}_0 - \mathbf{\Omega}\|_\infty \leq c'\epsilon_n/p\right) \gtrsim (c'\epsilon_n/p)^{p+s}. \tag{4.34}$$

The prior concentration rate condition thus gives,

$$(p+s)(\log p + \log\frac{1}{\epsilon_n}) \asymp n\epsilon_n^2, \tag{4.35}$$

so as to get $\epsilon_n = n^{-1/2}(p+s)^{1/2}(\log n)^{1/2}$.

Next, we need to construct tests for $H_0\colon \mathbf{\Omega} = \mathbf{\Omega}_0$ against the alternative $H_1\colon \|\mathbf{\Omega}_0 - \mathbf{\Omega}\|_2 \geq \epsilon_n$. Let $p_{\mathbf{\Omega},n}(X_1, \ldots, X_n)$ denote the joint density of the observations and let Π be the prior. From the results of Birgé (1984) and Le Cam (1986), we know that for a probability measure P_0 and any convex set $\mathcal{P}$ of probability measures, there exists tests ϕ_n such that

$$P_0^n\phi_n \leq e^{-nh^2(P_0,\mathcal{P})/2}, \sup_{P\in\mathcal{P}} P^n(1-\phi_n) \leq e^{-nh^2(P_0,\mathcal{P})/2}, \tag{4.36}$$

where $h^2(P_0, \mathcal{P}) = \min\{h^2(P_0, P)\colon P \in \mathcal{P}\}$.

Let $\mathcal{P}_2 = \{p_{\mathbf{\Omega}_2} \colon \|\mathbf{\Omega}_2 - \mathbf{\Omega}_1\|_2 \leq c_0^{-1/2}\epsilon_n/2\}$, where c_0 is the constant appearing in Lemma 10 and $h(p_{\mathbf{\Omega}_0}, p_{\mathbf{\Omega}_1}) \geq \epsilon_n$. We claim that $h(p_{\mathbf{\Omega}}, p_{\mathbf{\Omega}_0}) > \epsilon_n/2$ for any $p_{\mathbf{\Omega}}$ in the convex hull $\mathrm{conv}(\mathcal{P}_2)$ of $\mathcal{P}_2$. To see this, represent $p_{\mathbf{\Omega}}$ as

$$p_{\mathbf{\Omega}} = \int_{\mathbf{\Omega}_2 : \|\mathbf{\Omega}_2 - \mathbf{\Omega}_1\|_2 \leq c_0^{-1/2}\epsilon_n/2} p_{\mathbf{\Omega}_2} \, d\Phi(\mathbf{\Omega}_2), \tag{4.37}$$

where Φ is an arbitrary probability measure on $c_0^{-1/2}\epsilon_n/2$-ball around $\mathbf{\Omega}_1$ in terms of Frobenius distance. Then, for any $p_{\mathbf{\Omega}} \in \mathrm{conv}(\mathcal{P}_2)$, by Lemma 10 and the convexity of the squared Hellinger distance,

$$h^2(p_{\mathbf{\Omega}}, p_{\mathbf{\Omega}_1}) \leq \int_{\mathbf{\Omega}_2 : \|\mathbf{\Omega}_2 - \mathbf{\Omega}_1\|_2 \leq c_0^{-1/2}\epsilon_n/2} h^2(p_{\mathbf{\Omega}_1}, p_{\mathbf{\Omega}_2}) \, d\Phi(\mathbf{\Omega}_2) < \epsilon_n^2/4. \tag{4.38}$$

This implies that $h\{p_{\mathbf{\Omega}_0}, \mathrm{conv}(\mathcal{P}_2)\} > \epsilon_n/2$ by the triangle inequality. Thus, by (4.36), we can find tests for $\mathbf{\Omega} = \mathbf{\Omega}_0$ vs. $\mathbf{\Omega} \in \mathcal{P}_2$ such that the error probabilities are bounded by $\exp(-n\epsilon_n^2/8)$.

In order to get a test for H_0 vs. H_1 with similar error probability, we also need to cover the alternative with balls of size $\epsilon_n/2$ and satisfy the metric entropy condition

$$\log N(\epsilon_n/2, \mathcal{P}_n, \|\cdot\|_2') \leq c_1 n \epsilon_n^2, \tag{4.39}$$

where $\|\cdot\|_2'$ is the distance on $p_{\mathbf{\Omega}}$ induced by $\|\cdot\|_2$ on $\mathbf{\Omega}$, $c_1 > 0$ is a constant and $\mathcal{P}_n \subset \mathcal{P}$ is a suitable subset of $\mathcal{P}$, called a sieve, such that $\Pi(\mathcal{P}_n^c)$ is exponentially small. For a graph G with p vertices, consider the sieve $\mathcal{P}_n$ to be the space of all densities $p_{\mathbf{\Omega}}$ such that the graph corresponding to $\mathbf{\Omega}$ has maximum number of edges $\bar{r} < \binom{p}{2}/2$ and each off-diagonal entry of

$\mathbf{\Omega}$ is at most L. Then the metric entropy condition is given by

$$\log\left\{\sum_{j=1}^{\bar{r}}\left(\frac{L}{\epsilon_n}\right)^j\binom{\binom{p}{2}}{j}\right\} \leq \log\left\{\bar{r}\left(\frac{L}{\epsilon_n}\right)^{\bar{r}}\binom{\binom{p}{2}}{\bar{r}}\right\}, \tag{4.40}$$

where we choose $L \in [b_2 n\epsilon_n^2, b_2 n\epsilon_n^2 + 1]$ to ensure that $\binom{p}{2}\exp(-L) \leq \exp(-b_3 n\epsilon_n^2)$, for some constants b_2 and b_3, and that b_3 can be made as large as we want by making b_2 larger. Thus the best solution of (4.40) leads to the relation

$$\log\bar{r} + \bar{r}\log p + \bar{r}\log(\frac{1}{\epsilon_n}) + \bar{r}\log(n\epsilon_n^2) \asymp n\epsilon_n^2, \tag{4.41}$$

which is satisfied if we choose $\bar{r} = b_1 n\epsilon_n^2/\log n$, b_1 large. Also, for this choice of $\bar{r}$, we have,

$$\mathrm{P}(\bar{R} > \bar{r}) \leq \exp(-a_2' b_1 n\epsilon_n^2), \tag{4.42}$$

where $a_2' b_1$ can be made as large as possible by making b_1 large. For the bound on the prior probability of the complement $\mathcal{P}_n^c$ of the above sieve, we have, using the condition on prior (4.8),

$$\Pi(\mathcal{P}_n^c) \leq \mathrm{P}(\bar{R} > \bar{r}) + \exp(-b_3 n\epsilon_n^2). \tag{4.43}$$

For the prior (4.9), the first term in (4.43) is exactly zero, and for the prior (4.8), from equation (4.42),

$$\Pi(\mathcal{P}_n^c) \leq \exp(-c_3 n\epsilon_n^2), \tag{4.44}$$

where c_3 is a constant which can be made as large as we please by making b_1, b_3 larger. Note that under the condition $n\epsilon_n^2/\log n \ll \binom{p}{2}$, the requirement $\bar{r} < \binom{p}{2}/2$ is satisfied as $n \to \infty$. Hence ϵ_n as found above is the desired posterior convergence rate. □

The following lemma establishes a norm equivalence necessary for finding posterior convergence rate and metric entropy calculations.

Lemma 10. *If $p_{\boldsymbol{\Omega}_k}$ is the density of $\mathrm{N}_p(\mathbf{0}, \boldsymbol{\Omega}_k^{-1})$, $k = 1, 2$, then for all $\boldsymbol{\Omega}_k \in \mathcal{U}(\varepsilon_0, s)$, $k = 1, 2$,*

(i) $c_0^{-1}\|\boldsymbol{\Omega}_1 - \boldsymbol{\Omega}_2\|_2^2 \leq h^2(p_{\boldsymbol{\Omega}_1}, p_{\boldsymbol{\Omega}_2})$, *when* $\|\boldsymbol{\Omega}_1 - \boldsymbol{\Omega}_2\|_2 < \varepsilon_0$,

(ii) $h^2(p_1, p_2) \leq c_0\|\boldsymbol{\Omega}_1 - \boldsymbol{\Omega}_2\|_2^2$,

for some universal constant $c_0 > 0$.

Proof. Let d_i, $i = 1, \ldots, p$ be the eigenvalues of the matrix $\boldsymbol{A} = \boldsymbol{\Omega}_1^{-1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{-1/2}$. Half squared Hellinger distance between p_1 and p_2 is given by

$$1 - \frac{\{\det(\boldsymbol{A})\}^{-1/4}}{\left(\det\left[\frac{1}{2}\{\boldsymbol{I} + \boldsymbol{A}^{-1}\}\right]\right)^{1/2}} = 1 - \frac{\prod_{i=1}^p d_i^{-1/4}}{\{\prod_{i=1}^p \frac{1}{2}(1 + d_i^{-1})\}^{1/2}}, \tag{4.45}$$

and the Frobenius norm of the difference between $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$ is given by, from Lemma 1,

$$\begin{aligned}
\|\boldsymbol{\Omega}_1 - \boldsymbol{\Omega}_2\|_2^2 &= \|\boldsymbol{\Omega}_1^{1/2}(\boldsymbol{I}_p - \boldsymbol{A})\boldsymbol{\Omega}_1^{1/2}\|_2^2 \\
&\leq \|\boldsymbol{\Omega}_1\|_{(2,2)}^2\|\boldsymbol{I}_p - \boldsymbol{A}\|_2^2 \\
&= \|\boldsymbol{\Omega}_1\|_{(2,2)}^2 \mathrm{tr}\,(\boldsymbol{I}_p - \boldsymbol{A})^2 \\
&= \|\boldsymbol{\Omega}_1\|_{(2,2)}^2 \sum_{i=1}^p (d_i - 1)^2 \\
&\leq \varepsilon_0^{-2} \sum_{i=1}^p (d_i - 1)^2. \qquad (4.46)
\end{aligned}$$

First we show that either $\|\boldsymbol{\Omega}_1 - \boldsymbol{\Omega}_2\|_2^2 \leq \delta^2$ or $h^2(p_{\boldsymbol{\Omega}_1}, p_{\boldsymbol{\Omega}_2}) \leq 2\delta^2$ implies $|d_i - 1| < 1$ for all $i = 1, \ldots, p$ for sufficiently small δ. This is necessary to expand d_i in powers of $(1 - d_i)$.

Let us consider the case $\|\boldsymbol{\Omega}_1 - \boldsymbol{\Omega}_2\|_2^2 \leq \delta^2$. Then,

$$\begin{aligned} \max_i |d_i - 1| &= \|\boldsymbol{A} - \boldsymbol{I}_p\|_{(2,2)} \\ &= \|\boldsymbol{\Omega}_1^{-1/2}(\boldsymbol{\Omega}_2 - \boldsymbol{\Omega}_1)\boldsymbol{\Omega}_1^{-1/2}\|_{(2,2)} \\ &\leq \|\boldsymbol{\Omega}_1^{-1}\|_{(2,2)}\|\boldsymbol{\Omega}_2 - \boldsymbol{\Omega}_1\|_{(2,2)} \\ &\leq \varepsilon_0^{-1}\delta < 1. \end{aligned} \tag{4.47}$$

Now let $h^2(p_{\boldsymbol{\Omega}_1}, p_{\boldsymbol{\Omega}_2}) \leq 2\delta^2$. This implies $1 - \{\prod_{i=1}^p \frac{1}{2}(d_i^{1/2} + d_i^{-1/2})\}^{-1/2} \leq \delta^2$. Rearranging the terms, we get, $\prod_{i=1}^p \frac{1}{2}(d_i^{1/2} + d_i^{-1/2}) \leq (1-\delta^2)^{-2} = 1 + \eta$, say. Since every term in the product exceeds 1, we have,

$$\max_i \frac{1}{2}(d_i^{1/2} + d_i^{-1/2}) \leq 1 + \eta. \tag{4.48}$$

The above equation, upon squaring and rearrangement of terms, gives, for all i,

$$(d_i - 1)^2 \leq 2d_i^{1/2}\eta. \tag{4.49}$$

Note that equation (4.48) gives that $d_i^{1/2} \leq 2(1+\eta)$. Hence, the above equation implies that $(d_i - 1)^2 \leq 4\eta(1+\eta)$. Choose $\eta < (\sqrt{2}-1)/2$ so that we get $|d_i - 1| < 1$ for all $i = 1, \ldots, p$.

Now, let us assume that $\frac{1}{2}h^2(p_{\boldsymbol{\Omega}_1}, p_{\boldsymbol{\Omega}_2}) \leq \delta^2$, for some $\delta > 0$. This implies, from equation (4.45),

$$\prod_{i=1}^p (d_i^{1/2} + d_i^{-1/2}) \leq 2^p(1-\delta^2)^{-2}.$$

Now, $\prod_{i=1}^p (d_i^{1/2} + d_i^{-1/2}) = 2^p[1 + O\{\sum_{i=1}^p (d_i - 1)^2\}]$ using Taylor's series expansion. Then, from the above equation after rearrangement of the terms, we get, $1 + O\left\{\sum_{i=1}^p (d_i - 1)^2\right\} \leq$

$(1-\delta^2)^{-2} \sim 1+2\delta^2$, so that,

$$\sum_{i=1}^{p}(d_i-1)^2 \leq c_0\delta^2, \text{ for some } c_0 > 0. \tag{4.50}$$

Now, equation (4.46) gives that $\|\boldsymbol{\Omega}_1 - \boldsymbol{\Omega}_2\|_2^2 \leq \|\boldsymbol{\Omega}_1\|_{(2,2)}^2 \sum_{i=1}^{p}(1-d_i)^2$. Choosing $\delta = h(p_{\boldsymbol{\Omega}_1}, p_{\boldsymbol{\Omega}_2})$, the first inequality follows.

To show the other way round, assume that $\|\boldsymbol{\Omega}_1 - \boldsymbol{\Omega}_2\|_2^2 \leq \delta^2$. Then,

$$\begin{aligned}
\frac{1}{2}h^2(p_{\boldsymbol{\Omega}_1}, p_{\boldsymbol{\Omega}_2}) &= 1 - \frac{\prod_{i=1}^{p} d_i^{-1/4}}{\{\prod_{i=1}^{p}\frac{1}{2}(1+d_i^{-1})\}^{1/2}} \\
&= 1 - \frac{1}{[1+O\{\sum_{i=1}^{p}(d_i-1)^2\}]^{1/2}} \\
&= O\left\{\sum_{i=1}^{p}(d_i-1)^2\right\} \leq c\delta^2, \text{ for some } c > 0.
\end{aligned} \tag{4.51}$$

Thus, if $\|\boldsymbol{\Omega}_1 - \boldsymbol{\Omega}_2\|_2^2 \leq \delta^2$, then $h^2(p_{\boldsymbol{\Omega}_1}, p_{\boldsymbol{\Omega}_2}) \leq c\delta^2$ for some $c > 0$. □

The following lemma is essential in proving the ignorability of the non-regular models for posterior probability evaluation.

Lemma 11. *Consider a non-regular model $\boldsymbol{\Gamma}$ with the corresponding regular submodel $\boldsymbol{\Gamma}'$, having identical graphical lasso estimate given by $\boldsymbol{\Omega}^*$. For $\boldsymbol{\Delta}_{\boldsymbol{\Gamma}} = (\!(u_{\boldsymbol{\Gamma},ij})\!)$ as defined in Section 4.3.2, for fixed values of $u_{\boldsymbol{\Gamma},ij} \in \{u_{\boldsymbol{\Gamma},ij} : \gamma_{ij} = \gamma'_{ij} = 1, \|\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}\|_2 \leq \epsilon_n\}$, we have,*

$$\log\det(\boldsymbol{\Delta}_{\boldsymbol{\Gamma}} + \boldsymbol{\Omega}^*) - \operatorname{tr}(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}) \leq \log\det(\boldsymbol{\Delta}_{\boldsymbol{\Gamma}'} + \boldsymbol{\Omega}^*) - \operatorname{tr}(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Delta}_{\boldsymbol{\Gamma}'}). \tag{4.52}$$

Proof. Consider maximization of the function

$$f(\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}) = \log\det(\boldsymbol{\Delta}_{\boldsymbol{\Gamma}} + \boldsymbol{\Omega}^*) - \operatorname{tr}(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}). \tag{4.53}$$

with respect to the elements $u_{\mathbf{\Gamma},ij}$ where $(i,j) \in \{(i,j)\colon \gamma_{ij} = 1, \gamma'_{ij} = 0\}$. Differentiating the above function for a particular value of u_{ij} gives,

$$\frac{\partial f(\mathbf{\Delta_\Gamma})}{\partial u_{\mathbf{\Gamma},ij}} = \mathrm{tr}\left[\left\{(\mathbf{\Delta_\Gamma} + \mathbf{\Omega}^*)^{-1}\boldsymbol{E}_{(i,j)} - \widehat{\mathbf{\Sigma}}\boldsymbol{E}_{(i,j)}\right\}\right]. \tag{4.54}$$

The maximizer $\widehat{u}_{\mathbf{\Gamma},ij}$ satisfies $\mathrm{tr}\left[\left\{(\mathbf{\Delta_\Gamma} + \mathbf{\Omega}^*)^{-1}\boldsymbol{E}_{(i,j)} - \widehat{\mathbf{\Sigma}}\boldsymbol{E}_{(i,j)}\right\}\right] = 0$. Now consider the function $g(\mathbf{\Delta_\Gamma})$ as defined earlier. The derivative of $g(\mathbf{\Delta_\Gamma})$ with respect to $u_{\mathbf{\Gamma},ij}$ satisfies

$$\left.\frac{\partial g(\mathbf{\Delta_\Gamma})}{\partial u_{\mathbf{\Gamma},ij}}\right|_{u_{\mathbf{\Gamma},ij}=0^+, u_{\mathbf{\Gamma},lm}=0, \forall (l,m)\neq(i,j)} \geq 0, \tag{4.55}$$

and,

$$\left.\frac{\partial g(\mathbf{\Delta_\Gamma})}{\partial u_{\mathbf{\Gamma},ij}}\right|_{u_{\mathbf{\Gamma},ij}=0^-, u_{\mathbf{\Gamma},lm}=0, \forall (l,m)\neq(i,j)} \leq 0. \tag{4.56}$$

The above two conditions give,

$$\begin{aligned}
\mathrm{tr}\left[\left\{(\mathbf{\Delta_\Gamma} + \mathbf{\Omega}^*)^{-1}\boldsymbol{E}_{(i,j)} - \widehat{\mathbf{\Sigma}}\boldsymbol{E}_{(i,j)}\right\}\right]\Big|_{u_{\mathbf{\Gamma},ij}=0^+, u_{\mathbf{\Gamma},lm}=0, \forall (l,m)\neq(i,j)=0} &\leq \frac{2\lambda}{n}, \\
\mathrm{tr}\left[\left\{(\mathbf{\Delta_\Gamma} + \mathbf{\Omega}^*)^{-1}\boldsymbol{E}_{(i,j)} - \widehat{\mathbf{\Sigma}}\boldsymbol{E}_{(i,j)}\right\}\right]\Big|_{u_{\mathbf{\Gamma},ij}=0^-, u_{\mathbf{\Gamma},lm}=0, \forall (l,m)\neq(i,j)=0} &\geq -\frac{2\lambda}{n}.
\end{aligned}$$

If the first derivative of $f(\mathbf{\Delta_\Gamma})$ is continuous at 0, then, we have $\widehat{u}_{\mathbf{\Gamma},ij} = 0$. This immediately implies the result stated in the lemma. □

We now prove the result on the bound of the remainder term in the Taylor series expansion of the function $h(\mathbf{\Omega})$.

Proof of Lemma 9. The Taylor series expansion of $h(\mathbf{\Omega})$ gives,

$$h(\mathbf{\Omega}) = h(\mathbf{\Omega}^*) + \frac{1}{2}\mathrm{vec}(\mathbf{\Delta})^T \boldsymbol{H}_{\mathbf{\Omega}^*}\mathrm{vec}(\mathbf{\Delta}) + R_n, \tag{4.57}$$

where R_n is the remainder term in the expansion. Using the integral form of the remainder, we have,

$$h(\boldsymbol{\Omega}) = h(\boldsymbol{\Omega}^*) + \text{vec}(\boldsymbol{\Delta})^T \left\{ \int_0^1 (1-\nu) \boldsymbol{H}_{\boldsymbol{\Omega}^* + \nu\boldsymbol{\Delta}} d\nu \right\} \text{vec}(\boldsymbol{\Delta}). \tag{4.58}$$

Subtracting (4.58) from (4.57) gives,

$$\begin{aligned}
R_n &= \text{vec}(\boldsymbol{\Delta})^T \left\{ \int_0^1 (1-\nu) \boldsymbol{H}_{\boldsymbol{\Omega}^* + \nu\boldsymbol{\Delta}} d\nu \right\} \text{vec}(\boldsymbol{\Delta}) - \frac{1}{2} \text{vec}(\boldsymbol{\Delta})^T \boldsymbol{H}_{\boldsymbol{\Omega}^*} \text{vec}(\boldsymbol{\Delta}) \\
&= \text{vec}(\boldsymbol{\Delta})^T \left\{ \int_0^1 (1-\nu) \left(\boldsymbol{H}_{\boldsymbol{\Omega}^* + \nu\boldsymbol{\Delta}} - \boldsymbol{H}_{\boldsymbol{\Omega}^*} \right) d\nu \right\} \text{vec}(\boldsymbol{\Delta}) \\
&\leq \|\boldsymbol{\Delta}\|_2^2 \left\| \int_0^1 (1-\nu) \left(\boldsymbol{H}_{\boldsymbol{\Omega}^* + \nu\boldsymbol{\Delta}} - \boldsymbol{H}_{\boldsymbol{\Omega}^*} \right) d\nu \right\|_{(2,2)} \\
&\leq \|\boldsymbol{\Delta}\|_2^2 \int_0^1 (1-\nu) \| \boldsymbol{H}_{\boldsymbol{\Omega}^* + \nu\boldsymbol{\Delta}} - \boldsymbol{H}_{\boldsymbol{\Omega}^*} \|_{(2,2)} d\nu \\
&\leq \frac{1}{2} \|\boldsymbol{\Delta}\|_2^2 \max_{0 \leq \nu \leq 1} \| \boldsymbol{H}_{\boldsymbol{\Omega}^* + \nu\boldsymbol{\Delta}} - \boldsymbol{H}_{\boldsymbol{\Omega}^*} \|_{(2,2)} \\
&\leq \frac{1}{2} \|\boldsymbol{\Delta}\|_2^2 (p+s) \max_{0 \leq \nu \leq 1} \| \boldsymbol{H}_{\boldsymbol{\Omega}^* + \nu\boldsymbol{\Delta}} - \boldsymbol{H}_{\boldsymbol{\Omega}^*} \|_{\infty}.
\end{aligned} \tag{4.59}$$

The above bound involves the maximum of the absolute differences between the elements of the Hessian matrices $\boldsymbol{H}$ computed at two different values $\boldsymbol{\Omega}^* + \nu\boldsymbol{\Delta}$ and $\boldsymbol{\Omega}^*$. We first show that, with probability tending to one,

$$\|(\boldsymbol{\Omega}^* + \nu\boldsymbol{\Delta})^{-1} - \boldsymbol{\Omega}^{*-1}\|_{\infty} \leq K \|\boldsymbol{\Delta}\|_2. \tag{4.60}$$

Using the matrix norm relations in Lemma 1, we get,

$$\begin{aligned}
\|(\boldsymbol{\Omega}^* + \nu\boldsymbol{\Delta})^{-1} - \boldsymbol{\Omega}^{*-1}\|_{(2,2)} &= \| \left((\boldsymbol{I} + \nu\boldsymbol{\Omega}^{*-1}\boldsymbol{\Delta})^{-1} - \boldsymbol{I}\right) \boldsymbol{\Omega}^{*-1}\|_{(2,2)} \\
&= \|\nu\boldsymbol{\Omega}^{*-1}\boldsymbol{\Delta}(\boldsymbol{I} + \nu\boldsymbol{\Omega}^{*-1}\boldsymbol{\Delta})^{-1}\boldsymbol{\Omega}^{*-1}\|_{(2,2)} \\
&\leq \nu\|\boldsymbol{\Omega}^{*-1}\|^2_{(2,2)}\|\boldsymbol{\Delta}\|_{(2,2)} \\
&\leq \|\boldsymbol{\Omega}^{*-1}\|^2_{(2,2)}\|\boldsymbol{\Delta}\|_2 \leq K\|\boldsymbol{\Delta}\|_2, \qquad (4.61)
\end{aligned}$$

with probability tending to 1, using (4.3). Thus, noting that $\|(\boldsymbol{\Omega}^* + \nu\boldsymbol{\Delta})^{-1} - \boldsymbol{\Omega}^{*-1}\|_\infty \leq \|(\boldsymbol{\Omega}^* + \nu\boldsymbol{\Delta})^{-1} - \boldsymbol{\Omega}^{*-1}\|_{(2,2)}$, we prove the result in equation (4.60).

For any symmetric matrix $\boldsymbol{A}$ of order d, we note that $\mathrm{tr}\left\{\boldsymbol{A}\boldsymbol{E}_{(i,j)}\boldsymbol{A}\boldsymbol{E}_{(l,m)}\right\}$ has the form $a_1a_2 + a_3a_4 + a_5a_6 + a_7a_8$ where $i, j, l, m \in \{1, \ldots, d\}$, and a_js are some elements of $\boldsymbol{A}$. This can be derived easily by writing out the elements of the product of the matrices involved and noting that matrices like $\boldsymbol{E}_{(i,j)}$ have non-zero entries at only two places corresponding to (i, j). Hence the elements of $\boldsymbol{H}_{\boldsymbol{\Omega}^*+\nu\boldsymbol{\Delta}} - \boldsymbol{H}_{\boldsymbol{\Omega}^*}$ have the form $(a_1a_2 + a_3a_4 + a_5a_6 + a_7a_8) - (b_1b_2 + b_3b_4 + b_5b_6 + b_7b_8)$, where a_j's and b_j's are some elements of $(\boldsymbol{\Omega}^* + \nu\boldsymbol{\Delta})^{-1}$ and $\boldsymbol{\Omega}^{*-1}$ respectively. Then, using equation (4.60), we get, with probability tending to one,

$$\sum a_1a_2 - \sum b_1b_2 \leq C_1\|\boldsymbol{\Delta}\|_2\|\boldsymbol{\Omega}^{*-1}\|_\infty + C_2\|\boldsymbol{\Delta}\|_2^2. \qquad (4.62)$$

Since this holds true for any arbitrary element of $\boldsymbol{H}_{\boldsymbol{\Omega}^*+\nu\boldsymbol{\Delta}} - \boldsymbol{H}_{\boldsymbol{\Omega}^*}$, using (4.3) and (4.62), we get that with probability tending to one,

$$\|\boldsymbol{H}_{\boldsymbol{\Omega}^*+\nu\boldsymbol{\Delta}} - \boldsymbol{H}_{\boldsymbol{\Omega}^*}\|_\infty \leq C_1\|\boldsymbol{\Delta}\|_2 + C_2\|\boldsymbol{\Delta}\|_2^2, \qquad (4.63)$$

where C_1 and C_2 are suitable constants.

Using (4.59) and (4.63), with probability tending to one, we have,

$$R_n \leq \frac{1}{2}(p+s)\|\boldsymbol{\Delta}\|_2^2 \left(C_1\|\boldsymbol{\Delta}\|_2 + C_2\|\boldsymbol{\Delta}\|_2^2\right).$$

□

We now prove the result on the error in Laplace approximation of the posterior probabilities of graphical model structures.

Proof of Theorem 4. Using the Taylor series expansion of $h(\boldsymbol{\Omega})$ as in (4.57), we can write the posterior probability of the graphical structure indicator $\boldsymbol{\Gamma}$ given the data $\boldsymbol{X}^{(n)}$ as in equation (4.15) to be proportional to

$$\int_{\boldsymbol{\Delta}+\boldsymbol{\Omega}^* \in \mathcal{M}^+} \exp\left\{-\frac{n}{2}\left(h(\boldsymbol{\Omega}^*) + \frac{1}{2}\mathrm{vec}(\boldsymbol{\Delta})^T \boldsymbol{H}_{\boldsymbol{\Omega}^*}\mathrm{vec}(\boldsymbol{\Delta}) + R_n\right)\right\} \prod_{(i,j)\in\mathcal{V}_{\boldsymbol{\Gamma}}} du_{ij}. \tag{4.64}$$

We denote $\prod_{(i,j)\in\mathcal{V}_{\boldsymbol{\Gamma}}} du_{ij}$ by $d\boldsymbol{\Delta}$ for notational simplicity. Using (4.14), we get

$$\frac{\int_{\|\boldsymbol{\Delta}\|_2\leq\epsilon_n} \exp\left[-\frac{n}{2}\left\{h(\boldsymbol{\Omega}^*) + \frac{1}{2}\mathrm{vec}(\boldsymbol{\Delta})^T \boldsymbol{H}_{\boldsymbol{\Omega}^*}\mathrm{vec}(\boldsymbol{\Delta}) + R_n\right\}\right] d\boldsymbol{\Delta}}{\int_{\boldsymbol{\Delta}+\boldsymbol{\Omega}^* \in \mathcal{M}^+} \exp\left[-\frac{n}{2}\left\{h(\boldsymbol{\Omega}^*) + \frac{1}{2}\mathrm{vec}(\boldsymbol{\Delta})^T \boldsymbol{H}_{\boldsymbol{\Omega}^*}\mathrm{vec}(\boldsymbol{\Delta}) + R_n\right\}\right] d\boldsymbol{\Delta}} \to 1. \tag{4.65}$$

Also, for $\|\boldsymbol{\Delta}\|_2 \leq \epsilon_n$, $R_n \leq (p+s)\|\boldsymbol{\Delta}\|_2^2\epsilon_n/2$. Thus, the upper and lower bounds of the integral $\int_{\|\boldsymbol{\Delta}\|_2\leq\epsilon_n} \exp\left\{-\frac{n}{2}h(\boldsymbol{\Omega})\right\} d\boldsymbol{\Delta}$ are given by

$$\begin{aligned} & e^{-nh(\boldsymbol{\Omega}^*)/2} \int_{\|\boldsymbol{\Delta}\|_2\leq\epsilon_n} \exp\left\{-\frac{n}{2}\left(\frac{1}{2}\mathrm{vec}(\boldsymbol{\Delta})^T \boldsymbol{H}_{\boldsymbol{\Omega}^*}\mathrm{vec}(\boldsymbol{\Delta}) \mp \frac{1}{2}(p+s)\epsilon_n\|\boldsymbol{\Delta}\|_2^2\right)\right\} d\boldsymbol{\Delta}. \\ = \;& e^{-nh(\boldsymbol{\Omega}^*)/2} \int_{\|\boldsymbol{\Delta}\|_2\leq\epsilon_n} \exp\left[-\frac{n}{4}\mathrm{vec}(\boldsymbol{\Delta})^T \left\{\boldsymbol{H}_{\boldsymbol{\Omega}^*} \mp (p+s)\epsilon_n \boldsymbol{I}\right\}\mathrm{vec}(\boldsymbol{\Delta})\right] d\boldsymbol{\Delta}. \end{aligned} \tag{4.66}$$

Note that,

$$\int_{\|\boldsymbol{\Delta}\|_2>\epsilon_n} \exp\left[-\frac{n}{4}\text{vec}(\boldsymbol{\Delta})^T\left\{\boldsymbol{H}_{\boldsymbol{\Omega}^*} \mp (p+s)\epsilon_n \boldsymbol{I}\right\}\text{vec}(\boldsymbol{\Delta})\right] d\boldsymbol{\Delta} \to 0, \tag{4.67}$$

if $(p+s)\epsilon_n \to 0$ and the minimum eigenvalue of $\boldsymbol{H}_{\boldsymbol{\Omega}^*}$ is bounded away from zero, which we prove in Lemma 12 below. Hence, the bounds can be simplified to

$$e^{-nh(\boldsymbol{\Omega}^*)/2}\int_{\boldsymbol{\Delta}+\boldsymbol{\Omega}^*\in\mathcal{M}^+} \exp\left[-\frac{n}{4}\text{vec}(\boldsymbol{\Delta})^T\left\{\boldsymbol{H}_{\boldsymbol{\Omega}^*} \mp (p+s)\epsilon_n \boldsymbol{I}\right\}\text{vec}(\boldsymbol{\Delta})\right] d\boldsymbol{\Delta}. \tag{4.68}$$

Using the above bounds, the ratio of the actual integral to the approximate integral has upper and lower bounds given by

$$\begin{aligned} & \frac{\int_{\boldsymbol{\Delta}+\boldsymbol{\Omega}^*\in\mathcal{M}^+} \exp\left[-\frac{n}{4}\text{vec}(\boldsymbol{\Delta})^T\left\{\boldsymbol{H}_{\boldsymbol{\Omega}^*} \mp (p+s)\epsilon_n \boldsymbol{I}\right\}\text{vec}(\boldsymbol{\Delta})\right] d\boldsymbol{\Delta}}{\int_{\boldsymbol{\Delta}+\boldsymbol{\Omega}^*\in\mathcal{M}^+} \exp\left\{-\frac{n}{4}\text{vec}(\boldsymbol{\Delta})^T\boldsymbol{H}_{\boldsymbol{\Omega}^*}\text{vec}(\boldsymbol{\Delta})\right\} d\boldsymbol{\Delta}} \\ = & \left[\frac{\det\left\{\boldsymbol{H}_{\boldsymbol{\Omega}^*} \pm (p+s)\epsilon_n \boldsymbol{I}_p\right\}}{\det(\boldsymbol{H}_{\boldsymbol{\Omega}^*})}\right]^{-1/2}. \end{aligned} \tag{4.69}$$

The above expression is bounded between $[1 \mp \{\text{eig}_1(\boldsymbol{H}_{\boldsymbol{\Omega}^*})\}^{-1}(p+s)\epsilon_n]^{-(p+s)/2}$. Lemma 12 below gives, $\text{eig}_1(\boldsymbol{H}_{\boldsymbol{\Omega}^*}) \gg 0$, and hence the above bound on the ratio goes to 1 if $(p+s)^2\epsilon_n \to 0$, so that the error in Laplace approximation is asymptotically small.

□

We now prove the result that the eigenvalues of the Hessian $\boldsymbol{H}_{\boldsymbol{\Omega}^*}$ are bounded away from zero.

Lemma 12. *Given a graphical model with model indicator $\boldsymbol{\Gamma}$, the minimum eigenvalue of the Hessian $\boldsymbol{H}_{\boldsymbol{\Omega}^*}$ corresponding to the function $h(\boldsymbol{\Omega})$, evaluated at $\boldsymbol{\Omega}^*$, is bounded away from zero.*

Proof. The Hessian of the function $h(\boldsymbol{\Omega})$ corresponding to the full model with $p + \binom{p}{2}$ free

elements has the form $\boldsymbol{H}_{\boldsymbol{\Omega},\text{full}} = \boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}^{-1}$. The Hessian $\boldsymbol{H}_{\boldsymbol{\Omega}^*}$ evaluated at the graphical lasso solution $\boldsymbol{\Omega}^*$ corresponding to the graphical model with model indicator $\boldsymbol{\Gamma}$ is a principal minor of $\boldsymbol{H}_{\boldsymbol{\Omega}^*,\text{full}}$. Hence it suffices to prove that the minimum eigenvalue of $(\boldsymbol{\Omega}^*)^{-1} \otimes (\boldsymbol{\Omega}^*)^{-1}$ is bounded away from zero. Note that, $\text{eig}_1\{(\boldsymbol{\Omega}^*)^{-1} \otimes (\boldsymbol{\Omega}^*)^{-1}\} = [\text{eig}_1\{(\boldsymbol{\Omega}^*)^{-1}\}]^2$. Thus, using (4.3), $[\text{eig}_1\{(\boldsymbol{\Omega}^*)^{-1}\}]^2 = 1/\|\boldsymbol{\Omega}^*\|_2^2 > 0$. □

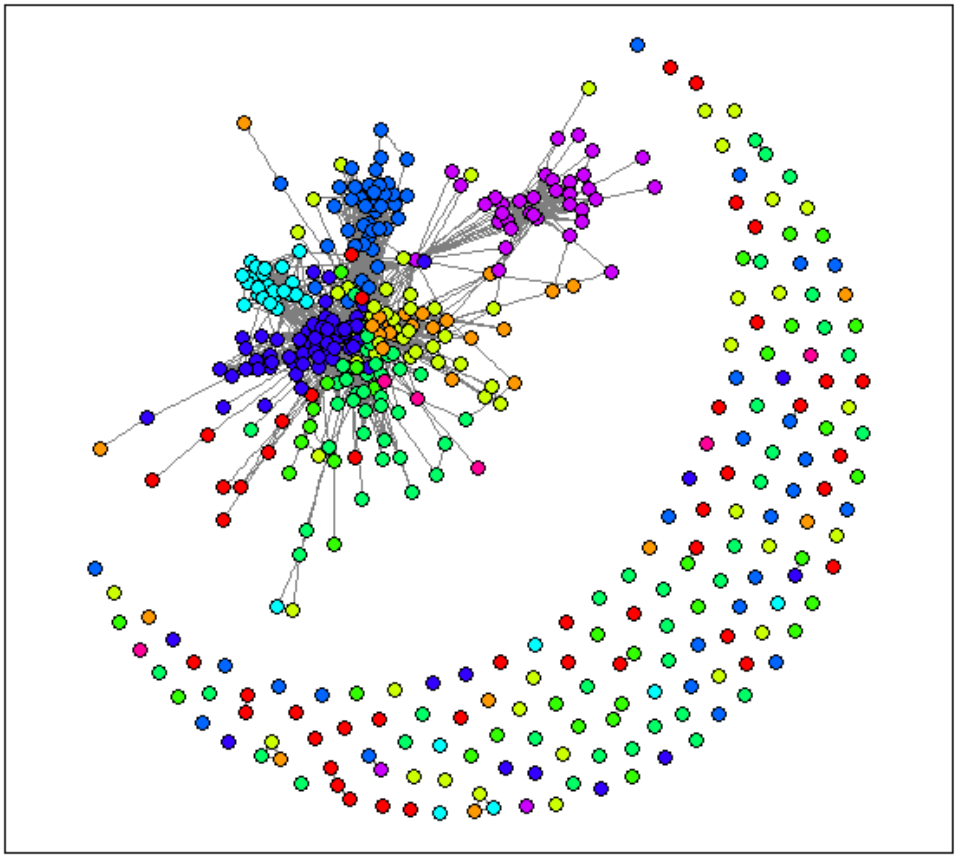


Figure 4.1: Graphical structure of the median probability model selected by the Bayesian graphical structure learning method.

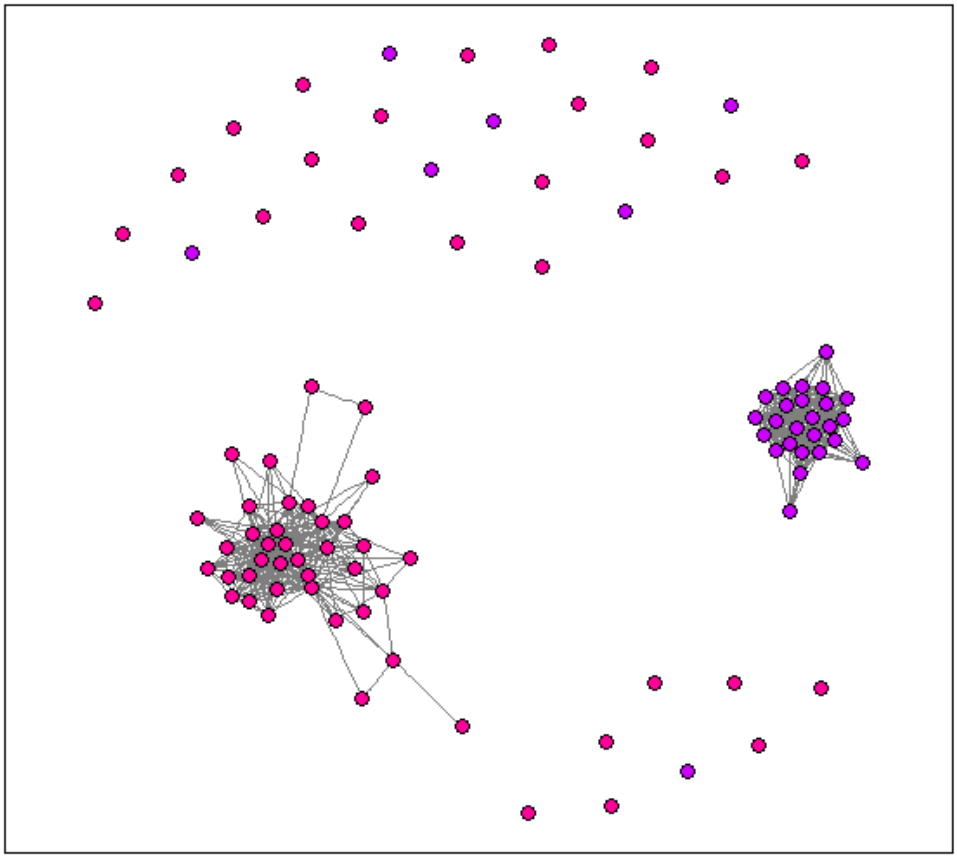


Figure 4.2: Graphical structure corresponding to the subgraph corresponding to the sectors “Utilities” [red] and “Information Technology”[violet].

REFERENCES

Atay-Kayis, A. and Massam, H. (2005). A Monte-Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models. *Biometrika*, 92(2):317–335.

Avalos, M., Grandvalet, Y., and Ambroise, C. (2003). Regularization methods for additive models. *Advances in Intelligent Data Analysis*, V:509–520.

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.*, 9:485–516.

Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. *Ann. Statist.*, 32(3):870–897.

Barron, A., Schervish, M. J., and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.*, 27(2):536–561.

Belitz, C. and Lang, S. (2008). Simultaneous selection of variables and smoothing parameters in structured additive regression models. *Comput. Statist. Data Anal.*, 53(1):61–81.

Bickel, P. and Levina, E. (2008a). Covariance regularization by thresholding. *Ann. Statist.*, 36(6):2577–2604.

Bickel, P. and Levina, E. (2008b). Regularized estimation of large covariance matrices. *Ann. Statist.*, 36(1):199–227.

Birgé, L. (1984). Sur un théorème de minimax et son application aux tests. *Probab. Math. Statist.*, 3:259–282.

Bondell, H. D. and Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics*, 64(1):115–123.

Boyd, S. P. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg. Methods, theory and applications.

Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models. *Ann. Statist.*, 17(2):453–555.

Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.*, 106(494):672–684.

Cai, T., Liu, W., and Luo, X. (2011). A constrained ℓ_1-minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.*, 106(494):594–607.

Cai, T. and Yuan, M. (2012). Adaptive covariance matrix estimation through block thresholding. *Ann. Statist.*, 40(4):2014–2042.

Cai, T., Zhang, C., and Zhou, H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.*, 38(4):2118–2144.

Carvalho, C., Massam, H., and West, M. (2007). Simulation of hyper-inverse Wishart distributions in graphical models. *Biometrika*, 94(3):647–659.

Carvalho, C. and Scott, J. (2009). Objective Bayesian model selection in Gaussian graphical models. *Biometrika*, 96(3):497–512.

Castillo, I. and van der Vaart, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *Ann. Statist.*, 40(4):2069–2101.

Chen, M.-H., Ibrahim, J. G., and Yiannoutsos, C. (1999). Prior elicitation, variable selection and bayesian computation for logistic regression models. *J. Roy. Statist. Soc. Ser. B*, 61(1):223–242.

Chen, Z. H. (1993). Fitting multivariate regression functions by interaction spline models. *J. Roy. Statist. Soc. Ser. B*, 55(2):473–491.

Curtis, S. M., Banerjee, S., and Ghosal, S. (2014). Fast Bayesian model assessment for nonparametric additive regression. *Comput. Statist. Data Anal.*, 71:347–358.

Dawid, A. and Lauritzen, S. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.*, 21(3):1272–1317.

Dellaportas, P. and Forster, J. J. (1999). Markov chain monte carlo model determination for hierarchical and graphical log-linear models. *Biometrika*, 86(3):615–633.

Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates. *Ann. Statist.*, 14(1):1–67. With a discussion and a rejoinder by the authors.

Dobra, A., Hans, C., Jones, B., Nevins, J., Yao, G., and West, M. (2004). Sparse graphical models for exploring gene expression data. *J. Multivariate Anal.*, 90(1):196–212.

Dobra, A., Lenkoski, A., and Rodriguez, A. (2011). Bayesian inference for general Gaussian graphical models with application to multivariate lattice data. *J. Amer. Statist. Assoc.*, 106(496):1418–1433.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360.

Faure, H., Preziosi, P., Roussel, A., Bertrais, S., Galan, P., Hercberg, S., and Favier, A. (2006). Factors influencing blood concentration of retinol, α-tocopherol, vitamin c, and β-carotene in the French participants of the SU. VI. MAX trial. *European Journal of Clinical Nutrition*, 60(6):706–717.

Freedman, D. A. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *Ann. Math. Statist.*, 34:1386–1403.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.*, 85(410):398–409.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741.

George, E. and McCulloch, R. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.*, 88:881–889.

George, E. I. (2000). The variable selection problem. *J. Amer. Statist. Assoc.*, 95(452):1304–1308.

George, E. I. and McCulloch, R. E. (1997). Approaches for bayesian variable selection. *Statistica Sinica*, 7(2):339–373.

Ghosal, S. (2000). Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity. *J. Multivariate Anal.*, 74(1):49–68.

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.*, 27(1):143–158.

Ghosal, S., Ghosh, J. K., and Van Der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15.

Gustafson, P. (2000). Bayesian Regression Modeling with Interactions and Smooth Effects. *J. Amer. Statist. Assoc.*, 95:795–806.

Harville, D. A. (2008). *Matrix Algebra from a Statistician's Perspective*. Springer.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*, volume 43 of *Monographs on Statistics and Applied Probability*. Chapman and Hall Ltd., London.

Huang, J., Horowitz, J. L., and Wei, F. (2010). Variable selection in nonparametric additive models. *Ann. Statist.*, 38(4):2282–2313.

Huang, J., Liu, N., Pourahmadi, M., and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98.

Hwang, W. Y., Zhang, H. H., and Ghosal, S. (2009). FIRST: combining forward iterative selection and shrinkage in high dimensional sparse linear regression. *Stat. Interface*, 2(3):341–348.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *Ann. Statist.*, 33(2):730–773.

James, L. F. (2008). Large sample asymptotics for the two-parameter Poisson-Dirichlet process. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, volume 3 of *Inst. Math. Stat. Collect.*, pages 187–199. Inst. Math. Statist., Beachwood, OH.

Karoui, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.*, 36(6):2717–2756.

Kim, Y. and Lee, J. (2001). On posterior consistency of survival models. *Ann. Statist.*, 29(3):666–686.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann Statist.*, 37(6B):4254.

Lauritzen, S. (1996). *Graphical Models*, volume 17. Oxford University Press, USA.

Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer, New York.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.*, 88(2):365–411.

Lenkoski, A. and Dobra, A. (2011). Computational aspects related to inference in Gaussian graphical models with the g-Wishart prior. *J. Comput. Graphical Statist.*, 20(1):140–157.

Letac, G. and Massam, H. (2007). Wishart distributions for decomposable graphs. *Ann. Statist.*, 35(3):1278–1323.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *J. Amer. Statist. Assoc.*, 103(481):410–423.

Lin, Y. and Zhang, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.*, 34(5):2272–2297.

Linton, O. and Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, 82(1):93–100.

Liu, H., Lafferty, J., and Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.*, 10:2295–2328.

Liu, X., Wang, L., and Liang, H. (2011). Estimation and variable selection for semiparametric additive partial linear models. *Statistica Sinica*, 21(3):1225.

Lokhorst, J., Venables, B., Turlach, B., and Maechler, M. (2013). lasso2: L1 constrained estimation aka lasso. R package version 1.2–18.

Marra, G. and Wood, S. N. (2011). Practical variable selection for generalized additive models. *Comput. Statist. Data Anal.*, 55(7):2372–2387.

Marx, B. D. and Eilers, P. H. (1998). Direct generalized additive modeling with penalized likelihood. *Comput. Statist. Data Anal.*, 28(2):193–209.

McCullagh, P. and Nelder, J. A. (1983). *Generalized linear models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London.

Meier, L., van de Geer, S., and Bühlmann, P. (2009). High-dimensional additive modeling. *Ann. Statist.*, 37(6B):3779–3821.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann Statist.*, 34(3):1436–1462.

Meyer, M. C. and Laud, P. W. (2002). Predictive variable selection in generalized linear models. *J. Amer. Statist. Assoc.*, 97(459):859–871.

Miller, A. (2002). *Subset selection in regression*, volume 95 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, Boca Raton, FL, Second edition.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.*, 83(404):1023–1036. With comments by James Berger and C. L. Mallows and with a reply by the authors.

Muirhead, R. (2005). *Aspects of Multivariate Statistical Theory*. Wiley, New York.

Nierenberg, D. W., Stukel, T. A., Baron, J. A., Dain, B. J., and Greenberg, E. R. (1989). Determinants of plasma levels of beta-carotene and retinol. *American Journal of Epidemiology*, 130(3):511–521.

Ntzoufras, I., Dellaportas, P., and Forster, J. J. (2003). Bayesian variable and link determination for generalised linear models. *J. Statist. Plann. Inference*, 111(1):165–180.

Park, T. and Casella, G. (2008). The Bayesian Lasso. *J. Amer. Statist. Assoc.*, 103(482):681–686.

Pati, D., Bhattacharya, A., Pillai, N., and Dunson, D. (2012). Posterior contraction in sparse Bayesian factor models for massive covariance matrices. *arXiv preprint arXiv:1206.3627*.

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, 83(2):251–266.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *J. Amer. Statist. Assoc.*, 92(437):179–191.

Raftery, A. E. and Richardson, S. (1993). Model selection for generalized linear models via GLIB, with application to epidemiology. *Bayesian Biostatistics. New York: Marcel Dekker*.

Rajaratnam, B., Massam, H., and Carvalho, C. (2008). Flexible covariance estimation in graphical Gaussian models. *Ann Statist.*, 36(6):2818–2849.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009). Sparse additive models. *J. Roy. Statist. Soc. Ser. B*, 71(5):1009–1030.

Reich, B. J., Storlie, C. B., and Bondell, H. D. (2009). Variable selection in Bayesian smoothing spline ANOVA models: application to deterministic computer codes. *Technometrics*, 51(2):110–120.

Rothman, A., Bickel, P., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Statist.*, 2:494–515.

Rothman, A., Levina, E., and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc.*, 104(485):177–186.

Roverato, A. (2000). Cholesky decomposition of a hyper inverse Wishart matrix. *Biometrika*, 87(1):99–112.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*, volume 12. Cambridge University Press.

Schwartz, L. (1965). On Bayes procedures. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 4:10–26.

Shi, P. and Tsai, C.-L. (1999). Semiparametric regression model selections. *J. Statist. Plann. Inference*, 77(1):119–139.

Shively, T. S., Kohn, R., and Wood, S. (1999). Variable selection and function estimation in additive nonparametric regression using a data-based prior. *J. Amer. Statist. Assoc.*, 94(447):777–806.

Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E., and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. ii. radical prostatectomy treated patients. *The Journal of Urology*, 141(5):1076–1083.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288.

Wang, H. (2012). Bayesian graphical lasso models and efficient posterior computation. *Bayesian Analysis*, 7(4):867–886.

Wang, L., Liu, X., Liang, H., and Carroll, R. J. (2011). Estimation and variable selection for generalized additive partial linear models. *Ann. Statist.*, 39(4):1827–1851.

Wang, X. and George, E. I. (2007). Adaptive bayesian criteria in variable selection for generalized linear models. *Statistica Sinica*, 17(2):667–690.

Witten, D. M., Friedman, J. H., and Simon, N. (2011). New insights and faster computations for the graphical lasso. *J. Comput. Graph. Statist.*, 20(4):892–900.

Wood, S., Kohn, R., Shively, T., and Jiang, W. (2002). Model selection in spline nonparametric regression. *J. Roy. Statist. Soc. Ser. B*, 64(1):119–139.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *J. Amer. Statist. Assoc.*, 99(467).

Wood, S. N. (2006). *Generalized additive models*. Texts in Statistical Science Series. Chapman & Hall/CRC, Boca Raton, FL.

Yang, R. and Berger, J. O. (1994). Estimation of a covariance matrix using the reference prior. *Ann. Statist.*, 22(3):1195–1211.

Yuan, M. and Lin, Y. (2005). Efficient empirical bayes variable selection and estimation in linear models. *J. Amer. Statist. Assoc.*, 100(472).

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. Ser. B*, 68(1):49–67.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, 6:233–243.

Zellner, A. and Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. *Trabajos de estadística y de investigación operativa*, 31(1):585–603.

Zhao, T., Liu, H., Roeder, K., Lafferty, J., and Wasserman, L. (2012). The huge package for High-dimensional Undirected Graph Estimation in R. *J. Mach. Learn. Res.*, 98888:1059–1062.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B*, 67(2):301–320.