Bayesian Methods for Graphical Models with Limited Data

Zehang Li

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Tyler H. McCormick, Chair

Samuel J. Clark

Johannes Lederer

Jon Wakefield

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Bayesian Methods for Graphical Models with Limited Data

Zehang Li

Chair of the Supervisory Committee:
Associate Professor Tyler H. McCormick
Department of Statistics and Department of Sociology

Scientific studies in many fields involve understanding and characterizing dependence relationships among large numbers of variables. This can be challenging in settings where data is limited and noisy. Take survey data as an example, understanding the associations between questions may help researchers better explain themes amongst related questions and impute missing values. Yet, such data typically contains a combination of binary, continuous, and categorical variables, high proportions of missing values, and complex data structures. In this dissertation, we develop flexible models and algorithms to estimate Gaussian and latent Gaussian graphical models from noisy data. First, we develop a latent Gaussian graphical model for mixed data that takes advantage of informative prior beliefs on the marginal distribution of variables. Next, we propose several shrinkage priors for precision matrices and develop estimation procedures for fast posterior explorations of a single and multiple graphical models. This work is motivated by modeling survey-based cause of death instruments, known as verbal autopsies (VAs). Our methods provide new perspectives in improving model performance while recovering useful dependencies in the VA data.

TABLE OF CONTENTS

Page

List of Figures iii

List of Tables vii

Glossary viii

Chapter 1: Introduction 1
1.1 An overview 1
1.2 Motivating Examples 3
1.3 Organization of the dissertation 4

Chapter 2: Background 6
2.1 Gaussian graphical models 6
2.2 Verbal autopsy 8

Chapter 3: Latent Gaussian graphical models with informative marginal priors . . 12
3.1 Introduction 12
3.2 Latent Gaussian graphical model for mixed data 16
3.3 Posterior inference 23
3.4 Cause-of-death assignment using latent Gaussian mixture model 27
3.5 Simulation evidence 30
3.6 Analysis of verbal autopsy data 34
3.7 Discussion 41

Chapter 4: Gaussian graphical models using ECM algorithm 45
4.1 Introduction 45
4.2 Spike-and-slab prior for Gaussian graphical models 46

4.3 ECM algorithm for graph selection 50
4.4 ECM algorithm for copula graphical models 53
4.5 Informative priors 56
4.6 Posterior summary of the ECM output 58
4.7 Simulation 59
4.8 Analysis of sales data 60
4.9 Discussion 61

Chapter 5: Joint estimation of multiple Gaussian graphical models 65
5.1 Introduction 65
5.2 Preliminaries 66
5.3 Bayesian joint graphical lasso priors 68
5.4 Bayesian joint spike-and-slab graphical lasso priors 70
5.5 Model estimation 73
5.6 Dynamic posterior exploration 76
5.7 Numerical results 79
5.8 Discussion 83

Chapter 6: Discussion and Future Work 85

Appendix A: Appendix for Chapter 3 102
A.1 Derivation of the spike-and-slab prior 102
A.2 Implied prior sparsity with different hyperparameters 103
A.3 Posterior sampling for the classification model 105

Appendix B: Appendix for Chapter 5 107
B.1 Gibbs sampler of the proposed models 110
B.2 Additional simulation evidence 112
B.3 Details on the verbal autopsy data analysis 116
B.4 Details on prediction of missing mortality rates 117

LIST OF FIGURES

Figure Number Page

3.1 Implied prior edge probability with $\lambda = 10$ for $p = 100$ graph. The dots represent the median prior probabilities and the error bars represent the 0.025 and 0.975 quantiles. The densities are derived from sampling $1,000$ draws using MCMC from the prior distribution after $1,000$ iterations of burn-in. . . 24

3.2 Classification and CSMF accuracy for mixed data. Average classification accuracy and CSMF accuracy for different methods with correct and misspecified priors and different proportion of missing data for *mixed data*. Top row: CSMF accuracy. Bottom row: Accuracy of individual most likely class assignment. The accuracy is evaluated in a dataset with a total $n = 800$ observations and $p = 50$ variables including 5 continuous variables from $C = 20$ classes, with or without additional labeled data. 35

3.3 Classification and CSMF accuracy for PHMRC cross-validation study. The metrics are evaluated on $1,000$ randomly selected deaths for InterVA, Naive Bayes classifier, and the proposed model without any training data (GM: w/o training). An additional $1,000$ randomly selected labeled death is used as training data in the last case (GM: w/i training). The labeled data are not assumed to have the same distribution of causes. 37

3.4 Distributions of causes-of-death in Karonga dataset by year. The integers in each cell show the number of deaths in the corresponding period, and the shading represents the proportion of causes in each year. The data before 2008 are used as prior information in the experiment and thus are combined in this figure. 38

3.5 Scatter plot of the estimated CSMF against true CSMF for Karonga data from 2008 to 2014 using different methods. Causes with true fractions larger than 0.05 are labeled in the plot. The vertical bars correspond to the 95% posterior credible intervals estimated from the proposed model. The proposed Gaussian mixture model shows smaller bias. 40

3.6 Posterior mean correlation (left), inverse correlation (middle), and the inclusion probability (right) matrix for Karonga data. The cells with orange color are the known edges from the questionnaire structure that is not estimated. . . . 41

3.7 Classification accuracy and CSMF accuracy for Karonga physician coded data through cross-validation. The marginal probabilities are calculated with data from 2002 to 2007. The training data and testing data are randomly sampled from the rest of the data from 2008 to 2014. The proposed method consistently outperform Naive Bayes classifier using either prior conditional probabilities or conditional probabilities derived from training data. 44

4.1 Different marginal priors on $\boldsymbol{R}$ induced by the spike-and-slab prior on $\boldsymbol{\Omega}$ with $p = 50$ and $\lambda = 2$. A complete graph, i.e. $v_0 = v_1$ is assumed. The densities are derived from sampling $2,000$ draws using MCMC from the prior distribution after $2,000$ iterations of burn-in. 48

4.2 Comparison of specified marginal prior distribution and induced marginal prior distributions for $\boldsymbol{\Omega}$ with $p = 50$, $\lambda = 2$, $v_1 = 1$ and varying v_0 values. The underlying graph is fixed to be an AR(2) graph. Left: diagonal elements $\boldsymbol{\Omega}_{ii}$. Right: Non-zero off-diagonal elements (slab) $\boldsymbol{\Omega}_{ij}, i \neq j$. The densities are derived from sampling $2,000$ draws using MCMC from the prior distribution after $2,000$ iterations of burn-in. 49

4.3 Comparing graph selection path using EMGS and graphical lasso, on a 10-node graph. The red dashed line at 0.5 is the true value for the non-zero partial correlations. The blue solid lines represent the non-zero off-diagonal elements and the green dashed lines represent the zero off-diagonal elements. 54

4.4 Comparing the estimated and true precision matrix using with exchangeable prior and structured prior for block-wise rescaling. In each plot of the precision matrix comparison, the upper triangle shows the estimated matrix and the lower triangle shows the true precision matrix. All the tunning parameters or selected so that the estimated graph has the closest number of edges compared to the true graph. The last line plot shows the change of $\hat{\tau}_{gg'}$ over different choices of v_0. The blocks are labeled 1 to 3 from top left to bottom right. . . 59

4.5 Comparing the estimated precision matrix from cross validation. The blocks correspond to product categories. From upper left to lower right: pretzels, cold cereal, frozen pizza, and mouthwash. Left column: estimated standardized precision matrix. Middle column: estimated standardized precision matrix with highlighted graph selection. Edges with less than 0.5 probability of being from the slab distributions in the first two plots, and exact zeros in graphical lasso output are marked with gray color. Right column: average negative log likelihood on the validation sets. 62

5.1 The solution paths and estimated precision matrices of FGL (upper row), SS-FGL (middle row) and DSS-FGL (lower row). The red nodes correspond to true edges and the gray nodes correspond to 0's. The two vertical lines in the FGL solution path indicate the model that best matches the true sparsity (left) and the model with the lowest AIC (right). The block containing the edges is plotted for the estimated values (upper triangular) against the truth (lower triangular). The model that best matches the true graphs is plotted for FGL. The off-diagonal values are rescaled and negated to partial correlations, and 0's are colored with light gray background for easier visual comparison. The bias of the estimated precision matrix measured by the Frobenius norm, $||\hat{\mathbf{\Omega}}_g - \mathbf{\Omega}_g||_F$, is also printed in the captions. 77

5.2 Performance of FGL, GGL, DSS-FGL, and DSS-GGL over 100 replications. The dots represent the metrics for the 100 selected models under DSS-FGL and DSS-GGL, and the lines represent the average performance of FGL and GGL over 100 replications under different tuning parameters. 80

5.3 Estimated edges between the symptoms under the three causes using DSS-FGL (top row) and DSS-GGL (bottom row). The width of the edges are proportional to the size of $|\omega_{jk}^{(g)}|$. Common edges across all groups are colored in blue, and the differential edges are colored in red. 82

A.1 Implied prior edge probability for $p = 100$ graph. The dots represent the median prior probabilities and the error bars represent the 0.025 and 0.975 quantiles The rows in the panel represent the value of v_0, and the columns represent the choice of v_1/v_0. For each combination of v_0 and v_1, the edge probabilities induced by different λ and $\pi_{\boldsymbol{\delta}}$ are plotted. The densities are derived from sampling $1,000$ draws using MCMC from the prior distribution after $1,000$ iterations of burn-in. 104

B.1 Graph structure of the simulated dataset. The edges between the red nodes are removed from the second class, and edges between both the red and blue nodes are removed from the third class. 113

B.2 Performance of FGL, GGL, DSS-FGL, and DSS-GGL over 100 replications, $p = 100$. The dots represent the metrics for the 100 selected models under DSS-FGL and DSS-GGL, and the lines represent the average performance of FGL and GGL over 100 replications under different tuning parameters. . . . 114

B.3 Performance of FGL, GGL, DSS-FGL, and DSS-GGL over 100 replications, $p = 200$. The dots represent the metrics for the 100 selected models under DSS-FGL and DSS-GGL, and the lines represent the average performance of FGL and GGL over 100 replications under different tuning parameters. . . . 114

B.4 The density plot of true positive edges against false positive edges for DSS-FGL (top row), and DSS-GGL (bottom row) under different choices of λ_2. λ_1 is set to 1. 115

B.5 The density plot of true positive differential edges against false positive positive edges for DSS-FGL (top row), and DSS-GGL (bottom row) under different choices of λ_2. λ_1 is set to 1. 115

B.6 Estimated edges between the symptoms under the three causes using FGL using AIC (first row), GGL using AIC (second row), FGL with the same number of edges as selected by DSS-FGL (third row), and GGL with the same number of edges as selected by DSS-GGL (last row). The width of the edges are proportional to the size of $|\omega_{jk}^{(g)}|$. Common edges across all groups are colored in blue, and the differential edges are colored in red. 118

B.7 Estimated partial correlation matrix using one cross-validation dataset. The partial correlations among the 101 age groups are estimated using FGL with the same number of edges as selected by DSS-FGL (top row), and DSS-FGL (bottom row). DSS-FGL estimates 197 and 199 edges respectively for female and male. The closet configuration of FGL estimates 157 and 241 edges respectively. The precision matrices are rescaled and negated to partial correlations for easier interpretation. 119

LIST OF TABLES

Table Number Page

3.1 Simulation with binary and mixed $\boldsymbol{X}$ under different scenarios. The proposed latent Gaussian graphical model approach (Spike-and-Slab prior) outperforms the semi-parametric alternatives and the marginal uniform prior (Uniform prior) in both scenarios. The Spike-and-slab prior performs especially well in scenarios with a high proportion of missing data. 33

3.2 CSMF accuracy, Top 1 to 3 cause assignment accuracy for Karonga physician coded data. The marginal probabilities are calculated with data from 2002 to 2007. The training data consist of all the data from 2002 to 2007. The testing data are the rest of the data from 2008 to 2014. The proposed Gaussian mixture model consistently outperform Naive Bayes classifier and InterVA. . 40

3.3 List of conditional dependent symptom pairs. The non-zero elements in the inverse correlation matrix are selected by the estimated median probability graph. 42

4.1 Comparing estimation of the standardized precision matrix for Gaussian graphical model and copula graphical model with mixed variables. The final graphs are chosen so that the sparsity level is closest to the truth. 64

5.1 Average and standard deviation of the mean squared errors from 50 cross-validation experiments. The FGL model is selected to have the same number of edges as the DSS-FGL. 83

B.1 List of symptoms considered in this analysis. 116

GLOSSARY

VA: Verbal autopsy.

CSMF: Cause-specific morality fraction

COD: Cause of death

HDSS: Health & demographic surveillance system

PHMRC: Population Health Medical Research Consortium

EMGS: EM graph selection

JGL: Joint graphical lasso

GGL: Group graphical lasso

FGL: Fused graphical lasso

SS-GGL: Spike-and-slab group graphical lasso

SS-FGL: Spike-and-slab fused graphical lasso

DSS-GGL: Doubly spike-and-slab group graphical lasso

DSS-FGL: Doubly spike-and-slab fused graphical lasso

ACKNOWLEDGMENTS

I am extraordinarily fortunate to have had several great mentors throughout my graduate career. First and foremost, I would like to thank my advisor, Tyler McCorcmick, for his patience and encouragement throughout the last few years. The work in this dissertation and beyond would not have been possible without the countless inspiring discussions with Tyler or the motivation and joy he helped me identify in research.

I also owe a debt of gratitude to my committee members for their contribution to both my research and professional development. I attribute my interest in demography largely to the collaborations with Sam Clark, who has also been an amazing source of wisdom and inspiration in my career. I was fortunate to take my first course in Bayesian statistics from Jon Wakefield. An incredible amount of my knowledge in Bayesian and spatial statistics comes from working alongside Jon and is imprinted with stories from his huge repertoire of anecdotes. I am also grateful for the spirited discussions on graphical models with Johannes Lederer, which always provide new ideas for my research.

From the Department of Statistics, I wish to thank Adrian Dobra, Daniela Witten, Elena Erosheva, Marina Meilă, Peter Hoff, Mathias Drton, Thomas Richardson, and Paul Sampson for their profound influence on my perspective of statistics. I would also like to thank, in particular, Daniela Witten and Caren Marzban, for showing me the passion of teaching and making me a better educator.

I have been blessed with the opportunity to work with many wonderful collaborators over the years. My thanks go to Clara Calvert, Clarissa Surek-Clark, Tsuyoshi Kunihama, Jason

Thomas, Martin Bratschi, Basia Zaba, and many more colleagues working at Data for Health Initiative, WHO, and various countries, for pushing me to better understand the field of verbal autopsy and for the many memorable meetings all over the world. I wish to thank Laina Mercer, Yuan Hsiao, Jessica Godwin, Bryan Martin, Geir-Arne Fuglstad, and Andrea Riebler for sharing their time and enthusiasm in small-area estimation. I also appreciate the mentorship from Matt Goldman and Matt Taddy during my summer at Microsoft Research. Finally I would like to thank Chun Yip Yau for introducing to me the whole new world of statistics during my undergraduate years.

There are also many other individuals without whom I cannot imagine any of my achievements being possible. I would like to thank Ellen Reynolds, Eileen Heimer, and Mee-Ling Hon for making all the administrative procedures easy and worry-free; and Asa Sourdiffe for tolerating and maintaining the heavy computational burden I brought to the computer clusters. I am also grateful to my fellow students from the Department of Statistics at University of Washington and the Chinese University of Hong Kong for walking the journey with me while making it so much fun: Rebecca Ferrell, Ted Westling, Maryclare Griffin, Bowen Wang, Qiyang Han, Mingwei Tang, Peiran Liu, Johnny Paige, Yali Wan, Yanjun He, Xiaohan Yan, Hui Yu, Qi Gao, Xiangnan Feng, Zijian Guo, Zifeng Zhao, and many more. I am blessed with the many fond memories from our friendship.

Finally, my thanks go to my parents for always encouraging me to pursue my dreams and explore the world; and above all, I could not have arrived here without the love, support, and food from Nanxun. I am extremely grateful for all our great adventures.

DEDICATION

To my patents, and to Nanxun

Chapter 1
INTRODUCTION

1.1 An overview

Modeling dependence structures among multivariate data is a challenging task in a variety of scientific domains, especially as data are collected with increasing complexity. Over the last several decades, an extensive collection of methodological research centered around *graphical models* have been studied and successfully implemented in many fields including biological sciences, economics, disease modeling, epidemiology, etc. The key assumption behind the graphical model framework is that the true dependence relationships among variables are sparse, and thus the joint distribution of many variables can be characterized by a small number of parameters. This is especially useful in high dimensional settings, where maximum likelihood estimators can be too noisy and difficult to interpret.

In fields such as social science and demography, however, although high dimensional data are more and more common, applying standard graphical model techniques are usually challenging due to many practical constraints. To name a few,

1. Data quality: Many problems from the social science domains need to deal with data with high proportion of missing values and limited sample sizes. Thus it is usually impractical to estimate models based on only the complete observations.

2. Mixed data types: Non-Gaussian data and data containing a mixture of discrete and continuous variables are very common in many studies, while models that deal with high-dimensional mixed data is still limited.

3. Informative prior information: Domain knowledge and expert opinions typically exist in many forms, such as the prior beliefs about network structure, or the marginal distribution of variables, etc. It is often nontrivial to efficiently incorporate such prior information to improve inference for dependence structures.

4. Heterogeneity: Many observational studies collect data from heterogeneous classes or groups. For example, survey data are usually conducted over time, different locations, or from distinct subpopulations. Ignoring such heterogeneity in the data can lead to erroneous inference of the dependence structures, while inference based on each subset of data separately may be computationally infeasible due to the low data quality.

5. Hierarchical models: While uncovering dependence relationships is useful in understanding the data, rarely do research questions stop at identifying conditional dependent pairs of random variables. Association models usually need to be embedded in hierarchical, multilevel framework, in order to provide realistic descriptions of the data generating mechanism, and to be useful for answering research questions through tasks such as prediction and classification.

To address these challenges, we need both new models to characterize associations among variables under various constraints, and novel algorithms for efficient computation. This dissertation focuses on both the modeling and the computational aspects and describes different approaches to deal with these challenges in practice. For the rest of this chapter, we will include a brief introduction to three motivating examples and describe the organization of the rest of the dissertation.

1.2 Motivating Examples

1.2.1 Symptom dependencies in verbal autopsy analysis

In many regions without complete coverage of civil registration and vital statistics systems, verbal autopsies (VAs) are widely used to provide cause-specific mortality estimates. VAs are a standardized questionnaire administered to caregivers or family members of a recently deceased person. Since all of these deaths happen outside of hospitals, physical autopsy is usually not available. Assigning causes of death and estimating the population fraction of death by causes from the collected surveys has been a challenging task. The lack of consistent gold-standard training data rules out most commonly used machine learning approaches, and make cause-of-death assignment rely heavily on the informative prior knowledge obtained from physicians, as well as the assumption that symptoms are conditional independent given the cause-of-death. However, this assumption is always violated in practice. None of the existing methods successfully characterize the joint distribution of symptoms on the VA questionnaires, due to the high dimensionality and their mixed types. More details of VA and existing models is discussed in Chapter 2. In Chapter 3, we will propose a new framework for modeling symptom dependencies in VA data. In Chapter 5, we will further discuss extensions that estimate cause-specific dependence structures among symptoms.

1.2.2 Sales of many competing products

Economists often model consumer choice between a set of alternatives via random utility models. The essential assumption for the generic demand model lies in the condition of independence of irrelevant alternatives (IIA). This assumption, however, is very strong and is often left untested in economic models of discrete choice. In some cases, IIA is unlikely to hold across the entire universe of competing products. Thus one alternative is to divide the

choice problem into a series of nested decisions, which is often based primarily on subjective beliefs. Therefore it is desirable to develop data driven approaches to search for cross-product competition automatically from the time series of aggregated sales. In Chapter 4, we will analyze one such dataset of sales that is publicly available.

1.2.3 Age-specific mortality rates

Age-specific mortality schedules are one of the most fundamental indicators in many formal demography models. However, in many low income countries without full civil registration and vital statistics system, it is common to have missing values in age-specific death rates. Many models have been proposed to impute missing death rates as a function of age, and they are typically estimated for each country, year, and sex combination separately, assuming independent errors after removing the mean model. This independence assumption of residuals, however, is unlikely to hold in practice. For example, consider one-year age groups from 0 to 100, residuals from one age group are usually correlated with residuals with adjacent age groups. Ignoring such correlations can lead to suboptimal inference or prediction of the missing mortality rates, while on the other hand, estimating an unconstrained covariance matrix among all 101 age groups for each country and sex with a small number of data can be unstable. In Chapter 5, we will describe a model to estimate multiple sparse inverse covariance matrices among the residuals of the 101 age groups for multiple groups of mortality rate series, and show its performance in predicting missing values.

1.3 Organization of the dissertation

In this dissertation, we aim to provide several modeling and computational strategies to estimating graphical models under the Bayesian framework. Chapter 2 provides a review of Gaussian graphical models, and a brief introduction to the literature on verbal autopsy

analysis. The core methodological chapters will address the challenges described before.

Chapter 3 focuses on the scenario where data contains both continuous and binary variables. We propose a new spike-and-slab prior for sparse inverse correlation matrices, and an efficient Markov chain Monte Carlo algorithm to sample from the resulting posterior distribution. This approach allows us to incorporate informative priors on the marginal distribution of variables directly. We further extend the framework to mixtures of latent Gaussian models for semi-supervised classification tasks with marginal informative priors, with a focus on learning symptom dependencies in VA data and improving cause-of-death classification.

Chapter 4 proposes a new computational tool for Gaussian and Gaussian copula graphical model estimation using an Expectation Conditional Maximization (ECM) algorithm, extended from the EM approach from Bayesian variable selection in linear regression literature. We show that the ECM approach enables fast posterior exploration under a sequence of mixture priors, and can incorporate multiple sources of information. We demonstrate this model using a dataset of aggregated sales of multiple convenient store products.

Chapter 5 combines the mixtures of Gaussian model framework in Chapter 3 with the idea of fast posterior exploration in Chapter 4. We propose a new class of priors for Bayesian inference with multiple Gaussian graphical models with similar precision matrices. We first introduce fully Bayesian treatments of two popular procedures, the group graphical lasso and the fused graphical lasso, and extend them to a continuous spike-and-slab framework that allows self-adaptive shrinkage and model selection simultaneously. We discuss both fully Bayesian inference and a fast EM algorithm for dynamic explorations of posterior modes. We demonstrate the performance of our methods through the modeling of both symptom structures in VA and imputation of missing mortality rates.

Finally, in Chapter 6 we discuss directions for future research.

Chapter 2

BACKGROUND

2.1 Gaussian graphical models

Graphical models (Lauritzen, 1996) describe the conditional dependence structures among random variables by a graph. Consider a random vector $\boldsymbol{x} \in \mathbb{R}^p$. An undirected graph G, represented by a vertex set $V = \{1, 2, ..., p\}$ and an edge set $E = \{(j, k), j, k \in V\}$, defines the conditional independence structure of the probability density of $\boldsymbol{x}$, in the sense that $x_j \perp x_k | \boldsymbol{x}_{\setminus\{j,k\}}$ for any $(j, k) \notin E$. Under the simple setting where $\boldsymbol{x}$ follows a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$, the zeros in off-diagonal elements of precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ correspond to pairs of variables that are conditionally independent.

2.1.1 Inference of Gaussian graphical models using penalized likelihood

For a data matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ consisting of n i.i.d multivariate normal random variables, many algorithms have been proposed to estimate a sparse $\boldsymbol{\Omega}$ under the regularization framework (e.g., Meinshausen and Bühlmann, 2006; Yuan and Lin, 2007; Friedman et al., 2008; Rothman et al., 2008; Friedman et al., 2010; Cai et al., 2010, etc.). One of the most popular procedures is the graphical lasso (Yuan and Lin, 2007). Graphical lasso finds the estimator of the precision matrix that maximizes the likelihood under the l_1-penalization, i.e.,

$$\boldsymbol{\Omega} = \arg\max_{\boldsymbol{\Theta}} n \log(\det \boldsymbol{\Theta}) - \mathrm{tr}(\boldsymbol{S}\boldsymbol{\Theta}) - \lambda ||\boldsymbol{\Theta}||_1$$

where $\boldsymbol{S} = \boldsymbol{X}^T\boldsymbol{X}$ and λ is a tunning parameter, and $||\cdot||_1$ denote the l_1 norm. Several algorithms have been proposed to solve the graphical lasso problem efficiently for high-dimensional data (Mazumder and Hastie, 2012b; Witten et al., 2011; Mazumder and Hastie, 2012a).

2.1.2 Bayesian inference of Gaussian graphical models

In the Bayesian literature, structure learning in high-dimensional Gaussian graphical models has also been widely studied. Broadly speaking, two main classes of priors have been used for the inference of precision matrices in Gaussian graphical models, namely the G-Wishart prior and the shrinkage priors. The G-Wishart prior (Roverato, 2002), developed from the early work of hyper inverse Wishart prior for decomposable graphs (Dawid and Lauritzen, 1993; Giudici and Green, 1999), extends the Wishart distribution by restricting its support to the space of positive definite matrices with zeros specified by a graph, and thus is attractive in Bayesian modeling due to its conjugacy to the Gaussian likelihood. Posterior inference for the graph structure under the G-Wishart distribution, however, poses computational difficulties because the normalizing constant is intractable. Different algorithms have been proposed to sample from the G-Wishart distribution more efficiently, such as shotgun stochastic search (Jones et al., 2005), reversible jump MCMC (Lenkoski and Dobra, 2011; Dobra et al., 2011; Wang and Li, 2012), and more recently birth-death MCMC (Mohammadi et al., 2017).

More closely related to the topics discussed in this dissertation, a rather different line of work on shrinkage priors has gained more popularity in the last few years. As a Bayesian analogy to the widely used graphical lasso, Bayesian graphical lasso has been proposed in Wang (2012) and Peterson et al. (2013). Wang (2015) later draws the connection between the Bayesian variable selection problem (George and McCulloch, 1993) to Bayesian graphical model estimation, and proposed a new class of spike-and-slab prior for precision and covariance

matrices by putting normal mixture priors on the off-diagonal elements. This type of spike-and-slab prior enables a fast block Gibbs sampler that significantly improves the scalability of the model. Several different variants of similar shrinkage priors and their properties have also been studied in the literature (e.g., Talluri et al., 2014; Lin et al., 2017; Banerjee and Ghosal, 2015; Li et al., 2017a).

2.1.3 Graphical models for non-Gaussian and mixed data

The most widely used framework for non-Gaussian graphical model is provided by copulas (Nelsen, 1999). Gaussian copula models the association between the marginally transformed variables using a multivariate Gaussian distribution (e.g., Hoff, 2007; Liu et al., 2012; Dobra and Lenkoski, 2011). Many Bayesian modeling strategies have been proposed to model the univariate marginal transformations together with the covariance matrix of the latent Gaussian distribution (e.g., Pitt et al., 2006). Dobra and Lenkoski (2011) provides a general framework of modeling latent Gaussian graphical model using G-Wishart prior while bypassing the estimation of the marginal transformations with extended rank likelihood (Hoff, 2007). More recently, following the semiparametric approaches propose in Liu et al. (2012) and Xue and Zou (2012), Fan et al. (2016) proposes a two-step procedure for estimating latent graphical model for data with continuous and binary variables.

2.2 Verbal autopsy[1]

Data describing cause of death are critical to formulate, implement and evaluate public health policy. Fewer than one-third of deaths worldwide are assigned a cause, with the most impoverished nations having the least information (Horton, 2007). Verbal autopsy (VA) is

[1]The contents of this section are based on some of the discussions from two papers not included as the main chapters of this dissertation: "Probabilistic cause-of-death assignment using verbal autopsies." (McCormick et al., 2016) and "The `openVA` toolkit for Verbal Autopsies" (Li et al., 2017c).

a well established approach to ascertain cause-of-death when medical certification and full autopsies are not feasible or practical (Garenne, 2014; Taylor et al., 1983). After a death is identified, a specially-trained fieldworker interviews the caregivers (usually family members) of the decedent. A typical VA interview includes a set of structured questions with categorical or quantitative responses and a narrative section that records the 'story' of the death from the respondent's point of view (World Health Organization, 2012a). Currently, there are multiple commonly used questionnaires with overlapping, but not identical questions. The resulting data contains a mixture of binary, numerical, categorical and narrative data. The data are then usually pre-processed into a standard set of binary indicators for which many methods have been proposed to automatically assign cause of death.

After VA data are collected, inferring a cause has two additional components. First, there must be some external information about the relationship between causes and symptoms. One means of obtaining this information is directly through expert opinion. A common practice is to ask groups of physicians to rate the likelihood of a symptoms occurring given a particular cause of death, which can be converted into a set of probabilities of observing a symptom given a particular cause. Alternatively, external information can take the form of a small training set of cases with cause of death assigned by in person autopsy. This process is extremely time and resource-intensive, however, and requires assumptions about the generalizability of deaths in the training set to the population of interest. A more common practice is to obtain training data using clinically trained, experienced physicians read a fraction of the interviews and determine causes. To address the fact that physicians frequently do not agree on causes, VA interviews are often read by two physicians, and sometimes three, and the final causes are determined through a consensus mechanism (e.g. Kahn et al., 2012). Second, actually assigning a cause of death requires an algorithmic or statistical method that combines the external information on the cause-symptom relationship with the symptoms

observed in the VA interviews.

2.2.1 A brief survey of cause-of-death assignment algorithms

Due to the typical lack of labeled gold-standard training data and the high dimensionality of both the VA data (signs and symptoms) and the set of causes, a widely used strategy popularized by the InterVA (Byass et al., 2012) software is to use the information describing the relationship between VA data and causes of death provided by physicians in the form of conditional probabilities of symptoms given causes of death. That is, for a particular symptom s, physicians provide the propensity of observing it in a death due to a particular cause c, i.e., $P_{s|c}$. By using such physician-provided information for a pre-defined set of symptoms and causes, VA methods can be applied when gold-standard data are unavailable. In the original InterVA algorithm, the conditional probabilities are provided in the form of ranked lists. The prior information $\boldsymbol{P}_{s|c}$ is gathered as marginal distributions because it is impractical and time-consuming (probably impossible) to ask physicians about the joint distribution of all combinations of hundreds of symptoms. Consequently, the items collected in verbal autopsy questionnaires are considered to be independent, which means the joint distribution of observing a set of symptoms given a particular cause of death is modeled by the product of the corresponding entries in the $\boldsymbol{P}_{s|c}$ matrix.

Several methods have been proposed to automate the assignment of cause of death from VA data using essentially similar conditional probabilities either provided by physicians or calculated from training data. The Institute for Health Metrics and Evaluation (IHME) proposed the Tariff method (James et al., 2011) based on the counts of co-occurrence in the training data. Miasnikof et al. (2015) proposed a similar algorithm to InterVA using a Naive Bayes classifier. All of the above methods are deterministic in the sense that there is no measurement of uncertainty associated with any input information. In our earlier work, we

proposed a probabilistic statistical framework, *InSilicoVA* (McCormick et al., 2016), that infers both individual's cause of death and the population cause of death distribution explicitly, while quantifying the uncertainties in each model component. In this framework, the values of the entries in the $\boldsymbol{P}_{s|c}$ matrix are re-estimated in a data-driven way with truncated priors that preserves their relative rankings.

All of the widely used methods require the assumption of independent symptoms conditional on a given underlying cause of death. While greatly simplifying the problem, violation of such conditional independence assumption could bias the inference on the outcome. Unfortunately, the violation of this assumption has been widely overlooked in practice. The only method where combinations of symptoms are considered is the early work by King and Lu (King and Lu, 2008). This approach, however, relies on regressing the cause of death on stochastic samples of combinations of symptoms in the gold-standard training data. Even when the symptom set is of small to moderate size, exploring the entire space of all possible combinations is effectively computationally infeasible.

Moreover, from the practical perspective, three main components are required to analyze VA data: (i) VA survey data, (ii) inputs that give information about the association between symptoms and causes, and (iii) a statistical or algorithmic method for assigning a likely cause. The current state of VA literature does not distinguish between these three, in part because existing software for algorithmic and statistical methods require a specific set of inputs and survey format. This restriction prevents robust comparison between methods and contexts. For more discussions and recommendations of implementing VA algorithms, we refer interesting readers to the `openVA` package (Li et al., 2017c) and its component packages, which are all open-sourced and publicly available on CRAN.

Chapter 3

LATENT GAUSSIAN GRAPHICAL MODELS WITH INFORMATIVE MARGINAL PRIORS[1]

3.1 Introduction

In this chapter we propose a Bayesian framework to infer latent graphical models from data that consist of both continuous and binary variables. We show that our method improves estimation of both the underlying correlation matrix and the discovery of the latent graph structure. Our approach allows the incorporation of informative priors on the marginal distribution of variables directly, which can be useful when sample size is small and contains many missing values. Such marginal informative priors also allows us to extend our method to estimate latent Gaussian mixture models in a semi-supervised fashion, when there are insufficient or no labeled data.

Our model is motivated by estimating the distribution of deaths by cause using verbal autopsy (VA) surveys (Garenne, 2014). VA is a commonly used tool to assess cause of death in areas without complete-coverage civil registration (Horton, 2007; Jha, 2014). Data describing the signs and symptoms leading up to a death are elicited through an interview with caregivers of the decedent. Inferring cause of death from VA data is extremely difficult which has led practitioners to use external information gathered from clinicians and public health experts about the relationships between causes of death and symptom profiles. We demonstrate that with such informative prior knowledge, we could perform cause-of-death

[1]The contents of this chapter are based on the paper "Bayesian inference of latent Gaussian graphical models for mixed data" (Li et al., 2017b).

assignments with little or no training data using the latent Gaussian mixture model.

Our work builds on a rich literature on learning dependence structures under the framework of graphical models (Lauritzen, 1996). In particular, the properties and estimation of the Gaussian graphical model have been extensively studied by many authors (e.g., Yuan and Lin, 2007; Witten et al., 2011; Peterson et al., 2013; Wang, 2015; Dobra, 2014; Mohammadi et al., 2017, to name a few.). The study of the multivariate Gaussian model also provides the basis for modeling non-Gaussian data through copulas (Nelsen, 1999). Copula Gaussian graphical models impose a multivariate Gaussian distribution on the association between the marginally transformed variables. Such marginal transformations may be estimated parametrically (e.g., Pitt et al., 2006) or considered as nuisance parameters (e.g., Hoff, 2007; Dobra and Lenkoski, 2011). More recently, following the semiparametric approaches proposed in Liu et al. (2012) and Xue and Zou (2012), Fan et al. (2016) propose a two-step procedure for estimating latent graphical models for data with continuous and binary variables, which is most similar to the situation addressed by this chapter.

External knowledge about the variables can be extremely valuable in successfully recovering the dependence structure. Sometimes, information on the interactions between variables is known to researchers and can be utilized directly via tuning parameters or hyper priors. For example, Peterson et al. (2013) infers cellular metabolic networks based on prior reference information on the network structures, and Bu and Lederer (2017) improve estimation of brain connectivity networks by incorporating the distance between regions of the brain. In other contexts prior information may be available but not immediately on the same domain as the tuning parameters. In VA, each piece of external information requires a substantial time commitment from expert clinicians or public health officials, many of whose time would otherwise be spent caring for patients. Consequently external information is only available for a subset of marginal distributions of the variables of interest. This situation is common in

demographic surveys where it can be very difficult or impossible to elicit prior beliefs about the joint distribution of all the variables, but the marginal distribution of some variables is available from previous surveys or census data (Schifeling and Reiter, 2016). In this chapter, we are concerned with discovering *associations* among high-dimensional multivariate mixed data by allowing researchers to leverage any available prior knowledge about the *marginal* distributions of variables when data are insufficient, e.g., when the sample size is small or there are many missing values. In particular, in the scenarios where the observed data can be viewed as coming from a mixture of several components, *marginal* (or more precisely, *conditional*) prior information on each predictor can play a significant role in classifying unlabeled observations when training data is scarce or unbalanced. It also allows us to extend current probabilistic cause-of-death assignment algorithms for VA to explicitly incorporate symptom dependence structures. In the remainder of this section we describe the background for VA analysis, current practice, and limitations. In Section 3.2 we describe the proposed latent Gaussian graphical model to characterize the dependence structure in mixed data and present two different prior choices of the latent correlation matrix, reflecting different types of prior beliefs. In Section 3.3 we describe the details of the posterior sampling algorithms. In Section 3.4 we show how the latent Gaussian model could be extended to Gaussian mixture models and integrated into existing VA methods for cause-of-death assignment. Section 3.5 examines the performance of correlation matrix estimation, structure learning, and prediction performance with extensive numerical simulation. In Section 3.6 we apply our methods to a gold standard dataset and data from a health and demographic surveillance system (HDSS) site where only physician coded causes are available. Finally, in Section 3.7 we discuss the remaining limitations of the approach and some future directions for improvement.

3.1.1 Symptom dependence in VA data

VA is a tool used to assign causes to individual deaths and estimate cause-specific mortality fractions (CSMF) for collections of deaths in regions of the world without full-coverage civil registration and vital statistics systems. Typically VA surveys are conducted by interviewing caregivers or family members of a recently deceased person using a standardized questionnaire (Nichols et al., 2018; World Health Organization, 2012a). The resulting data describe the decedent's health history leading up to death with a mixture of binary, numerical, categorical and narrative data. The data are then usually pre-processed into a standard set of binary indicators for which many methods have been proposed to automatically assign cause(s) of death. InterVA (Byass et al., 2012), one of the most extensively used methods, preprocesses the 2012 WHO Standard Instrument (World Health Organization, 2012a) into 245 binary indicators and classifies deaths into a pre-defined list of 60 causes.

In many lower middle-income countries (LMIC), many or most deaths occur at home instead of a hospital or clinic so that traditional medical autopsies are impossible. Even when deaths occur in a health facility, standard autopsies and medical certification of cause of death are either not possible or often prohibitively costly. Consequently, there are very few examples of labeled deaths (deaths with both VA and a cause assigned through traditional autopsy or other medical certification) that can be used as training data for VA. Further, even when some deaths are labeled, either with medical autopsy or by having a clinician review the VA survey data, the fraction of labeled deaths is typically small, leaving substantial room to improve performance with reliable external information. A widely used strategy, is to poll expert clinicians about the relationship between symptoms (as reported by VA surveys) and causes of death. For a particular symptom s, physicians provide the propensity of observing it if a death results from cause c, i.e. $P_{s|c}$. In the original InterVA algorithm the conditional probabilities are provided in the form of ranked lists. The prior information $\boldsymbol{P}_{s|c}$ only consists

of marginal distributions because it is impractical and time-consuming (probably impossible) to ask clinicians about the joint distribution of all combinations of hundreds of symptoms. Without information about the associations between symptoms, methods that only use this expert input, $\boldsymbol{P}_{s|c}$, must assume that questions are independent (typically conditional on a cause).

Despite being influential in practice, associations between symptoms have been largely ignored in the VA literature. The only method that considers combinations of symptoms is work by King and Lu (King and Lu, 2008). Their approach relies on regressing the cause of death on stochastic samples of combinations of symptoms in a gold-standard training dataset, but even when the symptom set is of small to moderate size, exploring the entire space of all possible combinations is computationally infeasible. Like the work of King and Lu, we also consider associations between symptoms. Our approach uses a latent Gaussian graphical model to infer associations while also incorporating any available data about priors on marginal distributions.

3.2 Latent Gaussian graphical model for mixed data

We begin by considering the characterization and estimation of dependence structures in mixed data. Let $\boldsymbol{X} = (X_1, ..., X_n)^T$ denote the data with n observations of p-dimensional random variables. For example in survey data, X_{ij} may represent the response of respondent i on question j. In the VA context, we have p symptoms measured on n VA interviews and each X_{ij} is the response to question j regarding decedent i. We propose to use a latent Gaussian representation to encode the dependence between the variables. We assume that the observed data $\boldsymbol{X}$ can be represented by a set of multivariate Gaussian random variables

$\boldsymbol{Z}$ under some monotone transformation:

$$X_{ij} = f_j(Z_{ij}) \quad \boldsymbol{Z}_i \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{R}) \tag{3.1}$$

where $\boldsymbol{R}$ is a correlation matrix, and $f_j(\cdot)$'s are non-decreasing functions. The latent Gaussian distribution provides a simplistic description of the conditional independence relationship for $\boldsymbol{Z}$. Zeros in off-diagonal elements of the inverse correlation matrix, $\boldsymbol{R}^{-1}$, correspond to pairs of latent variables that are conditionally independent given other latent variables.

When the marginal transformation functions are unknown, this formulation is usually referred to as the Gaussian copula model (e.g., Xue and Zou, 2012). For continuous variables, a popular strategy to deal with the marginal transformation f_j is to first estimate it by $\hat{f}_j(z) = \tilde{F}_j^{-1}(\Phi(z))$, where $\tilde{F}_j$ is typically taken to be the empirical marginal cumulative distribution function of the j-th variable (e.g. Klaassen and Wellner, 1997; Liu et al., 2009). Inference on $\boldsymbol{R}$ is then performed with pseudo-data $\hat{Z}_{ij} = \hat{f}_j^{-1}(X_{ij})$. However, this strategy is problematic for discrete data, since directly applying monotonic marginal transformations changes only the sample space instead of the distribution of the observed data (Hoff, 2007). Therefore, for data with mixed variable types, it is common to adopt the semi-parametric marginal likelihood approach (Hoff, 2007). Inference on the correlation matrix is then carried out based on the marginal likelihood of the observed ordering of the variables, with the marginal transformation functions considered as nuisance parameters.

Moving now to binary variables, the marginal distribution can be characterized by the marginal probability, a single parameter, and direct estimation of the transformation functions can be reduced to estimating cutoffs of the latent Gaussian variables (Fan et al., 2016). Conceptually this provides a way to incorporate marginal prior probabilities for binary variables. For example in a VA survey, the marginal probability of observing a rare symptom

can be difficult to estimate empirically from data, but a reasonable prior can be obtained easily using expert input from physicians or data collected elsewhere. Since the latent mean variable $\boldsymbol{\mu}$ is one of the parameters we wish to estimate, the marginal transformation of the binary variables is no longer identifiable. Thus we can fix the marginal transformation and let

$$\begin{aligned} X_{ij} = f_j(Z_{ij}) &= \begin{cases} I(Z_{ij} > 0) & \text{if } X_{ij} \text{ binary} \\ Z_{ij} & \text{if } X_{ij} \text{ continuous} \end{cases} \\ \boldsymbol{Z_i}|\boldsymbol{\mu}, \boldsymbol{R} &\sim \text{Normal}(\boldsymbol{\mu}, \tilde{\boldsymbol{R}}), \end{aligned}$$

where the dependence between covariates is characterized by the covariance matrix, $\tilde{\boldsymbol{R}}$, with diagonal elements corresponding to binary variables fixed at 1. That is, we can write $\tilde{\boldsymbol{R}} = \boldsymbol{\Lambda}\boldsymbol{R}\boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda}$ is a diagonal matrix that contains marginal standard deviations for the continuous variables, and $\boldsymbol{R}$ is a correlation matrix. The marginal prior probabilities for binary variables can then be specified though the priors for $\boldsymbol{\mu}$, since the expectation of X_{ij} given $\boldsymbol{\mu}$ is $Pr(X_{ij} = 1) = \Pr(Z_{ij} > 0) = 1 - \Phi(-\mu_j) = \Phi(\mu_j)$. Thus when $p_j = \Pr(X_{ij} > 1)$ is available *a priori*, we can let

$$\boldsymbol{\mu}|\boldsymbol{\mu}_0 \sim \text{Normal}(\boldsymbol{\mu}_0, \sigma^2\boldsymbol{I}_p), \quad \text{and} \quad \mu_{0j} = \Phi^{-1}(p_j).$$

For simplicity throughout this chapter we assume the continuous variables are marginally Gaussian, similar to the scenario considered in Fan et al. (2016). The extension to the case where the continuous variables exhibit non-Gaussian marginal patterns is straightforward by first preprocessing the raw continuous variables into pseudo-data using their marginal prior distributions (Liu et al., 2009), $\tilde{F}_j$, so that $X_{ij} = \Phi^{-1}(\tilde{F}_j(X_{ij}^{(raw)}))$. Specifying priors on the elements of $\boldsymbol{\Lambda}$ usually depends on the context. In this chapter we adopt the improper prior on the marginal standard deviations suggested in Gelman (2006), so that $\Lambda_{jj} \propto 1$.

The transformation of the marginal prior probabilities to $\boldsymbol{\mu}_0$ in the proposed model requires $\tilde{\boldsymbol{R}}$ to have unit variance for the binary variables, or equivalently, the submatrix of $\tilde{\boldsymbol{R}}$ corresponding to binary variables to be a correlation matrix. Posterior sampling on the space of the correlation matrices is generally more difficult than from the covariance matrices due to the constraint of unit diagonal elements, Further, conjugate priors do not exist for easy Bayesian inference. We adopt a parameter expansion (PX) scheme (Liu and Wu, 1999; Meng and Van Dyk, 1999), so that the correlation matrix $\boldsymbol{R}$ is first expanded to a covariance matrix and updated, and then projected back to the space of correlation matrices.

We discuss two classes of priors for $\boldsymbol{R} = \boldsymbol{\Lambda}^{-1}\tilde{\boldsymbol{R}}\boldsymbol{\Lambda}^{-1}$ that lead to efficient posterior inference: one with the standard conjugate priors for the covariance matrix and uniform marginal priors for $\boldsymbol{R}$, and one with a sparse structure in $\boldsymbol{R}^{-1}$. Similar priors for marginally uniform $\boldsymbol{R}$ were proposed in Talhouk et al. (2012) for the multivariate probit model. Their direct generalization to sparse $\boldsymbol{R}^{-1}$ uses a Metropolis-Hasting algorithm that is computationally expensive and imposes an additional decomposability constraint on the graph structure. A major advantage of the proposed model, summarized in Section 3.3, is the computational simplicity of posterior sampling, as well as the removal of the decomposability constraint.

3.2.1 Marginally uniform prior for the correlation matrix

First, we illustrate a marginally uniform prior on the correlation matrix, and the corresponding parameter expansion scheme. Without any additional knowledge about the structure of the latent correlation matrix, the marginal uniform prior on all the elements of $\boldsymbol{R}$ (Barnard et al., 2000) is

$$p(\boldsymbol{R}) \propto |\boldsymbol{R}|^{-(p+1)} \prod_j (r^{jj})^{-\frac{p+1}{2}}, \quad r^{jj} = \{\boldsymbol{R}^{-1}\}_{jj}.$$

For the model $\boldsymbol{Z}_i \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{R})$, sampling from the posterior distribution $p(\boldsymbol{R}|\boldsymbol{Z}, \boldsymbol{\mu})$ is not straightforward. However, with parameter expansion, we can expand the correlation matrix

into the covariance matrix by $\boldsymbol{\Sigma} = \boldsymbol{DRD}$, where $\boldsymbol{D} = \text{diag}(d_1, ..., d_p)$, and the observed data model into $\boldsymbol{DZ}_i \sim \text{Normal}(\boldsymbol{D\mu}, \boldsymbol{\Sigma})$. By carefully constructing the augmentation of the expansion parameters, the expanded covariance or precision matrix can be much easier to sample from. Following Talhouk et al. (2012), we put an inverse gamma prior on the expansion parameters,

$$d_j^2 | R \sim \text{InvGamma}((p+1)/2, r^{jj}/2),$$

that induces an inverse Wishart prior on the expanded precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1} \sim \text{Wishart}(p+1, \boldsymbol{I}_p)$. The conjugacy allows easy posterior updating of $\boldsymbol{\Sigma}$. This marginally uniform prior does not directly impose any sparsity constraints on the precision matrix. To summarize the conditional independence structure in a more concise manner, one option would be to estimate a sparse representation of $\hat{\boldsymbol{R}}^{-1}$ using a two-stage procedure similar to Fan et al. (2016) with the posterior mean $\hat{\boldsymbol{R}}$ as input. Alternatively, we could incorporate sparsity directly into the prior, which we describe in the next section.

3.2.2 Spike-and-slab prior for the inverse correlation matrix

The marginally uniform prior for $\boldsymbol{R}$ is sometimes inappropriate for settings where sparse structure in $\hat{\boldsymbol{R}}^{-1}$ is strongly suspected *a priori*. For example with VA data, we expect a small number of symptoms to be strongly correlated. We would expect, for example, that pregnancy-related symptoms would be correlated but would be conditionally independent of other clusters of symptoms. Several priors for sparse precision matrices have been proposed. The G-Wishart prior (Roverato, 2002) extends the Wishart distribution by restricting cells in the precision matrix that correspond to non-edges in a graph to be exact zeros, and has been extensively studied in existing literature (Jones et al., 2005; Lenkoski and Dobra, 2011; Mohammadi et al., 2017). More recently shrinkage priors have become more popular, in part due to their computational simplicity. Bayesian analogies to penalized precision matrix

estimators have been proposed for Lasso (Wang, 2012; Peterson et al., 2013), horseshoe (Li et al., 2017a) and spike-and-slab mixture penalties (Wang, 2015; Li and McCormick, 2017; Deshpande et al., 2017). In this work we adapt the spike-and-slab prior idea proposed in Wang (2015) and propose a mixture prior for the inverse correlation matrix. The supplement material contains a brief introduction to Wang's original proposal and its relationship to Wishart priors. The spike-and-slab framework is appealing because it performs graph selection and parameter inference simultaneously, in contrast to other shrinkage priors that require a further thresholding step after shrinkage. We put independent Gaussian priors on each off-diagonal element of the inverse correlation matrix, $\boldsymbol{R}^{-1}$, i.e.

$$\begin{aligned} p(\boldsymbol{R}|\boldsymbol{\delta}) &= C_{\boldsymbol{\delta}}^{-1}|\boldsymbol{R}|^{-(p+1)} \prod_{j<k} \text{Normal}(r^{jk}|0, v_{\delta_{jk}}^2) \prod_{j} \text{Exp}(r^{jj}|\lambda/2)\mathbf{1}_{\boldsymbol{R}\in R^+} \\ p(\boldsymbol{\delta}|\pi_\delta) &\propto C_{\boldsymbol{\delta}} \prod_{j<k} \pi_\delta^{\delta_{jk}}(1-\pi_\delta)^{1-\delta_{jk}} \end{aligned}$$

where R^+ denotes the space of correlation matrices, and $C_{\boldsymbol{\delta}}$ is the normalizing constant, which cancels out to result in the marginal prior (3.2) on the expanded parameter space, similar to Wang (2015). We show in the supplementary material that $C_{\boldsymbol{\delta}}$ is finite and thus both distributions are proper. The proposed setup differs from current literature on shrinkage priors in two ways. First, we restrict the support of $\boldsymbol{R}$ to the space of the correlation matrix, so that working with latent variables that cannot be normalized does not create identifiability issues. In the next section we show that this additional restriction does not increase computational cost by much. Second, we add a $|\boldsymbol{R}|^{-(p+1)}$ term to ensure that the prior assigns no weight to degenerate $\boldsymbol{R}^{-1}$. This term also allows the marginal distribution of $\boldsymbol{\Omega}$ after parameter expansion to be in a form similar to the spike-and-slab prior defined in Wang (2015). In general, any $|\boldsymbol{R}|^m$ with nonzero m can be used in place of this term. The optimal choice for m represents a potential topic for future research.

Finally, we complete the parameter expansion scheme by defining the expansion parameter $\boldsymbol{D}$ such that $d_j^2 \sim \text{InvGamma}((p+1)/2, 1/2)$. The expanded precision matrix $\boldsymbol{\Omega} = (\boldsymbol{DRD})^{-1}$ has the following marginal prior distribution:

$$p(\boldsymbol{\Omega}|\pi_\delta) \propto \prod_{j<k} \pi_\delta^{\delta_{jk}}(1-\pi_\delta)^{1-\delta_{jk}} \prod_{j<k} \exp(-\frac{\omega_{jk}^2}{2v_{\delta_{jk}^2}/\sigma_j^2\sigma_k^2}) \prod_j \exp(-\frac{\lambda\sigma_j^2}{2}\omega_{jj} - \frac{1}{2\sigma_j^2})\mathbf{1}_{\boldsymbol{\Omega}\in M^+} \quad (3.2)$$

where σ_j^2 is the j-th diagonal element of $\boldsymbol{\Omega}^{-1}$. This expanded prior can be derived with a standard change of variables, as described in more detail in the supplementary material. The dependence between $\boldsymbol{\Omega}$ and $\{\sigma_j^2\}_{j=1,\ldots,p}$ makes the posterior sampling seem complicated. However, it turns out that it can be efficiently sampled with a block update. We fully describe our sampling scheme in detail in Section 3.3.1.

3.2.3 Choosing the shrinkage parameters

The proposed prior for $\boldsymbol{R}$ has several hyperparameters, v_0, v_1, λ, and π_δ, that jointly determine the prior scales and sparsity of $\boldsymbol{R}^{-1}$. The relationship between the implied prior sparsity, i.e., $p(\delta = 1)$ and the hyperprameters, however, cannot be easily obtained, because of the constrained space of R^+ and the intractable normalizing constant $C_{\boldsymbol{\delta}}$. We follow a similar practice to Wang (2015) in choosing the hyperparameters by simulating the implied prior edge probabilities from different combination of hyperparameters. We use the sampler in Section 3.3 and choose the values that lead to the desired prior sparsity.

Generally, v_1/v_0 needs to be large so that it gives enough separation between the spike-and-slab densities. The choice of v_0 also needs to be carefully considered: an extremely small v_0 leads to a density that approaches the point-mass and thus can slow the mixing of the Markov chain, while a larger v_0 may absorb many elements of $\boldsymbol{R}^{-1}$ and assigns a heavy portion of prior mass on the 'sparse' models with many small values. The choice of v_0 may be roughly guided

by comparing the marginal distributions implied by the prior to a pre-specified threshold for practical significance. We let $v_0 = 0.01$ in our experiments, as it can be seen from the prior simulation in Figure 3.1 that it assigns reasonable weights to graphs with edge probability between 0.05 to 0.2 under various choice of v_1 and π_δ. Because of the linear constraints on the elements of $\boldsymbol{R}^{-1}$ imposed by the space of R^+, the hyperparameter π_δ typically differs from the implied marginal edge probability significantly, and also needs to be determined from numerical simulation. From Figure 3.1, the prior sparsity is relatively consistent for $v_0 = 0.01$ when $v_1/v_0 > 50$ and $\pi_\delta < 0.001$ We chose $v_1/v_0 = 100$ and $\pi_\delta = 0.0001$ in our experiments.

It is also worth noting that λ also contributes to the prior sparsity directly, as it regularizes the diagonal elements of $\boldsymbol{R}^{-1}$. Since the support of diagonal elements of $\boldsymbol{R}^{-1}$ are $(1, \infty)$, large λ restricts r^{jj} to be closer to 1, leading the correlation between the j-th variable and other variables to be closer to 0, and thus sparser models. From our prior simulation, we found the choice of $\lambda = 10$ usually leads to reasonable prior sparsity. We include more discussion of the relationships between the proposed prior and that of Wang (2015) in the supplementary materials.

3.3 Posterior inference

Inference using the full model can be performed using Markov Chain Monte Carlo with mostly Gibbs steps and elliptical slice sampling (ESS), a rejection-free MCMC technique (Murray et al., 2010). We first describe in detail the sampling procedure with the spike-and-slab prior, and then describe how this step fits into the the full inference procedure in Section 3.3.2.

3.3.1 Posterior sampling with the spike-and-slab prior

We begin by describing sampling with the spike-and-slab prior. We update $\boldsymbol{\Omega}$ with the prior defined in (3.2) in a column-wise fashion. Consider the j-th row and column of $\boldsymbol{\Omega}$, if we

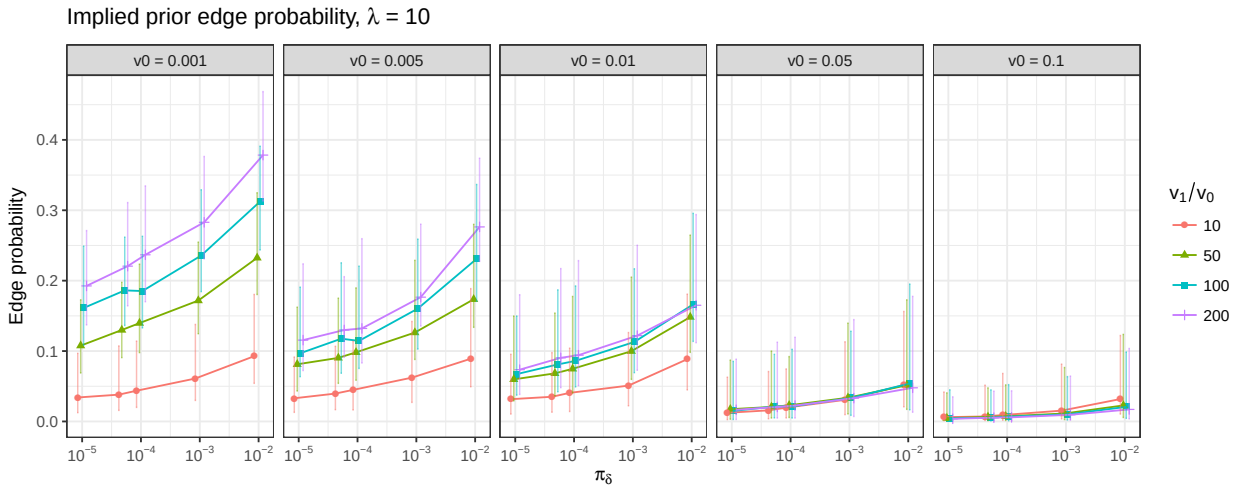


Figure 3.1: Implied prior edge probability with $\lambda = 10$ for $p = 100$ graph. The dots represent the median prior probabilities and the error bars represent the 0.025 and 0.975 quantiles. The densities are derived from sampling $1,000$ draws using MCMC from the prior distribution after $1,000$ iterations of burn-in.

denote $\boldsymbol{u} = \boldsymbol{\Omega}_{[j,-j]}$ and the Schur complement $v = \boldsymbol{\Omega}_{[j,-j]} - \boldsymbol{\Omega}^T_{[j,-j]}\boldsymbol{\Omega}^{-1}_{[-j,-j]}\boldsymbol{\Omega}_{[j,-j]}$, then given the expanded sample covariance matrix, $\boldsymbol{S} = \sum_{i=1}^n \boldsymbol{D}(\boldsymbol{Z}_i - \boldsymbol{\mu})'(\boldsymbol{Z}_i - \boldsymbol{\mu})\boldsymbol{D}$, and the variance specified by the latent indicators, $\boldsymbol{V} = \{v^2_{\delta_{jk}}\}_{jk}$, the joint distribution of $\boldsymbol{u}$ and v can be calculated as

$$p(\boldsymbol{u}, v|\boldsymbol{S}, \boldsymbol{V}) \propto v^{\frac{n}{2}} \exp(-\frac{1}{2}(\boldsymbol{u}'\tilde{\boldsymbol{V}}\boldsymbol{u} + 2\boldsymbol{s}'_{[j,-j]}\boldsymbol{u} + (s_{jj} + \lambda\sigma_j^2)(v + \boldsymbol{u}'\boldsymbol{\Omega}_{[-j,-j]}\boldsymbol{u})))$$

where $\tilde{\boldsymbol{V}} = \{v^2_{\delta_{jk}}/\sigma_j^2\sigma_k^2\}_{jk}$. Notice that $\sigma_j^2 = 1/v$, and for all $k \neq j$, σ_k^2 depends on both $\boldsymbol{u}$ and v, rendering the block Gibbs update scheme in Wang (2015) inapplicable. However, the full conditional distribution for $\boldsymbol{u}$ and v can both be written as the product of a standard distribution and an additional correction term. We let

$$\hat{\boldsymbol{D}} = \text{diag}(\{\frac{d_k^2}{v^2_{\delta_{jk}}}\}_{k\neq j}) \quad \text{and} \quad \tilde{\boldsymbol{D}}(\boldsymbol{u}, v) = \text{diag}(\{\frac{\sigma_k^2 - d_k^2}{v^2_{\delta_{jk}}}\}_{k\neq j}),$$

then we have the full conditional distributions

$$\begin{aligned} p(\boldsymbol{u}|v,\boldsymbol{S},\boldsymbol{V}) &\propto \text{Normal}(\boldsymbol{u};-\boldsymbol{C}\boldsymbol{S}_{[j,-j]},\boldsymbol{C})\exp(-\frac{1}{2v}\boldsymbol{u}'\tilde{\boldsymbol{D}}(\boldsymbol{u},v)\boldsymbol{u}-\frac{1}{2}\sum_{k\neq j}\frac{1}{\sigma_k^2}) \\ p(v|\boldsymbol{u},\boldsymbol{S},\boldsymbol{V}) &\propto \text{Gamma}(v;\frac{n}{2},\frac{s_{jj}+1}{2})\exp(-\frac{1}{2v}\boldsymbol{u}'(\hat{\boldsymbol{D}}+\tilde{\boldsymbol{D}}(\boldsymbol{u},v)+\lambda\boldsymbol{\Omega}^{-1}_{[-j,-j]})\boldsymbol{u}) \end{aligned}$$

where $\boldsymbol{C}=((s_{jj}+\lambda/v)\boldsymbol{\Omega}^{-1}_{[-j,-j]}+\hat{\boldsymbol{D}}))^{-1}$. To sample from $p(\boldsymbol{u}|\cdot)$, we use elliptical slice sampling (ESS) (Murray et al., 2010) to sample from both distributions by treating the normal distribution part as "prior" and the later term as "likelihood." For $\boldsymbol{u}$, ESS first generates an elliptical locus from the normal prior and then searches for acceptable points for slice sampling. ESS typically sticks to the same posterior region when strong signals are provided in the "prior" Gaussian distribution, as is the case here. Additionally when $\boldsymbol{\Omega}^{-1}$ is sparse, σ_k^2 and d_k^2 should be close to each other, and thus the signal from the "prior" part is typically much stronger. To implement ESS for v, we approximate the Gamma likelihood in $p(v|\cdot)$ by Normal$(v;\frac{n}{s_{jj}+1},\frac{2n}{(s_{jj}+1)^2})$. This approximation is typically reasonable given the size of n in the data we consider, and this again allows easy use of ESS. Furthermore, the added computational burden of ESS over the block Gibbs sampler in Wang (2015) is minimal, as the $\{\sigma_k^2\}$'s can be easily calculated by the fact that $\boldsymbol{\Sigma}_{[-j,-j]}=\boldsymbol{\Omega}^{-1}_{[-j,-j]}+\frac{1}{v}\boldsymbol{\Omega}^{-1}_{[-j,-j]}\boldsymbol{u}'\boldsymbol{u}\boldsymbol{\Omega}^{-1}_{[-j,-j]}$, and $\sigma_j^2=1/v$, without any additional computation of a matrix inversion. Finally, each time a block update is performed, all latent indicators can be updated with the corresponding conditional posterior inclusion probabilities,

$$\Pr(\delta_{jk}=1|\boldsymbol{R})=\frac{\pi_\delta\phi(r^{jk}|0,v_1^2)}{\pi_\delta\phi(r^{jk}|0,v_1^2)+(1-\pi_\delta)\phi(r^{jk}|0,v_0^2)}\ .$$

3.3.2 Sampling from the posterior

Given suitable initial values, the full sampling scheme updates each parameter in turn.

Update $\boldsymbol{Z}$. The conditional posterior distributions of the latent variables conditional on the observed data are truncated Normal($\boldsymbol{\mu}, \tilde{\boldsymbol{R}}$) distributions with the truncation defined by domain I_{ij} where $I_{ij} = (-\infty, 0)$ if X_{ij} binary and $X_{ij} = 0$, $(0, +\infty)$ if X_{ij} binary and $X_{ij} = 1$, and $(-\infty, +\infty)$ if X_{ij} is missing or continuous. To sample from the multivariate truncated normal posterior, we draw approximate samples by iteratively sampling $Z_{ij}|\boldsymbol{Z}_{i,-j}$ by

$$Z_{ij}|\boldsymbol{Z}_{[i,-j]}, \tilde{\boldsymbol{R}}, \boldsymbol{\mu}, \boldsymbol{X} \sim \text{TruncNorm}(\tilde{\boldsymbol{\mu}}_0, \tilde{\sigma}, I_{ij})$$

where $\tilde{\boldsymbol{\mu}}_0 = \boldsymbol{\mu}_j + (\boldsymbol{Z}_{[i,-j]} - \boldsymbol{\mu}_{-j})(\tilde{\boldsymbol{R}}_{[j,-j]}\tilde{\boldsymbol{R}}^{-1}_{[-j,-j]})^T$, $\tilde{\sigma} = \sqrt{1 - \tilde{\boldsymbol{R}}_{[j,-j]}\tilde{\boldsymbol{R}}^{-1}_{[-j,-j]}\tilde{\boldsymbol{R}}_{[-j,j]}}$, and the truncated domain I_{ij} is defined above.

Update $\boldsymbol{\Lambda}$. We perform the conditional update of $\boldsymbol{\Lambda}$ by sampling from $p(\Lambda^{-1}_{jj}|\boldsymbol{\Lambda}_{[-j,-j]}, \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{R})$ iteratively. The improper uniform prior on Λ_{jj} is equivalent to $p(\Lambda^{-1}_{jj}) \propto \Lambda^2_{jj}$, leading to the conditional posterior distribution

$$p(\Lambda^{-1}_{jj}|\boldsymbol{\Lambda}_{[-j,-j]}, \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{R}) \propto \Lambda^{-(n-2)}_{jj}\text{Normal}(\Lambda^{-1}_{jj}; \frac{\sum_i b_i(z_{ij} - \mu_j)}{\sum_i (z_{ij} - \mu_j)^2}, \frac{c}{\sum_i (z_{ij} - \mu_j)^2})$$

where the constant terms are

$$b_i = \boldsymbol{\Lambda}_{[-j,-j]}\boldsymbol{R}_{[-j,j]}\boldsymbol{R}^{-1}_{[-j,-j]}(z_{i,-j} - \boldsymbol{\mu}_{-j}), \quad c = \boldsymbol{\Lambda}_{[-j,-j]}\boldsymbol{R}_{[-j,j]}\boldsymbol{R}^{-1}_{[-j,-j]}\boldsymbol{R}_{[j,-j]}\boldsymbol{\Lambda}_{[-j,-j]} \ .$$

These conditional distributions can be efficiently sampled with ESS (Murray et al., 2010).

Update $\boldsymbol{\mu}$. The conditional posterior distribution for the mean parameters is also multivariate normal,

$$\boldsymbol{\mu}|\tilde{\boldsymbol{R}}, \boldsymbol{X} \sim \text{Normal}\left((\frac{1}{\sigma^2}\boldsymbol{I}_p + n\tilde{\boldsymbol{R}}^{-1})^{-1}(\frac{1}{\sigma^2}\boldsymbol{\mu}_0 + n\tilde{\boldsymbol{R}}^{-1}\bar{z}), (\frac{1}{\sigma^2}\boldsymbol{I}_p + n\tilde{\boldsymbol{R}}^{-1})^{-1}\right).$$

Update $\boldsymbol{R}$. To update the latent correlation matrix, we first draw the working expansion parameter with $d_j^2|\boldsymbol{R} \sim \text{InvGamma}((p+1)/2, \beta)$, where $\beta = r^{ii}/2$ for the marginally uniform prior, and $\beta = 1/2$ for the spike-and-slab prior. The inverse Gamma distribution is parameterized with shape and scale. We then construct the expanded observation $\boldsymbol{W} = \boldsymbol{ZD}$, where $\boldsymbol{D} = \text{diag}(d_1, d_2, ..., d_p)$, and compute the sample covariance matrix $\boldsymbol{S} = \sum_{i=1}^{n}(W_i - \boldsymbol{D\mu})'\boldsymbol{\Lambda}^{-2}(W_i - \boldsymbol{D\mu})$. For the marginally uniform prior, the posterior conditional distribution of the expanded precision matrix $\boldsymbol{\Omega}$ takes the conjugate form,

$$\boldsymbol{\Omega}|\boldsymbol{W}, \boldsymbol{\mu} \sim \text{Wishart}(\boldsymbol{I}_p + \boldsymbol{S}, n + p + 2) \ .$$

For the spike-and-slab prior, we sample the expanded precision matrix $\boldsymbol{\Omega}|\boldsymbol{W}, \gamma$ using ESS as described in Section 3.3.1. After a new $\boldsymbol{\Omega}$ is sampled, we can then compute the induced expansion parameter $\boldsymbol{D} = \text{diag}(\sigma_1^2, ..., \sigma_p^2)^{\frac{1}{2}}$ and the induced correlation matrix $\boldsymbol{R} = \boldsymbol{D}^{-1}\boldsymbol{\Omega}^{-1}\boldsymbol{D}^{-1}$. For problems with very large p, it may also be useful to perform the posterior sampling in two stages, where the first stage updates all the parameters, while in the second stage, $\boldsymbol{\delta}$ is fixed to be the posterior median graph estimated from the first stage. The two-stage procedure may improve the mixing of the chain by reducing the dimension of discrete parameters in the second stage, especially in the mixture model case discussed in the next section. For all the numerical examples used in this chapter, adding an extra post-selection stage does not change the posterior mean estimators of interest by much and thus all results are reported using MCMC with a single stage.

3.4 Cause-of-death assignment using latent Gaussian mixture model

In this section we extend the latent Gaussian graphical models to model data from a mixture of underlying distributions. This extension allows us to complete our model to simultaneously estimate the latent correlation matrix and assign causes of death using VA data. Before we

describe our model, it is worth noting that for many existing automated VA methods such as InSilicoVA (McCormick et al., 2016), InterVA (Byass et al., 2003), and the Naive Bayes Classifier (Miasnikof et al., 2015), the classification rule is closely related to the naive Bayes classifier under the assumption of (conditional) independence between symptoms, i.e.

$$\Pr(y_i = c|\boldsymbol{X}_i) = \frac{\pi_c \prod_j p(X_{ij}|y_i = c)}{\sum_{c=1}^{C} \pi_c \prod_j p(X_{ij}|y_i = c)}.$$

For algorithms using this conditional independence assumption, the information provided by training data (aside from a prior guess of π_c) can be summarized by the conditional relationships between a single sign/symptom and causes. In contexts without training data, expert clinicians provide the same information in the form of informative prior beliefs (e.g. Byass et al., 2003; McCormick et al., 2016). Thus to extend the latent Gaussian graphical model to the context of cause-of-death assignment, we hope to incorporate such conditional relationships as well, in order to make full use of the existing information. This can be achieved similarly as before. We let y_i denote the categorical indicator from a set of C causes of death for person i. A key goal of VA analysis is to associate unlabeled data with cause-of-death assignments. With a generative model similar to Section 3.2, we let $X_{ij} = f(Z_{ij})$, and the data generating mechanism to be

$$\begin{aligned} \boldsymbol{Z}_i|y_i = c &\sim \text{Normal}(\boldsymbol{\mu}_c, \tilde{\boldsymbol{R}}), \quad c = 1, 2, ..., C, \\ \boldsymbol{\mu}_c &\sim \text{Normal}(\boldsymbol{\mu}_{0c}, \sigma_c^2 \boldsymbol{I}_p), \end{aligned}$$

where the priors for $\boldsymbol{\mu}$ and $\tilde{\boldsymbol{R}}$ are the same as in Section 3.2. Following the setup presented in McCormick et al. (2016), we treat the causes of death for unlabeled observations as missing data, and the relationship between symptoms and causes are iteratively re-estimated until the distributions of individual cause-of-death probabilities are compatible with the population

cause-specific mortality fractions (CSMF). We model the distribution of the class assignment indicator with a conjugate Dirichlet prior, $y_i|\boldsymbol{\pi} \sim \text{Multi}(\boldsymbol{\pi})$, and $\boldsymbol{\pi}|\alpha \sim \text{Dirichlet}(\alpha)$.

To account for the different strength of prior information for each mixture, we can also put an additional hyper-prior on σ_c^2. In our experiments with unspecified σ_c^2, we use weak independent priors such that $\sigma_c^2 \sim \text{InvGamma}(0.001, 0.001)$, for $c = 1, ..., C$. Although not presented here, if marginal information on the continuous variable distributions is available in practice, we may also let $X_{ij}|y_i = c$ to be $f_{cj}(z) = \tilde{F}_{cj}^{-1}(\Phi(z))$, where $\tilde{F}_{cj}$ is the fixed marginal distribution function, and inference can be similarly carried out with one additional step to update the observed continuous variables each time an assignment changes.

The mixture model approach allows the joint distribution of symptoms in the data to further guide the estimation of the latent correlation matrix. The proposed model is ideally suited for settings with some, but not extensive, training data. In verbal autopsy this typically happens when a small subset of deaths are assigned a cause either by a traditional medical autopsy or, more commonly, when clinicians review the verbal autopsy data and assign a cause of death, so-called 'physician-coded' VAs. In most settings physician-coded VAs are comparatively (very) rare because physician coding is costly in terms of physician time and opportunity costs, e.g. physicians not seeing living patients. The informative prior setup we propose allows researchers to combine prior or clinician-derived expert information with training data. Conceptually, in the extreme case when no training data exist, the latent Gaussian mixture model can still be estimated given strong informative priors on $\boldsymbol{\mu}$, i.e. the conditional probabilities of symptoms, and the latent correlation matrix will be estimated dynamically based on cause assignments in each iteration. In the following sections we show the advantages of combining both strong priors and limited training data using both simulated and observed data.

Finally, if the labeled and unlabeled deaths come from different populations (e.g. the

labeled deaths occur in a high malaria region whereas the unlabled deaths do not), then one could let the labeled and unlabeled deaths follow two multinomial distributions with different $\boldsymbol{\pi}$, or further include additional subpopulation-specific $\boldsymbol{\pi}$. Posterior inference of $\boldsymbol{\pi}, \boldsymbol{\mu}$ and $\tilde{\boldsymbol{R}}$ can be similarly carried out as in Section 3.3.2 with minor modifications. We leave the detailed algorithms in the supplementary material. After obtaining the posterior mean estimators $\hat{\boldsymbol{\pi}}$, $\hat{\boldsymbol{\mu}}$, and $\hat{\boldsymbol{R}}$ through MCMC, the most likely cause-of-death assignments for each death can be obtained in a discriminant analysis fashion by marginalizing over the latent variable $\boldsymbol{Z}$ to obtain the Bayes classifier as

$$\Pr(y_i = c|\boldsymbol{X}_i) \propto \hat{\pi}_c \int_{z_{i1}\in S_{i1}} \cdots \int_{z_{ip}\in S_{ip}} \phi(\hat{\boldsymbol{\mu}}_c, \hat{\boldsymbol{R}}) dz_{i1} \cdots dz_{ip} \ . \tag{3.3}$$

3.5 Simulation evidence

In this section we conduct simulation experiments to characterize the performance of the proposed method for both the estimation of $\boldsymbol{R}$ under the latent Gaussian framework and classification under the mixture framework. We describe our data generation process and provide results for correlation matrix estimation and graph recovery in Section 3.5.1 and then for classification in Section 3.5.2.

3.5.1 Estimation error and graph recovery

To examine the performance of our method in recovering the latent correlation matrix under different scenarios, we follow a data generating procedure similar to those in Liu et al. (2012) and Fan et al. (2016). In all our simulations, we generate the sparse precision matrix $\boldsymbol{\Omega}$ so that $\omega_{jj} = 1$, and $\omega_{jk} = t a_{jk}$, where $a_{jk} \sim$ Bernoulli$((2\pi)^{-0.5} \exp(||z_j - z_k||_2)/(2c))$ and z_j's are independent bivariate uniform random variables sampled from $[0, 1]^2$. We set $c = 0.2$ so that on average each node has 6.4 edges in the graph, and set t so that the precision matrix is

positive definite. In all our examples we further rescale $\boldsymbol{\Omega}$ so that its inverse is a correlation matrix. We consider the following four scenarios using the assumed generative model:

(i) Let X contain all binary variables, and marginal means for the latent variables $\mu_j \sim \text{Unif}[-1, 1]$, and let the marginal prior $\boldsymbol{\mu}_0$ be the true $\boldsymbol{\mu}$.

(ii) Same as in case (i), except the marginal prior $\boldsymbol{\mu}_{0j}$ is misspecified to be $sign(\mu_{0j}) * \mu_{0j}^2$.

(iii) Assume X contains 10% continuous Gaussian variables and the rest of them are binary, with the correct marginal prior as in case (i).

(iv) Assume X contains 10% continuous Gaussian variables and the rest of them are binary, with a misspecified marginal prior for binary variables described as in case (ii), and further generate continuous variables from the misspecified marginal distribution so that X_{ij}^3 is marginally Gaussian.

Cases (ii) and (iv) reflects the practical scenario where more extreme marginal probabilities are relatively easier to solicit but may be provided on a different scale compared to the truth. In all our simulations we set $n = 200$, $p = 50$, and randomly remove $m\%$ of the entries in the data matrix to represent $m\%$ missing data. We repeat the simulation under each scenario 100 times. For both proposed models, we run the MCMC $3,000$ iterations and report the mean estimator for $\boldsymbol{R}$ from the second half of the posterior draws.

To benchmark the performance of our method in recovering the true correlation matrix, we compare our method with the semi-parametric estimator proposed in Fan et al. (2016). To obtain a fair comparison with our method that uses marginal priors, we calculate the rank-based estimator with the prior marginal probabilities, instead of the empirical marginal probabilities calculated from data. In our experiments described above, this approach leads to better estimation of $\boldsymbol{R}$. We note that this substitution may harm the estimator performance

when marginal priors are misspecified significantly. We compare the estimated correlation matrix error $\hat{\boldsymbol{R}} - \boldsymbol{R}$ in terms of the matrix element-wise maximum norm, spectral norm, and Frobenius norm. The results are in Tables 3.1. The posterior mean estimator $\hat{\boldsymbol{R}}$ from the proposed approach consistently outperforms the rank-based estimator for all three norms and is more robust to missing data and model misspecification.

To evaluate performance for graph recovery under the marginal uniform prior, we use the same two-stage procedure as in Fan et al. (2016) where we first obtain the posterior mean estimator of $\hat{\boldsymbol{R}}$ and then apply graphical Lasso to obtain a sparse $\widehat{\boldsymbol{R}^{-1}}$. For the spike-and-slab prior, we can directly threshold $\widehat{\boldsymbol{R}^{-1}}$ since the conditional posterior inclusion probability $\Pr(\delta_{jk}|\hat{r}^{jk})$ is a monotonically increasing function of $|\hat{r}^{jk}|$. We define the false positive rate and true positive rate in the same way as Fan et al. (2016):

$$\text{FPR} = \frac{\text{FP}}{p(p-1)/2 - |E|}, \qquad \text{TPR} = \frac{\text{TP}}{|E|}$$

where E is the number of edges in the graph. Tables 3.1 also shows the comparison of the ROC curve using AUC and maximum F1 score. Under all scenarios our estimator yields better AUC and F1 scores, especially when the fraction of missing data is high.

3.5.2 Classification error

In this section we illustrate the performance of our method for cause-of-death assignment in VA analysis. We generate $n = 800$ unlabeled data with $p = 50$ from $C = 20$ classes, where the class membership distributions are generated from Dirichlet(1). Data within all groups share the same latent correlation matrix but have different marginal mean vectors generated in the same way as described in 3.5.1. We compare the performance using mixed data, i.e., Case (iii) and (iv), in this subsection.

			$\|\|\hat{\boldsymbol{R}} - \boldsymbol{R}\|\|$			$\widehat{\boldsymbol{R}^{-1}}$	
Scenario	Missing	Estimator	M norm	S norm	F norm	AUC	max F1
Case (i)	0%	Semi-parametric	0.43	2.06	5.95	0.70	0.69
		Uniform prior	0.32	**1.67**	4.55	0.72	0.71
		Spike-and-Slab prior	**0.26**	1.70	**3.40**	**0.86**	**0.80**
	20%	Semi-parametric	0.52	2.45	6.95	0.61	0.67
		Uniform prior	0.36	1.96	5.11	0.66	0.68
		Spike-and-Slab prior	**0.27**	**1.85**	**3.62**	**0.81**	**0.76**
	50%	Semi-parametric	0.64	3.59	9.22	0.44	0.65
		Uniform prior	0.48	3.14	6.90	0.55	0.67
		Spike-and-Slab prior	**0.29**	**2.08**	**3.94**	**0.71**	**0.70**
Case (ii)	0%	Semi-parametric	0.38	1.79	5.39	0.72	0.69
		Uniform prior	0.32	**1.66**	4.55	0.72	0.71
		Spike-and-Slab prior	**0.26**	1.70	**3.40**	**0.86**	**0.80**
	20%	Semi-parametric	0.45	2.15	6.37	0.63	0.67
		Uniform prior	0.36	1.95	5.10	0.66	0.68
		Spike-and-Slab prior	**0.27**	**1.85**	**3.62**	**0.81**	**0.76**
	50%	Semi-parametric	0.59	3.21	8.59	0.47	0.65
		Uniform prior	0.47	3.12	6.85	0.55	0.67
		Spike-and-Slab prior	**0.29**	**2.09**	**3.94**	**0.71**	**0.70**
Case (iii)	0%	Semi-parametric	0.45	2.23	6.13	0.70	0.70
		Uniform prior	0.32	1.60	4.39	0.74	0.72
		Spike-and-Slab prior	**0.28**	**1.41**	**3.31**	**0.87**	**0.81**
	20%	Semi-parametric	0.53	2.60	7.11	0.61	0.67
		Uniform prior	0.35	1.88	4.93	0.68	0.69
		Spike-and-Slab prior	**0.30**	**1.54**	**3.65**	**0.83**	**0.77**
	50%	Semi-parametric	0.64	3.72	9.36	0.44	0.65
		Uniform prior	0.46	2.91	6.48	0.56	0.67
		Spike-and-Slab prior	**0.35**	**1.98**	**4.64**	**0.71**	**0.71**
Case (iv)	0%	Semi-parametric	0.42	1.98	5.62	0.72	0.70
		Uniform prior	0.32	**1.59**	4.39	0.74	0.72
		Spike-and-Slab prior	**0.26**	1.68	**3.37**	**0.87**	**0.81**
	20%	Semi-parametric	0.49	2.32	6.59	0.63	0.67
		Uniform prior	0.35	1.87	4.93	0.68	0.69
		Spike-and-Slab prior	**0.27**	**1.83**	**3.59**	**0.82**	**0.77**
	50%	Semi-parametric	0.61	3.37	8.79	0.46	0.65
		Uniform prior	0.46	2.89	6.46	0.56	0.67
		Spike-and-Slab prior	**0.29**	**2.07**	**3.92**	**0.72**	**0.71**

Table 3.1: Simulation with binary and mixed $\boldsymbol{X}$ under different scenarios. The proposed latent Gaussian graphical model approach (Spike-and-Slab prior) outperforms the semi-parametric alternatives and the marginal uniform prior (Uniform prior) in both scenarios. The Spike-and-slab prior performs especially well in scenarios with a high proportion of missing data.

We compare the average classification accuracy with that produced from the naive Bayes classifier and the underlying algorithm from InterVA (Byass et al., 2012), which is closely related to the naive Bayes classifier. To assess the performance in estimating class probability, as is a main goal in VA analysis, we also compared the estimation of $\boldsymbol{\pi}$ with the truth using 'CSMF accuracy' (Murray et al., 2011b) defined as $ACC_{\text{csmf}} = 1 - \frac{\sum_{c=1}^{C} |\pi_c^{\text{true}} - \hat{\pi}_c|}{2(1-\min \pi^{\text{true}})}$.

For the proposed model, we further investigate the scenario where 80, 100 and 200 labeled data exist. Intuitively, adding labeled data helps our model identify the dependence structure more quickly, especially in the presence of low sample size and high proportion of missing data. However, we do not impose the assumption that the labeled data shares the same class distribution as the testing data to maintain fair comparison. Figure 3.2 display the results. The proposed latent Gaussian model consistently outperforms both the naive Bayes classifier and InterVA model, and is more robust to misspecification.

3.6 Analysis of verbal autopsy data

In this section we present results comparing the proposed model and the naive Bayes classifier using VA data in two contexts. First, in Section 3.6.1, we compare our method to both InterVA and the Naive Bayes Classifier using a set of gold standard data. In this scenario, we have sufficient labeled data to obtain good estimates of the conditional distribution of each symptom given each cause. This setting mimics a scenario where informative prior information is available and of high quality, which is common but not ubiquitous in practice. In Section 3.6.2, we evaluate our methods using data from health and demographic surveillance system (HDSS) sites where the missing data proportion is much higher and the sample sizes are smaller. We compare different methods with physician-coded causes of death and show that the proposed approach is able to improve classification accuracy compared to both InterVA and the Naive Bayes classifier with noisy marginal priors that are poorly specified,

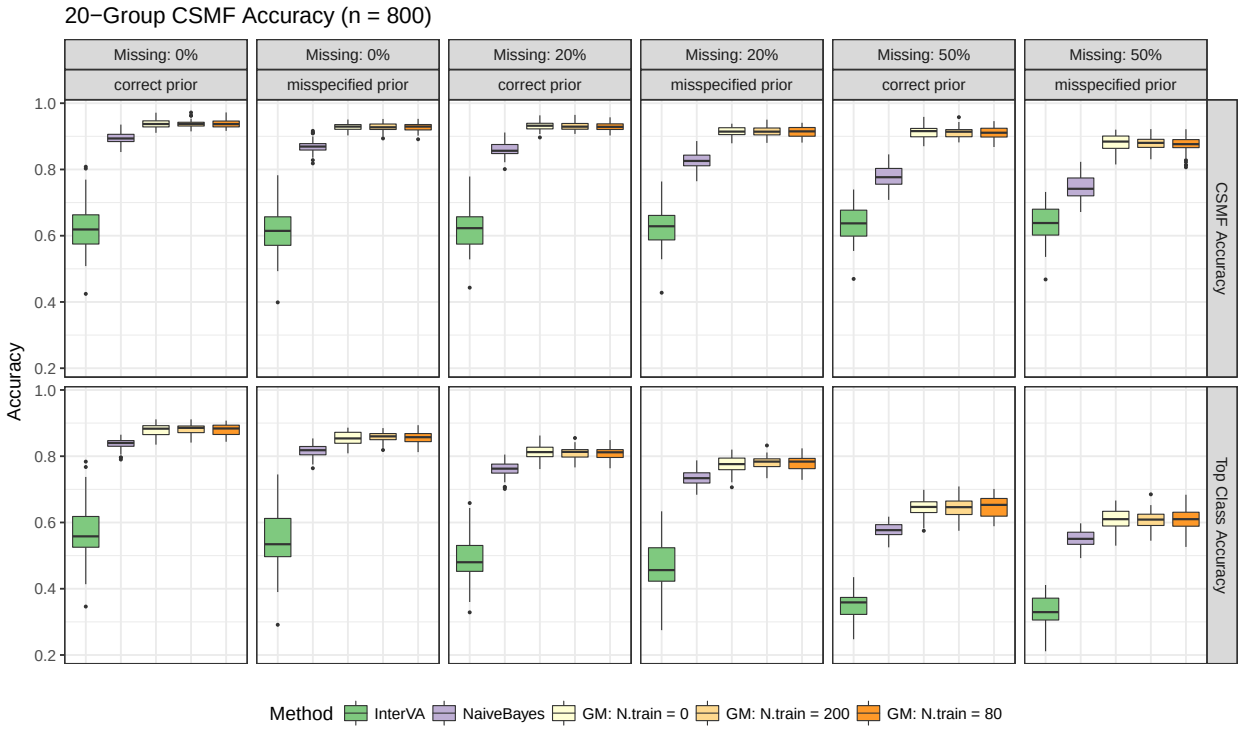


Figure 3.2: Classification and CSMF accuracy for mixed data. Average classification accuracy and CSMF accuracy for different methods with correct and misspecified priors and different proportion of missing data for *mixed data*. Top row: CSMF accuracy. Bottom row: Accuracy of individual most likely class assignment. The accuracy is evaluated in a dataset with a total $n = 800$ observations and $p = 50$ variables including 5 continuous variables from $C = 20$ classes, with or without additional labeled data.

in the scenarios where no or little labeled data are available.

3.6.1 PHMRC gold standard data

We first evaluate the performance of the proposed methods using the Population Health Metrics Research Consortium (PHMRC) 'gold standard' VA dataset (Murray et al., 2011a). The PHMRC dataset consists of about 7,000 deaths recorded in six sites across four countries (Andhra Pradesh, India; Bohol, Philippines; Dar es Salaam, Tanzania; Mexico City, Mexico;

Pemba Island, Tanzania; and Uttar Pradesh, India). Gold standard causes are assigned using a set of specific diagnostic criteria that use laboratory, pathology, and medical imaging findings. All deaths occurred in a health facility. For each death, a blinded verbal autopsy was also conducted. We removed all deaths due to external causes, e.g., homicide, road traffic, etc., since the conditional probabilities of many symptom given an external cause is less meaningful, and external causes are also much easier to identify with a deterministic screening procedure in practice. For the rest of the deaths from 26 causes, we randomly selected $1,000$ deaths as testing data, additional $1,000$ deaths as labeled data, and used the rest of the dataset to calculate the conditional probability matrix of each symptom given each cause as the informative prior. We fit the proposed model both with and without the labeled data, but do not assume the labeled deaths share the same distribution of causes. We repeated this experiment 50 times.

We compared our methods with both InterVA and the Naive Bayes classifier using the same prior information. We ran the MCMC chains for $3,000$ iterations and discarded the first half as burn-in. We put the hyper-prior described in Section 3.4 on σ^2. We used flat Dirichlet prior with $\alpha_c = 1$ for all $c = 1, ..., C$ and calculated the individual cause assignment using Equation 3.3, and compared with the truth in terms of the accuracy of most likely cause, top three most likely causes, and CSMF accuracy. Figure 3.3 shows clear improvements of the proposed method over alternatives that assume conditional independence.

3.6.2 HDSS sites

In this section, we apply our method to a dataset from the Karonga HDSS (Crampin et al., 2012). The Karonga site monitors a population of about $35,000$ in northern Malawi near the port village of Chilumba. The current system began with a baseline census from 2002–2004 and has maintained continuous demographic surveillance with verbal autopsy on all deaths

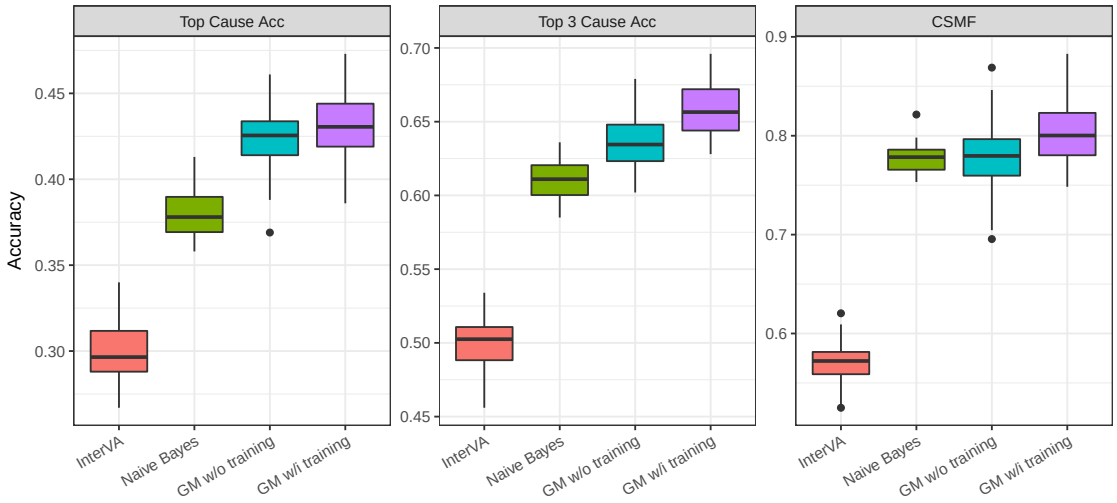


Figure 3.3: Classification and CSMF accuracy for PHMRC cross-validation study. The metrics are evaluated on $1,000$ randomly selected deaths for InterVA, Naive Bayes classifier, and the proposed model without any training data (GM: w/o training). An additional $1,000$ randomly selected labeled death is used as training data in the last case (GM: w/i training). The labeled data are not assumed to have the same distribution of causes.

since 2002. To validate the proposed method, we use 1,900 adult deaths from Karonga that occurred to people of both sexes from 2002–2014. All deaths have both a VA interview and a physician-assigned causes of death.

The Karonga VA data were first coded by two physicians, and if they disagreed, a third physician adjudicated and determined the final cause assignment. These assignments were originally coded into 88 cause categories. We removed the small fraction of deaths due to external causes (such as traffic accident and suicide) from this dataset since they are in practice easy to classify and may be conditionally independent from most of the symptoms. Given the limited sample size, we further aggregated the remaining causes into broader groups. We aggregated the assignments into 16 subcategories. A figure representation of the causes-of-death distribution in the Karonga dataset used in the experiments are presented in Figure 3.4.

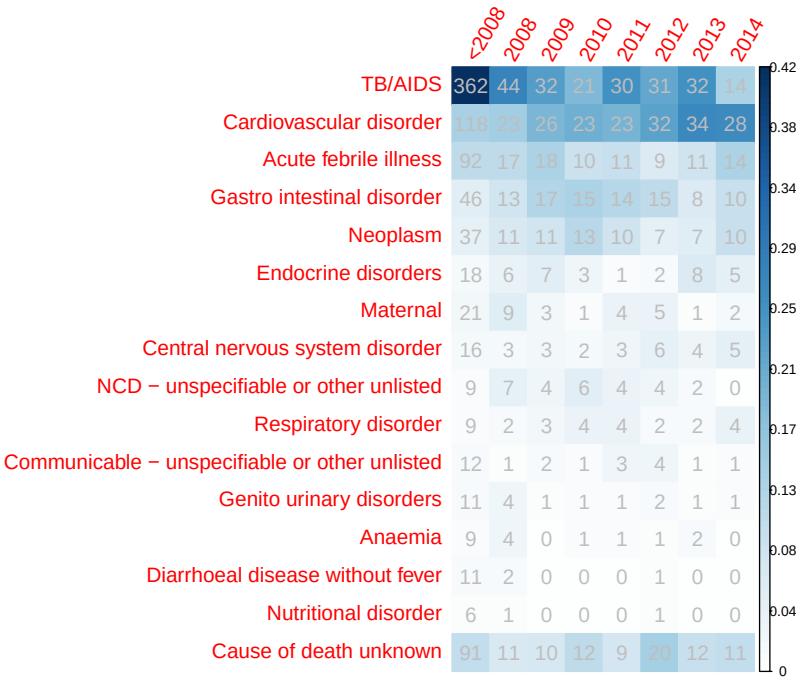


Figure 3.4: Distributions of causes-of-death in Karonga dataset by year. The integers in each cell show the number of deaths in the corresponding period, and the shading represents the proportion of causes in each year. The data before 2008 are used as prior information in the experiment and thus are combined in this figure.

We remove the symptoms that are missing for over 90% of the data which reduces the size of the symptom list to 92. Finally, we formed a "prior" dataset by taking all the deaths (VA symptoms and the physician-assigned causes) during 2002–2007 – about 50% of the entire dataset. Because the physician-provided conditional probabilities, P(symptom|cause), used in InterVA and InSilicoVA are defined with respect to a different cause list, we calculated the empirical P(symptom|cause) matrix from the training data so that P(symptom s|cause c) = (number of $s = 1$ occuring with c)/(number of c), and replace 0 and 1 in the prior probabilities with $0.5p_{min}$ and $1 - 0.5(1 - p_{max})$.

We first fit the model on all the data from 2008–2014 using this empirical P(symptom|cause) matrix. We used the same hyperparameter setup as the previous example with PHMRC. In the VA questionnaires, there are several groups of questions probing different aspects of the same symptom, for example "fever of any kind" and "fever lasting less than 2 weeks", or "male" and any pregnancy-related symptoms. Such questions are expected to be conditional dependent due to the structure of the questionnaire, and thus we fix the corresponding selection indices to be 1 in the inverse correlation matrix. We compare our method with InterVA and the Naive Bayes classifier using the same "prior" P(symptom|cause) matrix. Table 3.2 summarizes the performance of each algorithm, and Figure 3.5 shows the estimated CSMF compared to the truth. The estimated correlation matrix, inverse correlation matrix, and the posterior inclusion probabilities of edges are shown in Figure 3.6.

In addition to the structures induced by the questionnaire, we also recover interesting symptom pairs that are conditionally dependent on each other. For example, history of high blood pressure is strongly positively associated with paralysis of one side of the body across all cross-validation experiments, which is expected given the relatively high prevalence of cardiovascular diseases in the data. In our experiment, there are 3874 potential edges excluding the one knowns from the survey. Table 3.3 summarizes the list of top 25 symptom

	CSMF	Top1	Top2	Top3
InterVA	0.75	0.51	0.62	0.70
Naive Bayes	0.78	0.44	0.64	0.73
Gaussian Mixture	0.84	0.55	0.70	0.78

Table 3.2: CSMF accuracy, Top 1 to 3 cause assignment accuracy for Karonga physician coded data. The marginal probabilities are calculated with data from 2002 to 2007. The training data consist of all the data from 2002 to 2007. The testing data are the rest of the data from 2008 to 2014. The proposed Gaussian mixture model consistently outperform Naive Bayes classifier and InterVA.

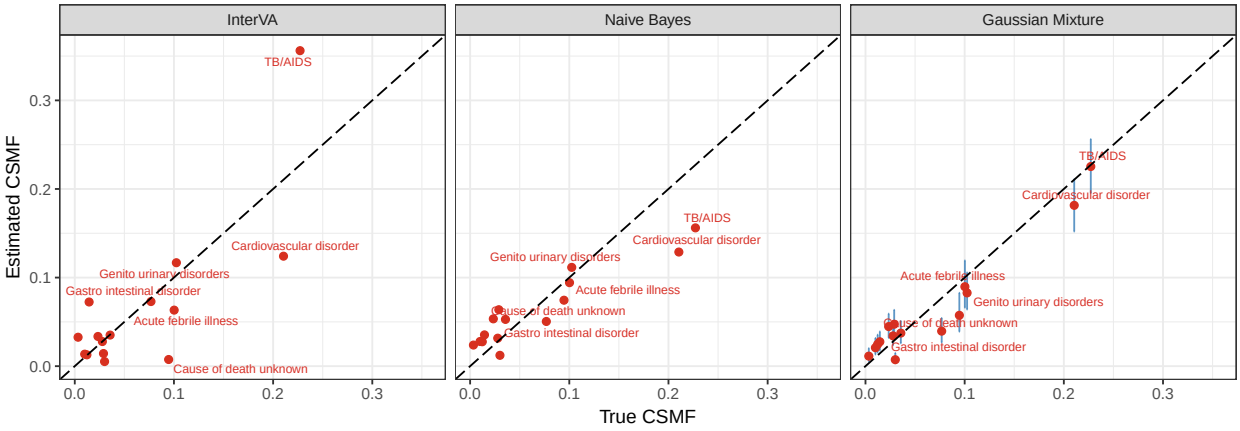


Figure 3.5: Scatter plot of the estimated CSMF against true CSMF for Karonga data from 2008 to 2014 using different methods. Causes with true fractions larger than 0.05 are labeled in the plot. The vertical bars correspond to the 95% posterior credible intervals estimated from the proposed model. The proposed Gaussian mixture model shows smaller bias.

pairs with highest posterior inclusion probability, $\hat{p}(\delta_{jk} = 1|\boldsymbol{X})$, greater than 0.5.

To further demonstrate the performance of the proposed model, we also conducted a cross-validation study with data from 2002 to 2007use to calculate the informative priors. We randomly selected from the rest of the data $\alpha\%$ of training set and use the rest as test set. We repeated the exercise for $\alpha = 5, 10, 20$. Our train-test split differs from standard out-of-sample cross-validation analysis in that we use the smaller fraction as training data, in order to reflect more closely the practical realities of VA data. We assumed the training and testing data

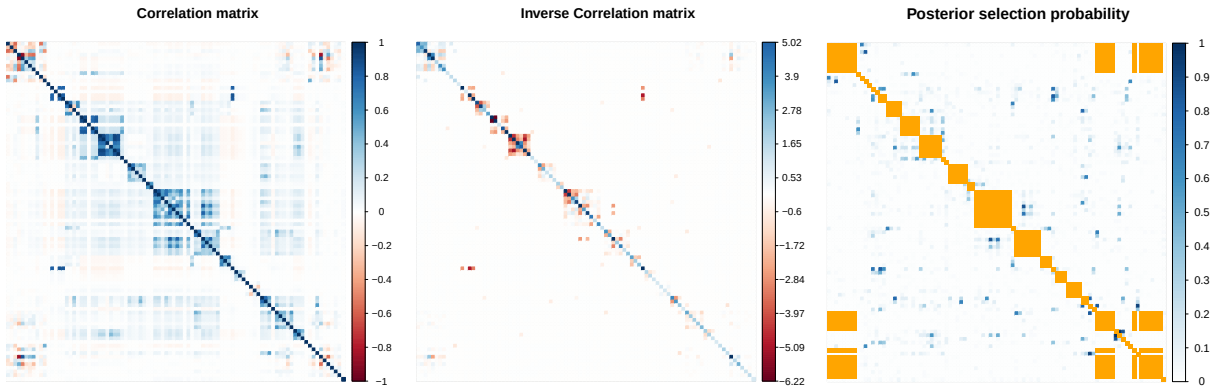


Figure 3.6: Posterior mean correlation (left), inverse correlation (middle), and the inclusion probability (right) matrix for Karonga data. The cells with orange color are the known edges from the questionnaire structure that is not estimated.

share the same CSMF, since they are both from the second period in time. Figure 3.7 shows the comparison based on the accuracy of the top-cause assignment and CSMF accuracy, both consistent with the patterns in the other experiments. Since we assumed the training data share the same CSMF, we included also a variation of Naive Bayes classifier derived from the training data only. It is worth noting that InterVA performs surprisingly well compared to the proposed model. This is misleading, however, as the experiment uses prior probabilities calculated from a small number of deaths, resulting in the probabilities for rare symptoms being very noisy. InterVA does not take into account the absence of symptoms, and therefore it is not subject to the impact of misspecified priors on rare symptom probabilities. We expect that if the priors are provided based on physician knowledge, or better estimated with a larger dataset as shown in the experiment with the PHMRC data, InterVA's classification rule will not be as effective as the Naive Bayes classifier or the proposed model.

3.7 Discussion

Understanding the correlation structure among mixed data is a challenging task, especially in the presence of missing data, a high dimensional parameter space, and small sample sizes.

Prob	Symptom	Symptom	Partial Corr
1.00	Swelling of the face (puffiness of face)	Both feet or ankles swollen	0.54
0.92	Sores or white patches in the mouth or tongue	Difficulty or pain while swallowing liquids	0.47
0.87	Abdominal distension	Any skin rash (non-measles)	0.32
0.87	Swelling of the face (puffiness of face)	Pale (thinning of blood) or pale palms/soles or nail beds	0.36
0.84	Sores or white patches in the mouth or tongue	Lumps/swellings	0.21
0.83	Age 15-49 years	History of high blood pressure	-0.18
0.82	History of mental confusion	Unconscious for at least 24 hours before death	0.48
0.79	Weight loss	Sores or white patches in the mouth or tongue	0.25
0.77	Fever lasting 2 weeks or more	Weight loss	0.15
0.73	Abdominal distension lasting 2 weeks or more	Any skin rash (non-measles)	0.28
0.72	History of asthma	Unconscious for at least 24 hours before death	0.3
0.68	Any skin problems	Lumps/swellings	0.19
0.67	Fever of any kind	Headache	0.14
0.67	Age 65+ years	History of HIV/AIDS	-0.31
0.66	Diarrhea lasting 4 weeks or more	Became very thin or wasted	0.23
0.63	History of high blood pressure	Paralysis of one side of the body	0.47
0.62	Fever lasting 2 weeks or more	Breathlessness lasting 2 weeks or more	0.09
0.61	Mental confusion for more than 3 months	Unconscious for at least 24 hours before death	0.34
0.61	History of tuberculosis	Productive cough with sputum	0.22
0.59	History of asthma	Mental confusion for more than 3 months	0.3
0.59	Severe abdominal pain lasting 2 weeks or more	Weight loss	0.07
0.57	Breathlessness lasting 2 weeks or more	Both feet or ankles swollen	0.16
0.57	Headache	Stiff or painful neck lasting 1 week or more	0.27
0.56	History of asthma	History of mental confusion	0.24
0.56	Coughed blood	Severe chest pain	0.23

Table 3.3: List of conditional dependent symptom pairs. The non-zero elements in the inverse correlation matrix are selected by the estimated median probability graph.

In this chapter, we propose a method that models the joint distribution of variables of mixed types and leverages marginal prior information. Using both simulation, gold-standard, and physician-coded VA data, we demonstrate that our new framework can significantly improve estimation of the latent correlation structure, graph recovery, and classification performance.

The proposed model can be extended in a few different ways. First, estimating the mixture model using MCMC may suffer from slow mixing when the sampler gets trapped in local modes. This is especially problematic with strong prior information on the extreme values,

i.e. conditional probabilities close to 0 and 1. An alternative approach would be to target the posterior modes directly with deterministic EM-type algorithms (e.g. Ročková and George, 2014), as we will explore more in the next two chapters. Second, symptom reduction in VA analysis is of key interest as a shorter set of symptoms can both reduce the cost as well as improve the quality of data collection. There has been active research on variable selection in Gaussian mixture models (Andrews and McNicholas, 2014), and consequently the proposed framework may also be extended to perform symptom selection in a data-driven way. Third, the model presented in this chapter focuses mostly on binary and continuous data. Extensions to ordinal data are also possible by specifying priors on additional cut-off points. With a normal prior on the log-scale differences between consecutive cutoffs, the proposed model can easily incorporate prior information on marginal probabilities of more than two levels. Finally, in this chapter we only consider the case where all mixtures follow the same correlation matrix. Direct extension to group-specific correlation matrices would be straightforward, but estimating several correlation matrices independently can be problematic where mixture probabilities are highly unbalanced, which we would expect in the context of VAs. Priors on joint distribution of multiple correlation matrix that allow them to borrow information needs to be developed. We will revisit this problem in Chapter 5.

Finally, we would like to draw attention to the fact that using marginal information to guide the modeling of joint associations is strongly related to stratified sampling. If we consider cause of death as an unknown stratification variable, the marginal informative prior helps smooth the potentially noisy estimates of the stratum effects from small samples. Thus the proposed approach might also be extended to improve inference with disproportionate samples, e.g. VA data collected from an HIV study site might have better samples of HIV deaths compared to deaths from other causes.

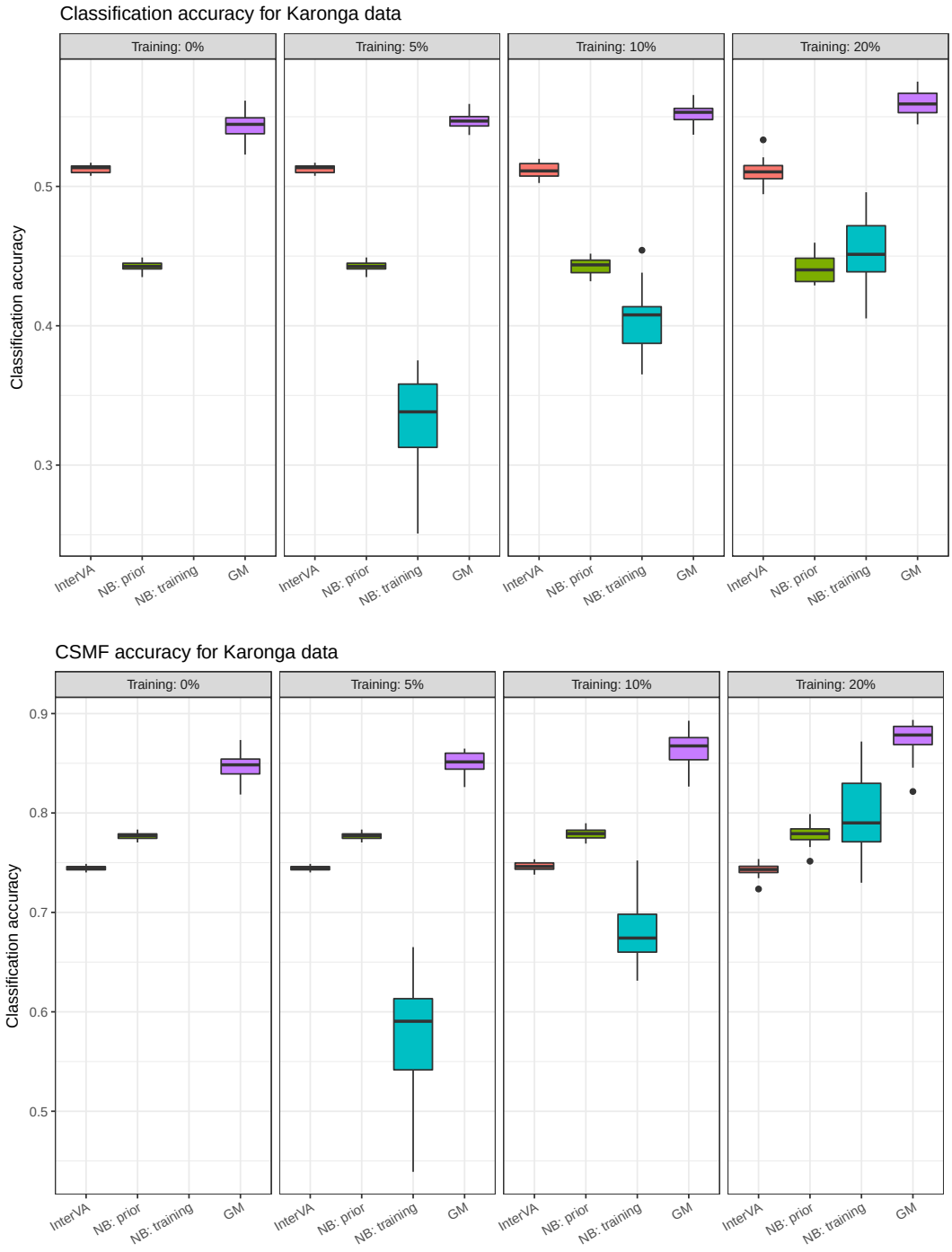


Figure 3.7: Classification accuracy and CSMF accuracy for Karonga physician coded data through cross-validation. The marginal probabilities are calculated with data from 2002 to 2007. The training data and testing data are randomly sampled from the rest of the data from 2008 to 2014. The proposed method consistently outperform Naive Bayes classifier using either prior conditional probabilities or conditional probabilities derived from training data.

Chapter 4

GAUSSIAN GRAPHICAL MODELS USING ECM ALGORITHM[1]

4.1 Introduction

Spike-and-slab priors for the precision matrices provides an useful framework to characterize graphical models. As discussed in the previous chapter, this type of spike-and-slab prior enables a fast block Gibbs sampler that significantly improves the scalability of the model, but such flexibility is at the cost of prior interpretability since the implied marginal distribution of each elements in the precision matrix is intractable due to the positive definite constraint. Wang (2015) provides some heuristics and discussions on prior choices, but it is still not clear how to choose the hyperparameters for practical problems or how those choices affect parameter estimation. In this chapter, we introduce a new algorithm to estimate sparse precision matrices with spike-and-slab priors (Wang, 2015) using a deterministic approach, EMGS (EM graph selection), based on the Expectation Conditional Maximization (ECM) algorithm (Meng and Rubin, 1993). We also show that a stochastic variation of the EMGS approach can be extended to copula graphical model estimation. Our work extends the EM approach to variable selection (EMVS) (Ročková and George, 2014) to general graphical model estimation.

The proposed ECM algorithm is closely connected to frequentist penalized likelihood methods. Similar to the algorithms with concave penalized regularization, such as SCAD (Fan et al., 2009), the spike-and-slab prior used in our method yields sparse inverse covariance

[1]The contents of this chapter are based on the paper "An Expectation Conditional Maximization approach for Gaussian graphical models" (Li and McCormick, 2017).

matrix where large values are estimated with less bias (see Figure 4.3). Similar work has been concurrently developed by Deshpande et al. (2017). The proposed approach in this paper uses a mixture of Gaussian distributions instead of the Laplace distributions used by Deshpande et al. (2017), enabling a simpler closed-form coordinate descent update for the CM-step. Our work also differs in scope, deriving the algorithm for non-Gaussian outcomes and informative priors.

The rest of the paper is organized as follows: In Section 4.2, we describe the spike-and-slab prior we use for the precision matrix. Section 4.3 presents the main ECM framework and algorithms for Gaussian graphical model estimation, and Section 4.4 proposes the extension to the copula graphical model and the modified stochastic ECM algorithm. Then in Section 4.5 we explore the incorporation of informative prior knowledge into the model. We discuss briefly about single model selection in Section 4.6. Section 4.7 examines the performance of our method through numerical simulations. Section 4.8 presents the result from our model using a dataset of aggregated sales of multiple convenient store products. Finally, in Section 4.9 we discuss the limitations of the approach and provide some future directions for improvements.

4.2 Spike-and-slab prior for Gaussian graphical models

First, we review the *Stochastic Search Structure Learning (SSSL)* prior proposed in Wang (2015) for sparse precision matrices. Consider the standard Gaussian graphical model setting, with observed data $\boldsymbol{X} \in \mathbb{R}^{n \times p}$. Each observation follows a multivariate Gaussian distribution, i.e., $\boldsymbol{x}_i \sim \text{Normal}(\boldsymbol{0}, \boldsymbol{\Omega}^{-1})$, where $\boldsymbol{x}_i$ is the i-th row of the $\boldsymbol{X}$, and $\boldsymbol{\Omega}$ is the precision matrix. Given hyperparameter v_0, v_1, and $\pi_{\boldsymbol{\delta}}$, the prior on $\boldsymbol{\Omega}$ is defined as:

$$
\begin{aligned}
p(\boldsymbol{\Omega}|\boldsymbol{\delta}) &= C_{\boldsymbol{\delta}}^{-1} \prod_{j<k} \text{Normal}(\omega_{jk}|0, v_{\delta_{jk}}^2) \prod_{j} \text{Exp}(\omega_{jj}|\lambda/2) \mathbf{1}_{\Omega \in M^+} \\
p(\boldsymbol{\delta}|\pi_{\boldsymbol{\delta}}) &\propto C_{\boldsymbol{\delta}} \prod_{j<k} \pi_{\boldsymbol{\delta}}^{\delta_{jk}} (1-\pi_{\boldsymbol{\delta}})^{1-\delta_{jk}}
\end{aligned}
$$

where δ_{jk} are latent indicator variables, and $\pi_{\boldsymbol{\delta}}$ is the prior sparsity parameter. The $C_{\boldsymbol{\delta}}$ term is the normalizing constant that ensures the integration of $p(\boldsymbol{\Omega}|\boldsymbol{\delta})$ on M^+, the space of positive definite matrices, is one. This formulation places a Gaussian mixture prior on the off-diagonal elements of $\boldsymbol{\Omega}$, similar to the spike-and-slab prior used in the Bayesian variable selection literature. By setting $v_1 \gg v_0$, the mixture prior imposes a different strength of shrinkage for elements drawn from the "slab" (v_1) and "spike" (v_0) respectively. This representation allows us to shrink elements in $\boldsymbol{\Omega}$ to 0 if they are small in scale, while not biasing the large elements significantly. The spike-and-slab prior has been extensively studied in the regression setting (e.g., Hans et al., 2007; Ishwaran and Rao, 2005, 2011; Hahn and Carvalho, 2015), but is less commonly used for graphical model estimation.

Due to the positive definiteness constraint, the normalizing constant for this prior distribution of $\boldsymbol{\Omega}$ is intractable, but we can glean insights about this prior distribution by simulating from the prior using the MCMC steps described in Wang (2015). Figure 4.1 shows the marginal prior distribution on the induced correlation coefficients, i.e., off-diagonal elements of the induced $\boldsymbol{R}$, under a complete graph. When the marginal shrinkage parameter v_1 is large, the marginal prior on $\boldsymbol{R}$ and $\boldsymbol{R}^{-1}$ induced by this spike-and-slab distribution becomes very similar to that of the marginal uniform prior. This is not surprising as it can be seen directly from the marginal distribution on the matrix elements of $\boldsymbol{\Omega}$ as well. For the j-th column of $\boldsymbol{\Omega}$, the spike-and-slab prior induces the conditional prior distribution on $\boldsymbol{\omega}_{[j,-j]}$ and the Schur complement $\omega_{j|-j} = \omega_{jj} - \boldsymbol{\omega}_{[j,-j]}^T \boldsymbol{\Omega}_{[-j,-j]}^{-1} \boldsymbol{\omega}_{[j,-j]}$ to be

$$\begin{aligned} \boldsymbol{\omega}_{[j,-j]}|\boldsymbol{\Omega}_{[-j,-j]} &\sim \text{Normal}(\mathbf{0}, (\lambda \boldsymbol{\Omega}_{[-j,-j]}^{-1} + \text{diag}(\boldsymbol{V}_{[j,-j]}^{-1}))^{-1}) \\ \omega_{j|-j}|\boldsymbol{\Omega}_{[-j,-j]} &\sim \text{Gamma}\left(1, \frac{\lambda}{2}\right) \end{aligned}$$

where $\boldsymbol{V} = \{v_{\delta_{jk}}^2\}_{jk}$ is the matrix of the "penalization" parameters determined by v_0, v_1 and a

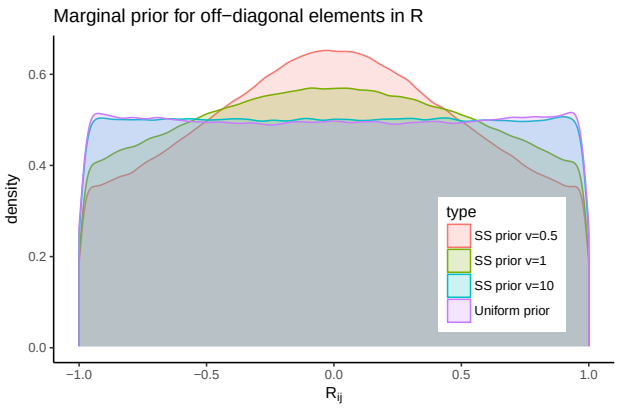


Figure 4.1: Different marginal priors on $\boldsymbol{R}$ induced by the spike-and-slab prior on $\boldsymbol{\Omega}$ with $p = 50$ and $\lambda = 2$. A complete graph, i.e. $v_0 = v_1$ is assumed. The densities are derived from sampling $2,000$ draws using MCMC from the prior distribution after $2,000$ iterations of burn-in.

given graph. This resembles the conditional prior distribution under the Wishart distribution in the previous section, i.e. when $\boldsymbol{\Omega} \sim \text{Wishart}(p+1, \boldsymbol{I}_p)$, the marginal prior distribution for the same quantities are

$$\begin{aligned} \boldsymbol{\omega}_{[j,-j]} | \boldsymbol{\Omega}_{[-j,-j]} &\sim \text{Normal}(\boldsymbol{0}, \boldsymbol{\Omega}_{[-j,-j]}) \\ \omega_{j|-j} | \boldsymbol{\Omega}_{[-j,-j]} &\sim \text{Gamma}\left(1, \frac{1}{2}\right) \end{aligned}$$

The Wishart prior induced on $\boldsymbol{\omega}_{[j,-j]}$ is the limiting case in the spike-and-slab prior as $v_0 = v_1 \to \infty$ and $\lambda = 1$. The spike-and-slab prior can be viewed, therefore, as a flexible prior in the middle ground between the Wishart prior and G-Wishart prior with exact zeros in the off-diagonal elements, while sharing both the easy computational properties of the former and the graph interpretation of the latter.

The main limitation of this spike-and-slab prior, however, also stems from its flexibility. Parameter estimation can be sensitive to the prior choices of the marginal variances. And unlike in variable selection problems, information on the scale of the elements in the precision

matrix cannot be easily solicited from domain knowledge. As shown in Wang (2015), there is no analytical relationship between the prior sparsity parameter $\pi_{\boldsymbol{\delta}}$ and the induced sparsity from the joint distribution. This complexity results from the positive definiteness constraint on the precision matrix. Thus even if the sparsity of the precision matrix is known before fitting the model, additional heuristics and explorations are required to properly select the prior $\pi_{\boldsymbol{\delta}}$. Similarly, the induced marginal distribution of the elements in $\boldsymbol{\Omega}$ is intractable as well. Figure 4.2 shows several induced marginal distributions for elements of $\boldsymbol{\Omega}$ when v_0 varies and all other parameters are held constant. This figure illustrates the difference between the specified marginal priors and the induced marginal priors. Thus although the fully Gibbs sampler is attractive for high dimensional problems, in practice researchers will usually need to evaluate the model fit under multiple prior choices, adding substantially to the computational burden.

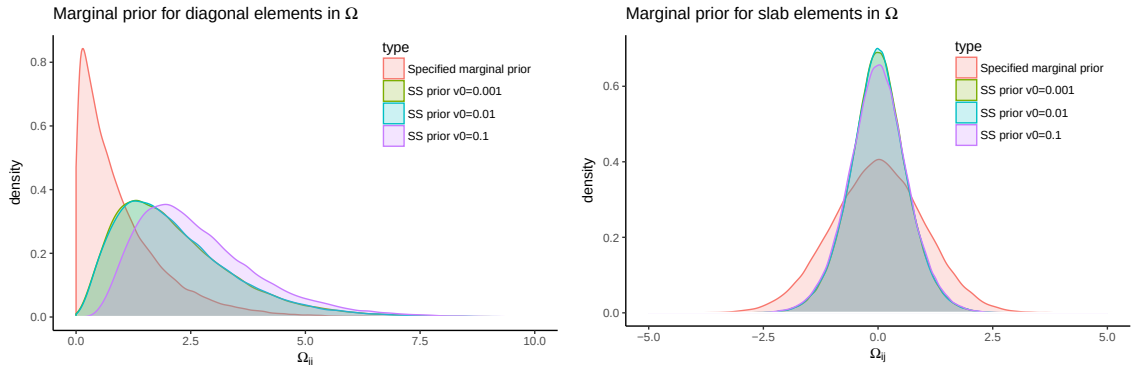


Figure 4.2: Comparison of specified marginal prior distribution and induced marginal prior distributions for $\boldsymbol{\Omega}$ with $p = 50$, $\lambda = 2$, $v_1 = 1$ and varying v_0 values. The underlying graph is fixed to be an AR(2) graph. Left: diagonal elements $\boldsymbol{\Omega}_{ii}$. Right: Non-zero off-diagonal elements (slab) $\boldsymbol{\Omega}_{ij}, i \neq j$. The densities are derived from sampling $2,000$ draws using MCMC from the prior distribution after $2,000$ iterations of burn-in.

4.3 ECM algorithm for graph selection

Consider spike-and-slab priors on $\boldsymbol{\Omega}$ as described in the previous section, the complete-data posterior distribution ccan be expressed as

$$p(\boldsymbol{\Omega}, \boldsymbol{\delta}, \pi_{\boldsymbol{\delta}}|\boldsymbol{X}) = p(\boldsymbol{X}|\boldsymbol{\Omega})p(\boldsymbol{\Omega}|\boldsymbol{\delta}, v_0, v_1, \lambda)p(\boldsymbol{\delta}|\pi_{\boldsymbol{\delta}})p(\pi_{\boldsymbol{\delta}}|a, b)$$

The block Gibbs algorithm proposed in Wang (2015) reduces the problem to iteratively sampling from $(p-1)$-dimensional multivariate Gaussian distributions for each column of $\boldsymbol{\Omega}$, which can be computationally expansive for large p or if the sampling needs to be repeated for multiple prior setups, which is oftern the case in practice as discussed before. Inspired by the EM approach for variable selection proposed in Ročková and George (2014), we propose a EMGS algorithm to identify the posterior mode of $p(\boldsymbol{\Omega}, \pi_{\boldsymbol{\delta}}|\boldsymbol{X})$ directly without the full stochastic search. We iteratively maximize the following objective function

$$\begin{aligned} Q(\boldsymbol{\Omega}, \pi_{\boldsymbol{\delta}}|\boldsymbol{\Omega}^{(k)}, \pi_{\boldsymbol{\delta}}^{(k)}) &= E_{\boldsymbol{\delta}|\boldsymbol{\Omega}^{(k)}, \pi_{\boldsymbol{\delta}}^{(k)}, \boldsymbol{X}}(\log p(\boldsymbol{\Omega}, \boldsymbol{\delta}, \pi_{\boldsymbol{\delta}}|\boldsymbol{X})|\boldsymbol{\Omega}^{(k)}, \pi_{\boldsymbol{\delta}}^{(k)}, \boldsymbol{X}) \\ &= \text{constant} + \frac{n}{2}\log|\boldsymbol{\Omega}| - \frac{1}{2}tr(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\Omega}) \\ &\quad -\frac{1}{2}\sum_{j<k}\omega_{jk}^2 E_{\cdot|\cdot}[\frac{1}{v_0^2(1-\delta_{jk}) + v_1^2\delta_{jk}}] - \frac{\lambda}{2}\sum_j \omega_{jj} \\ &\quad +\sum_{j<k}\log(\frac{\pi_{\boldsymbol{\delta}}}{1-\pi_{\boldsymbol{\delta}}}E_{\cdot|\cdot}[\delta_{jk}]) + \frac{p(p-1)}{2}\log(1-\pi_{\boldsymbol{\delta}}) \\ &\quad +(a-1)\log(\pi_{\boldsymbol{\delta}}) + (b-1)\log(1-\pi_{\boldsymbol{\delta}}) \end{aligned}$$

where $E_{\cdot|\cdot}[\cdot]$ denotes $E_{\boldsymbol{\delta}|\boldsymbol{\Omega}^{(k)}, \pi_{\boldsymbol{\delta}}^{(k)}, \boldsymbol{X}}[\cdot]$. This objective function can be easily estimated using ECM algorithm. We present the E-step and CM-step details in the next two subsections and then compare the algorithm with the coordinate ascent algorithm for solving graphical lasso problem in Section 4.3.3.

4.3.1 The E-step

The E-step computes the conditional expectations $E_{\boldsymbol{\delta}|\boldsymbol{\Omega}^{(k)},\pi_{\boldsymbol{\delta}}^{(k)},\boldsymbol{X}}[\delta_{jk}]$ and $E_{\boldsymbol{\delta}|\boldsymbol{\Omega}^{(k)},\pi_{\boldsymbol{\delta}}^{(k)},\boldsymbol{X}}[\frac{1}{v_0^2(1-\delta_{jk})+v_1^2\delta_{jk}}]$. This proceeds in the similar fashion as the standard EMVS,

$$E_{\boldsymbol{\delta}_{jk}|\boldsymbol{\Omega}^{(k)},\pi_{\boldsymbol{\delta}}^{(k)},\boldsymbol{X}}[\delta_{jk}] = p_{jk}^* \equiv \frac{a_{jk}}{a_{jk}+b_{jk}},$$

where $a_{jk} = p(\omega_{jk}|\delta_{jk}=1)\pi_{\boldsymbol{\delta}}^{(k)}$ and $b_{jk} = p(\omega_{jk}|\delta_{jk}=0)(1-\pi_{\boldsymbol{\delta}}^{(k)})$, and

$$E_{\boldsymbol{\delta}|\boldsymbol{\Omega}^{(k)},\pi_{\boldsymbol{\delta}}^{(k)},\boldsymbol{X}}[\frac{1}{v_0^2(1-\delta_{jk})+v_1^2\delta_{jk}}] = \frac{1-p_{jk}^*}{v_0^2} + \frac{p_{jk}^*}{v_1^2} \equiv d_{jk}^*.$$

4.3.2 The CM-step

After the E-step is performed, the CM-step performs the maximization of $(\boldsymbol{\Omega}, \pi_{\boldsymbol{\delta}})$ in a coordinate ascent fashion. First, the maximization of $\pi_{\boldsymbol{\delta}}$ has the close-form solution

$$\pi_{\boldsymbol{\delta}}^{(k+1)} = (a + \sum_{j<k}\delta_{jk} - 1)/(a+b+p(p-1)/2-2).$$

The joint maximization of $\boldsymbol{\Omega}$ has no closed-form solution, but thanks to an observation made by Wang (2015), if we denote

$$\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\omega}_{12} \\ \boldsymbol{\omega}_{12}^T & \omega_{22} \end{pmatrix} \quad \boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} \boldsymbol{S}_{11} & \boldsymbol{s}_{12} \\ \boldsymbol{s}_{12}^T & s_{22} \end{pmatrix},$$

the conditional distribution of the last column satisfies

$$\boldsymbol{\omega}_{12} \sim \text{Normal}(-\boldsymbol{C}\boldsymbol{s}_{12}, \boldsymbol{C}), \quad \boldsymbol{C} = ((s_{22}+\lambda)\boldsymbol{\Omega}^{-1} + \text{diag}(v_{\boldsymbol{\delta}_{12}}))^{-1},$$

where $v_{\boldsymbol{\delta}_{12}}$ are the inclusion indicators for ω_{12} and

$$\omega_{22} - \boldsymbol{\omega}_{12}^T \boldsymbol{\Omega}_{11}^{-1} \boldsymbol{\omega}_{12} \sim \text{Gamma}(1 + \frac{n}{2}, \frac{\lambda + s_{22}}{2}).$$

This enables us to perform conditional maximization (Meng and Rubin, 1993) for the last column holding the rest of $\boldsymbol{\Omega}$ fixed. That is, starting with $\boldsymbol{\Omega}^{(k+1)} = \boldsymbol{\Omega}^{(k)}$, we iteratively permute each column to the last and update it with

$$\boldsymbol{\omega}_{12}^{(k+1)} = ((s_{22} + \lambda)(\boldsymbol{\Omega}_{11}^{(k+1)})^{-1} + \text{diag}(d_{jk}^*))^{-1} \boldsymbol{s}_{12}$$

and

$$\omega_{22}^{(k+1)} = (\boldsymbol{\omega}_{12}^{(k+1)})^T (\boldsymbol{\Omega}_{11}^{(k+1)})^{-1} \boldsymbol{\omega}_{12}^{(k+1)} + \frac{n}{\lambda + s_{22}}.$$

Finally, be iterating between the E-step and the CM-steps until convergence, we obtain our estimator of the posterior mode $\hat{\boldsymbol{\Omega}}$ and $\hat{\pi}_{\boldsymbol{\delta}}$.

4.3.3 Connection to the graphical lasso

This column-wise update resembles the penalized likelihood representation in frequentist setting. In the graphical lasso algorithm (Mazumder and Hastie, 2012b) for example, the goal is to minimize the l_1-penalized negative log-likelihood:

$$f(\boldsymbol{\Omega}) = -\log|\boldsymbol{\Omega}| + \text{tr}(\boldsymbol{S}\boldsymbol{\Omega}) + \|\boldsymbol{\Omega}\|_1,$$

which can be solved via a block coordinate descent that iteratively solves the lasso problem

$$\boldsymbol{\omega}_{12} = \underset{\boldsymbol{\alpha} \in R^{m-1}}{\text{argmin}} \, \boldsymbol{\alpha}^T \Omega_{11}^{-1} \boldsymbol{\alpha} + \boldsymbol{\alpha}^T \boldsymbol{s}_{12} + \lambda \|\boldsymbol{\alpha}\|_1.$$

The updates at each iteration in the EMGS framework solve the optimization problem for $\boldsymbol{\omega}_{12}$ under an adaptive ridge penalty

$$\boldsymbol{\omega}_{12} = \underset{\boldsymbol{\alpha} \in R^{m-1}}{\operatorname{argmin}} \boldsymbol{\alpha}^T \boldsymbol{\Omega}_{11}^{-1} \boldsymbol{\alpha} + \boldsymbol{\alpha}^T \boldsymbol{s}_{12} + \sum_{j=1}^{m-1} \tilde{d}_j^* \alpha_j^2.$$

The penalty parameters $\tilde{d}_j^*$ are the corresponding d_{jk}^* estimated from the E-step and are informed by data. That is, instead of choosing a fixed penalty parameter for all precision matrix elements, the EMGS approach learns the element-wise penalization parameter at each iteration based on the magnitude of the current estimated $\boldsymbol{\Omega}$ and the hyperpriors placed on θ. Thus, as long as the signal from data is not too weak, the EMGS procedure can estimate large elements in the precision matrix with much lower bias than graphical lasso, as the adaptive penalties associated with large $\boldsymbol{\omega}_{jk}$ are small. To illustrate the diminished bias, we fit the EMGS algorithm to a simple simulated example, where $n = 100$, $p = 10$ and $\boldsymbol{\Omega}$ is constructed by $\omega_{jj} = 1$, and $\omega_{jk} = 0.5$ if $|j - k| = 1$. We fix $v_1 = 1000$ and compare the regularization path with various v_0 values with graphical lasso, as shown in Figure 4.3. The EMGS approach identifies the correct non-zero elements quickly and estimates the partial correlations correctly around 0.5, while graphical lasso shrinks the non-zero partial correlations downwards significantly when not many edges are selected.

4.4 ECM algorithm for copula graphical models

In this section, we extend the framework to non-Gaussian data with Gaussian copulas (Nelsen, 1999). Denote the observed data $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, and each of the p variables could be either continuous, ordinal, or binary. We model each observation as following a Gaussian copula model, i.e., there exists a set of monotonically increasing transformations $f = \{f_1, ..., f_p\}$ such that $\boldsymbol{Z} = f(\boldsymbol{X}) \sim \text{Normal}(\boldsymbol{0}, \boldsymbol{R})$, where $\boldsymbol{R}$ is a correlation matrix. Following the same

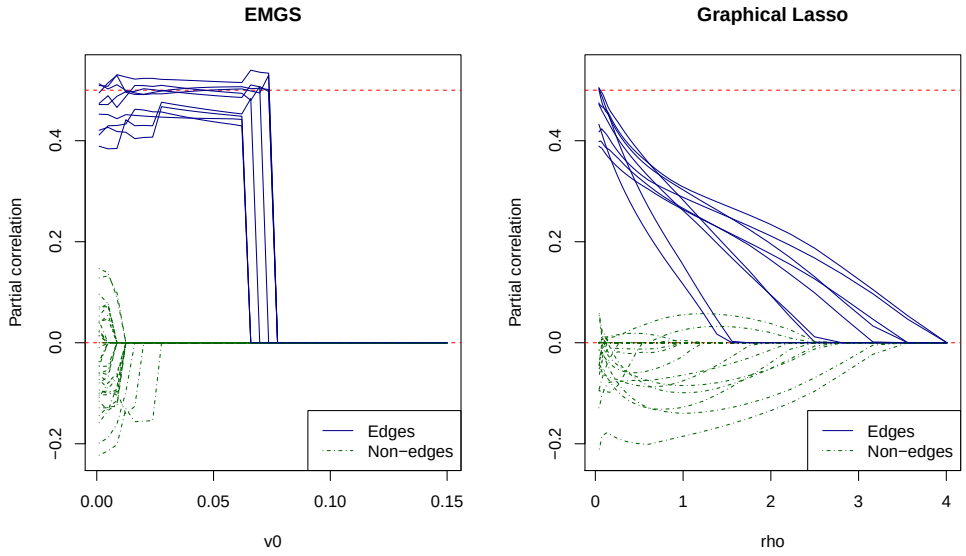


Figure 4.3: Comparing graph selection path using EMGS and graphical lasso, on a 10-node graph. The red dashed line at 0.5 is the true value for the non-zero partial correlations. The blue solid lines represent the non-zero off-diagonal elements and the green dashed lines represent the zero off-diagonal elements.

setup as before, we let $\boldsymbol{R}$ be the induced correlation matrix from $\boldsymbol{\Omega}$ with the spike-and-slab prior defined as before, i.e.,

$$\boldsymbol{R}_{[j,k]} = \boldsymbol{\Omega}^{-1}_{[j,k]} / \sqrt{\boldsymbol{\Omega}^{-1}_{[j,j]} \boldsymbol{\Omega}^{-1}_{[k,k]}}.$$

The explicit form of f is typically unknown, thus we impose no restrictions on the class of marginal transformations. Instead, we follow the extended rank likelihood method proposed in Hoff (2007), decomposing the complete data likelihood into

$$p(\boldsymbol{X}|\boldsymbol{R}, f) = Pr(\boldsymbol{Z} \in \boldsymbol{S}|\boldsymbol{R}) p(\boldsymbol{X}|\boldsymbol{Z} \in \boldsymbol{S}, \boldsymbol{R}, f), \tag{4.1}$$

where $\boldsymbol{S}$ is the support of $\boldsymbol{Z}$ induced by the ranking of $\boldsymbol{X}$ defined by

$$\boldsymbol{S}_{ij} = [\max\{z_{i'j'} : x_{i'j'} < x_{ij}\}, \min\{z_{i'j'} : x_{i'j'} > x_{ij}\}].$$

Since our goal is to recover the structure in $\boldsymbol{R}$, or equivalently $\boldsymbol{\Omega}$, we can estimate the parameters using only the first part of (4.1) without estimating the nuisance parameter f. Moreover, since the latent Gaussian variable $\boldsymbol{Z}$ is constructed to be centered at $\mathbf{0}$, the rank likelihood remains unchanged when multiplying columns of $\boldsymbol{X}$ by any constant. Thus, inference could be performed without restricting $\boldsymbol{R}$ to be an correlation matrix (Hoff, 2007). In this way, the target function to maximize is the extended rank likelihood function:

$$p(\boldsymbol{\Omega}, \boldsymbol{\delta}, \pi_{\boldsymbol{\delta}}, \boldsymbol{Z}|\boldsymbol{X}) = p(\boldsymbol{Z} \in \boldsymbol{S}|\boldsymbol{\Omega}, \boldsymbol{S})p(\boldsymbol{\Omega}|\boldsymbol{\delta})p(\boldsymbol{\delta}|\pi_{\boldsymbol{\delta}}).$$

This is immediately analogous to the EMGS framework with latent Gaussian variable $\boldsymbol{Z}$ as additional missing data. That is, we maximize the objective function defined as

$$\begin{aligned} Q(\boldsymbol{\Omega}, \pi_{\boldsymbol{\delta}}|\boldsymbol{\Omega}^{(k)}, \pi_{\boldsymbol{\delta}}^{(k)}) &= E_{\boldsymbol{\delta}, \boldsymbol{Z}|\boldsymbol{\Omega}^{(k)}, \pi_{\boldsymbol{\delta}}^{(k)}, \boldsymbol{X}}(\log p(\boldsymbol{\Omega}, \boldsymbol{\delta}, \pi_{\boldsymbol{\delta}}, \boldsymbol{Z}|\boldsymbol{X})|\boldsymbol{\Omega}^{(k)}, \pi_{\boldsymbol{\delta}}^{(k)}, \boldsymbol{X}) \\ &= \text{constant} + Q_1 - \frac{1}{2}\sum_{j<k} \omega_{jk}^2 E_{\cdot|\cdot}[\frac{1}{v_0^2(1-\delta_{jk}) + v_1^2\delta_{jk}}] - \frac{\lambda}{2}\sum_i \omega_{ii} \\ &\quad + \sum_{j<k} \log(\frac{\pi_{\boldsymbol{\delta}}}{1-\pi_{\boldsymbol{\delta}}} E_{\cdot|\cdot}[\delta_{jk}]) + \frac{p(p-1)}{2}\log(1-\pi_{\boldsymbol{\delta}}) \\ &\quad + (a-1)\log(\pi_{\boldsymbol{\delta}}) + (b-1)\log(1-\pi_{\boldsymbol{\delta}}) \end{aligned}$$

where $E_{\cdot|\cdot}[\cdot]$ denotes $E_{\boldsymbol{\delta}, \boldsymbol{Z}|\boldsymbol{\Omega}^{(k)}, \pi_{\boldsymbol{\delta}}^{(k)}, \boldsymbol{X}}[\cdot]$, and the only term different from the standard EMGS objective function is

$$\begin{aligned} Q_1 &= E_{\boldsymbol{Z}|\boldsymbol{\Omega}^{(k)}, \pi_{\boldsymbol{\delta}}^{(k)}, \boldsymbol{X}}(\log p(\boldsymbol{Z}|\boldsymbol{\Omega}, \boldsymbol{S})) \\ &= \text{constant} + \frac{n}{2}\log|\boldsymbol{\Omega}| - \frac{1}{2}E_{\boldsymbol{Z}|\boldsymbol{\Omega}^{(k)}, \boldsymbol{X}}[tr(\boldsymbol{Z}^T\boldsymbol{Z}\boldsymbol{\Omega})]. \end{aligned}$$

Exact computation for this expectation is intractable as $\boldsymbol{Z}|\boldsymbol{X}$ is a Gaussian random matrix where each row is conditionally Gaussian and the within column ranks are fixed by $\boldsymbol{S}$. Alter-

natively, posterior samples of $\boldsymbol{Z}$ are easy to obtain from the conditional truncated Gaussian distribution (Hoff, 2007), so we can adopt stochastic variants of the EM algorithm (Wei and Tanner, 1990; Delyon et al., 1999; Nielsen, 2000; Levine and Casella, 2001). We present one such algorithm in the subsequent subsection.

4.4.1 The SAE-step and the CM-step

Among the many variations of the EM with stochastic approximation, we discuss estimation steps using stochastic approximation EM (SAEM) algorithm (Delyon et al., 1999). SAEM calculates the E-step at each iteration as a weighted average of the current objective function and new stochastic samples using a decreasing sequence of weights for the stochastic averages, in a similar fashion as simulated annealing. In the stochastic E-step, we compute an additional term $Q(\boldsymbol{\Omega}^{(k)}) = E_{\boldsymbol{Z}|\boldsymbol{\Omega}^{(k)},\boldsymbol{X}}[\boldsymbol{Z}^T\boldsymbol{Z}]$ as

$$Q(\boldsymbol{\Omega}^{(k)}) = (1 - t_k)Q(\boldsymbol{\Omega}^{(k)}) + \frac{t_k}{B_k}\sum_{b=1}^{B_k} \boldsymbol{Z}_{(b)}^T\boldsymbol{Z}_{(b)}$$

where t_k is an decreasing step-size sequence such that $\sum t_k = \infty$, $\sum t_k^2 < \infty$,and B_k is the number of stochastic samples drawn at each iteration. The rank constrained Gaussian variables can be drawn using the same procedure described in Hoff (2007). The CM-step then proceeds as before, except that the empirical cross-product matrix $\boldsymbol{S}$ is replaced by its expectation $Q(\boldsymbol{\Omega}^k)$. For the numerical examples in this paper, we set fixed B_k and $t_k = 1/k$. Other weighting schemes could also be explored and may yield different rate of convergence.

4.5 Informative priors

The exchangeable beta-binomial prior discussed so far assumes no prior structure on $\boldsymbol{\Omega}$ and prior sparsity controlled by a single parameter for all off-diagonal elements. For many

problems in practice, informative priors may exist for pairwise interactions of the variables. For example, Peterson et al. (2013) infers cellular metabolic networks based on prior information in the form of reference network structures. Bu and Lederer (2017) improve estimation of brain connectivity network by incorporating the distance between regions of the brain. In problems with small sample sizes, such prior information can help algorithms identify the high probability edges more quickly and provide more interpretable model. To extend this framework, consider a situation where certain groupings exist among variables. For example, when the variables represent log sales of p products on the market, one might expect that the products within the same brand are more likely to be more strongly correlated. If we define a fixed index function $g_j \in \{1, ..., G\}, j \in \{1, ..., p\}$, where G denotes the total number of groups, then we propose the modification of the prior to be

$$\begin{aligned} p(\mathbf{\Omega}|\boldsymbol{\delta}) &= C_{\boldsymbol{\delta}}^{-1} \prod_{j<k} \text{Normal}(\omega_{jk}|0, \frac{v_{\delta_{jk}}^2}{\tau_{g_j g_k}}) \prod_j \text{Exp}(\omega_{jj}|\lambda/2) \mathbf{1}_{\Omega \in M^+} \\ p(\boldsymbol{\delta}|\pi_{\boldsymbol{\delta}}) &\propto C_{\boldsymbol{\delta}} \prod_{j<k} \pi_{\boldsymbol{\delta}}^{\delta_{jk}} (1-\pi_{\boldsymbol{\delta}})^{1-\delta_{jk}} \\ p(\boldsymbol{\tau}) &= \prod_{g<g'} \text{Gamma}(a_\tau, b_\tau) \end{aligned}$$

The block-wise rescaling parameter $\tau_{g_j g_k}$ of the variance parameter allows us to model within- and between-block elements of $\mathbf{\Omega}$ adaptively with different scales. This is particularly useful in applications where block dependence structures have different strengths. Take the example of sales of products for example. Products within the same brand or category are more likely to be conditional dependent, yet the within group sparsity and the scale of the off-diagonal elements may differ for different brands. The ECM algorithm discussed above only requires minor modifications to include the additional scale parameter so that the penalties for each block are allowed to vary (e.g., Ishwaran and Rao, 2003; Wakefield et al., 2010). The new objective function could be similarly estimated with ECM algorithm by including this

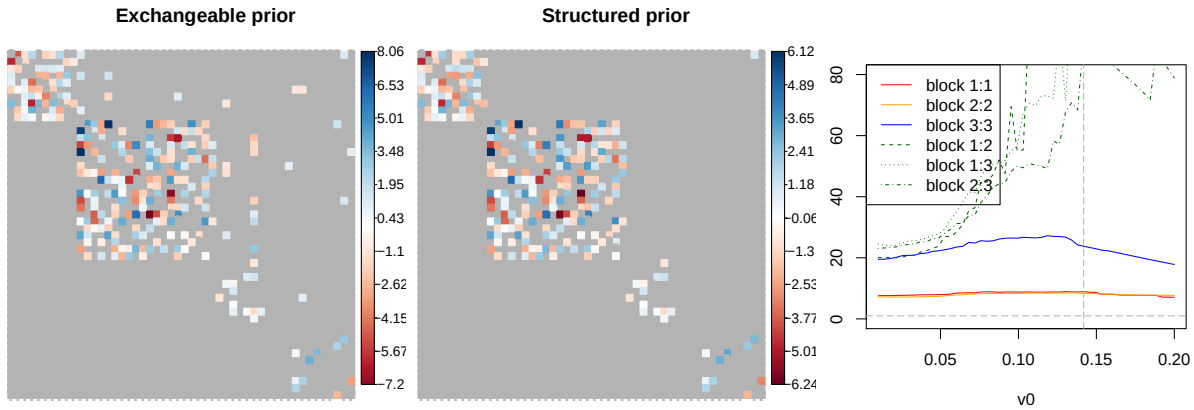
additional update in the CM-step:

$$\tau_{gg'}^{(k+1)} = \frac{a_\tau - 1 + \frac{1}{2}\sum_{j<k} \mathbf{1}_{j,k,g,g'}}{b_\tau + \frac{1}{2}\sum_{j<k} \omega_{jk}^2 d_{jk}^* \mathbf{1}_{j,k,g,g'}},$$

where $\mathbf{1}_{j,k,g,g'} = 1$ if $g_j = g, g_k = g'$, or $g_j = g', g_k = g$. To illustrate the behavior of this block rescaled prior, we simulate data with $n = 100$, $p = 20$, with the precision matrix to be block diagonal with block size $10, 20$, and 20 each. We then simulate the three block sub-matrices of $\mathbf{\Omega}$ to correspond to the following three graphs described in Section 4.7: random graph with sparsity 0.8, random grpah with sparsity 0.5, and two-cluster graph with sparsity 0.5. Figure 4.4 shows the effect of the structured prior. It can be seen that the estimated $\hat{\tau}_{gg'}$ are much larger where $g \neq g'$, which leads to stronger shrinkage effects for between cluster cells. $\hat{\tau}_{gg'}$ is also larger for the last block, which has ω_{jk} relatively smaller in scale. Accordingly the resulting graph using the structured prior shows fewer false positives for the off-diagonal blocks, and better discovery of the true positives within blocks.

4.6 Posterior summary of the ECM output

Finding the posterior mode with the ECM algorithm is computationally very fast. Thus in practice, we can fix v_1 to be a large constant and vary the choice of v_0 to reflect different levels of shrinkage on the off-diagonal elements of $\mathbf{\Omega}$ that are close to 0. Intuitively, a larger v_0 increases the probability of small parameters being drawn form the spike distribution and thus leads to sparse models. By fitting a sequence of v_0, we can create regularization plots, e.g., Figure 4.3, similar to that used in penalized regression literature to visually examine the influence of the prior choices. At any fixed (v_0, v_1), determining the final model could be achieved by thresholding the off-diagonal elements, ω_{jk}, as the posterior inclusion probability p_{jk}^* conditional on ω_{jk} is a monotone function of ω_{jk}. Choosing a single tuning parameter v_0 is possible with standard model selection criterion, such as AIC (Akaike, 1998), BIC (Schwarz,



.

Figure 4.4: Comparing the estimated and true precision matrix using with exchangeable prior and structured prior for block-wise rescaling. In each plot of the precision matrix comparison, the upper triangle shows the estimated matrix and the lower triangle shows the true precision matrix. All the tunning parameters or selected so that the estimated graph has the closest number of edges compared to the true graph. The last line plot shows the change of $\hat{\tau}_{gg'}$ over different choices of v_0. The blocks are labeled 1 to 3 from top left to bottom right.

1978), StARS (Liu et al., 2010), etc., or K-fold cross validation using the average log-likelihood of the validation sets.

4.7 Simulation

We follow a similar simulation setup to Mohammadi et al. (2017) with different graph structures. We compare the performance of our method with graphical lasso for Gaussian data and graphical lasso with nonparanormal transformation (Liu et al., 2009) for non-Gaussian data. We consider sparsity patterns:

- AR(1): A graph with $\sigma_{jk} = 0.7^{|j-k|}$.
- AR(2): A graph with $\omega_{jj} = 1$, $\omega_{j,j-1} = \omega_{j-1,j} = 0.5$, and $\omega_{j,j-2} = \omega_{j-2,j} = 0.25$, and $\omega_{jk} = 0$ otherwise.
- Random: A graph in which the edge set E is randomly generated from independent

Bernoulli distributions with probability 0.2 and the corresponding precision matrix is generated from $\mathbf{\Omega} \sim W_G(3, I_p)$.

- Cluster: A graph in which the number of clusters is $\max\{2, [p/20]\}$. Each cluster has the same structure as a random graph. The corresponding precision matrix is generated from $\mathbf{\Omega} \sim W_G(3, I_p)$.

For each sparsity pattern, we let the sample size $n \in \{50, 100, 500\}$, the dimension $p = 50$, and we generate both Gaussian and mixed data. For mixed data, the variables are randomly chosen to be continuous non-Gaussian, ordinal, or binary. We simulate graphs with the `R` package `BDgraph` (Mohammadi and Wit, 2015b). The graphical lasso estimation and nonparanormal transformation are implemented with the `R` package `huge` (Zhao et al., 2012).

For each generated graph, we fit our ECM algorithm with a sequence of 40 increasing v_0's, and select the graph at each v_0 using the median model. For a fair comparison, we select the tuning parameters so that the estimated graph has the closest number of edges to the true graph. We first evaluate the performance of the two methods in terms of the matrix error of the standardized precision matrix from the truth in terms of the matrix element-wise maximum norm, spectral norm and Frobenius norm. We then compare the graph selection performance using the F_1-score defined as $F_1 = \frac{2TP}{2TP+FP+FN}$. The results are summarized in Table 4.1. In almost all the cases of our study, we observe a smaller bias in estimated precision matrix, as well as better graph selection performance.

4.8 Analysis of sales data

In this section we consider the graph estimation for the *Breakfast at the Frat* dataset from the public *Dunnhumby* repository[2]. The dataset contains sales for the top five products from each of the top three brands within four selected categories: mouthwash, pretzels, frozen pizza, and

[2] https://www.dunnhumby.com/sourcefiles

boxed cereal, gathered from a sample of stores over 156 weeks. We aggregated the dataset into weekly sales of a total 55 products. We first performed a log transformation on the raw sales and carried out a autoregression on the log sales of previous week to remove autocorrelation across time. We then fit the EMGS with both the beta-binomial prior described in Section 4.3 and the group-wise rescaled prior on the standardized residuals from the autoregression. For comparison, we also applied graphical lasso. In all cases, we select the tuning parameter using 5-fold cross validation. The results are shown in Figure 4.5. Graphical lasso estimates many edges with small partial correlations, while EMGS is able to select sparse graphs while allowing small partial correlations to exist from the spike distribution. We also observe more edges within product categories from the EMGS estimator. For $p = 55$, convergence of the EMGS can be achieved within a few seconds for a given v_0 and v_1 on a laptop computer. With additional optimization of the implementation, we expect the computing time can be further reduced significantly.

4.9 Discussion

In this chapter, we propose a deterministic approach for graphical model estimation that builds upon the recently proposed class of spike-and-slab prior for precision matrix and the new advances in Bayesian variable selection. We also extend the algorithm to copula graphical models. The computational speed of the EGMS comes at the price of two potential limitations. First, characterization of posterior uncertainty is nontrivial due to the deterministic nature of the algorithm. As in Ročková and George (2014), one may choose to fit the full Bayesian locally from the posterior mode obtained by the ECM procedure, though this may still be challening in high-dimensional problems. Another limitation is that like the EM algorithm, ECM algorithm also converges only to local modes, thus the precision matrix initialization is critical. In this paper, we used the same initialization as the P-Glasso algorithm described

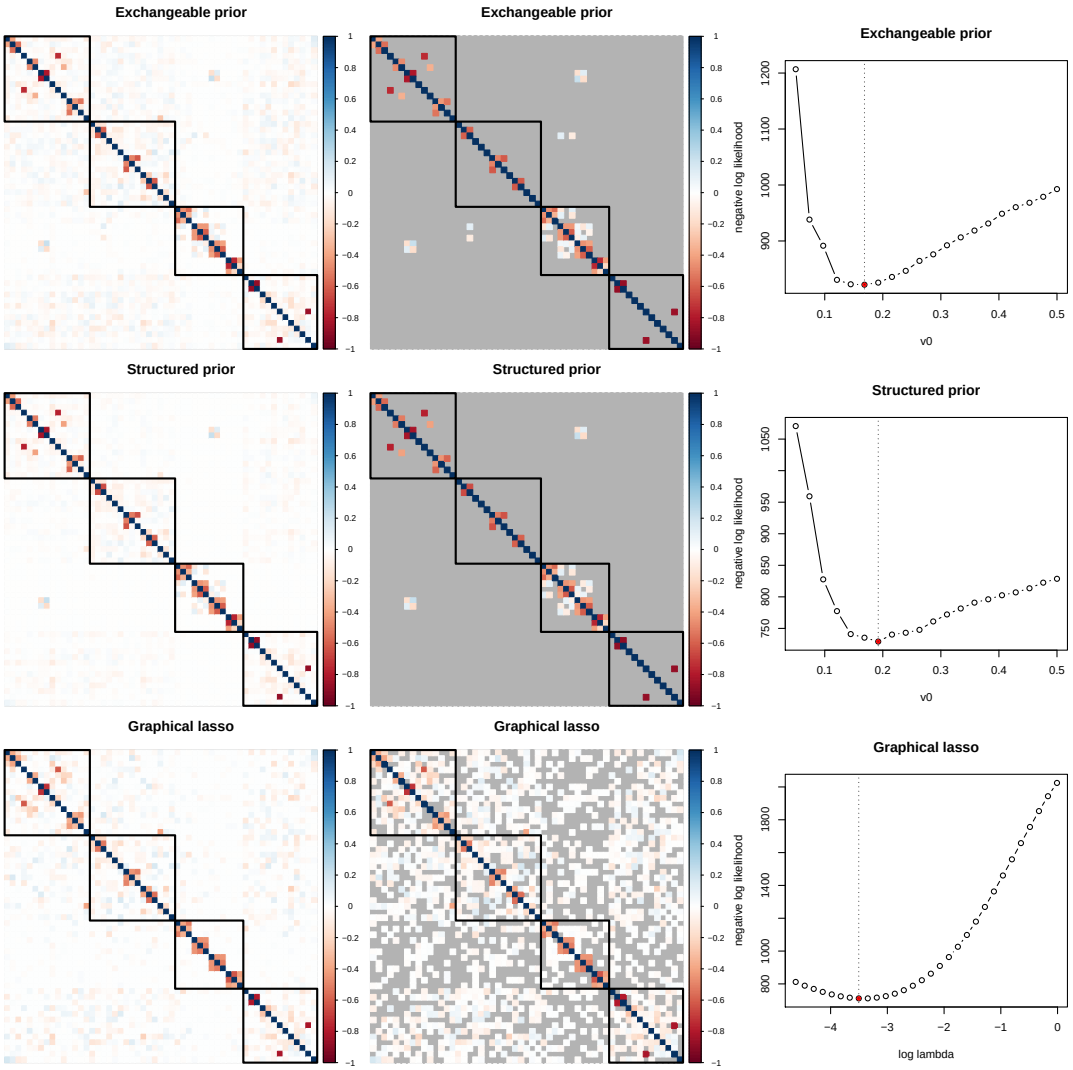


Figure 4.5: Comparing the estimated precision matrix from cross validation. The blocks correspond to product categories. From upper left to lower right: pretzels, cold cereal, frozen pizza, and mouthwash. Left column: estimated standardized precision matrix. Middle column: estimated standardized precision matrix with highlighted graph selection. Edges with less than 0.5 probability of being from the slab distributions in the first two plots, and exact zeros in graphical lasso output are marked with gray color. Right column: average negative log likelihood on the validation sets.

in Mazumder and Hastie (2012b). Multimodal posteriors are common. The proposed method could be extended to introduce perturbations in the algorithm, possibly drawing from the variable selection literature (see, e.g., Ročková and George, 2014; Rocková, 2018).

			M-norm		S-norm		F-norm		F_1 score	
Type	n	Graph	EMGS	Glasso	EMGS	Glasso	EMGS	Glasso	EMGS	Glasso
Gaussian	50	AR1	0.43	0.54	0.71	0.86	1.87	4.13	**0.89**	**0.89**
		AR2	0.47	0.50	1.25	1.39	4.41	4.87	0.22	**0.45**
		random	0.40	0.50	1.49	1.22	4.28	3.99	**0.42**	0.33
		cluster	0.41	0.54	0.99	0.98	2.83	3.05	0.46	**0.50**
	100	AR1	0.16	0.53	0.26	0.85	0.84	4.12	**1.00**	0.96
		AR2	0.33	0.50	0.68	1.31	1.94	4.54	**0.87**	0.58
		random	0.30	0.50	1.06	1.22	2.98	3.89	**0.53**	0.34
		cluster	0.30	0.52	0.68	0.96	2.00	2.97	**0.56**	0.53
	500	AR1	0.06	0.54	0.11	0.87	0.37	4.33	**1.00**	1.00
		AR2	0.09	0.42	0.16	1.23	0.55	4.20	**1.00**	0.65
		random	0.14	0.49	0.40	1.21	1.19	3.89	**0.73**	0.34
		cluster	0.14	0.54	0.28	0.99	0.83	3.06	**0.77**	0.54
Mixed	50	AR1	0.47	0.52	0.88	0.90	2.26	4.07	**0.85**	0.81
		AR2	0.51	0.50	1.43	1.43	4.25	4.98	**0.54**	0.41
		random	0.47	0.50	1.54	1.22	4.85	4.01	**0.36**	0.33
		cluster	0.51	0.56	1.19	1.00	3.42	3.14	0.42	**0.46**
	100	AR1	0.40	0.49	0.61	0.90	1.38	3.93	**0.96**	0.87
		AR2	0.49	0.50	1.09	1.39	2.78	4.69	**0.76**	0.54
		random	0.41	0.49	1.33	1.22	3.96	3.91	**0.44**	0.34
		cluster	0.40	0.53	0.94	0.96	2.56	3.00	0.50	**0.51**
	500	AR1	0.09	0.47	0.17	0.89	0.54	3.70	**1.00**	0.92
		AR2	0.18	0.49	0.31	1.35	0.78	4.42	**0.99**	0.64
		random	0.23	0.45	0.68	1.12	1.86	3.55	**0.64**	0.39
		cluster	0.20	0.54	0.46	0.95	1.13	2.99	**0.71**	0.54

Table 4.1: Comparing estimation of the standardized precision matrix for Gaussian graphical model and copula graphical model with mixed variables. The final graphs are chosen so that the sparsity level is closest to the truth.

Chapter 5

JOINT ESTIMATION OF MULTIPLE GAUSSIAN GRAPHICAL MODELS[1]

5.1 Introduction

Bayesian formulations of graphical models have been widely adopted as a way to characterize conditional independence structure among complex high-dimensional data. These models are popular in scientific domains including genomics (Briollais et al., 2016; Peterson et al., 2013), public health (Dobra, 2014; Li et al., 2017b), and economics (Dobra et al., 2010). In practice, data often come from several distinct groups. For example, data may be collected under various conditions, at different locations and time periods, or correspond to distinct subpopulations. Assuming a single graphical model in such cases can lead to unreliable estimates of network structure, whereas the alternative, estimating different graphical models separately for each group, may not be feasible for high dimensional problems.

Several approaches have been proposed to learn graphical models jointly for multiple classes of data. Much of this work extends the penalized maximum likelihood approach to incorporate additional penalty terms that encourage the class-specific precision matrices to be similar (Guo et al., 2011; Danaher et al., 2014; Saegusa and Shojaie, 2016; Ma and Michailidis, 2016). In the Bayesian literature, Peterson et al. (2015) and Lin et al. (2017) utilize Markov Random Field priors to model a super-graph linking different graphical models. Tan et al. (2017) uses a logistic regression model to link the connectivity of nodes to covariates specific

[1]The contents of this chapter are based on the paper "Bayesian joint spike-and-slab graphical lasso' (Li et al., 2018).

to each graph. These approaches only model the similarity of the underlying graphs, and thus are limited in their ability to borrow information when estimating the precision matrices. Borrowing strength is especially important when some classes have small sample sizes.

In this work, we introduce a new Bayesian formulation for estimating multiple related Gaussian graphical models by leveraging similarities in the underlying sparse precision matrices directly. We first present two shrinkage priors for multiple related precision matrices, as the Bayesian counterpart of joint graphical lasso estimators (Danaher et al., 2014). We then propose a doubly spike-and-slab mixture extension to these priors, which allows us to achieve simultaneous shrinkage and model selection, as well as hangle missing observations. In Section 5.5 and 5.6, we provide a fast Expectation-Maximization (EM) algorithm to quickly identify the posterior modes in a manner similar to (Li and McCormick, 2017) and (Deshpande et al., 2017). We also propose a procedure to sequentially explore a series of posterior modes. We then demonstrate the substantial improvements in both model selection and parameter estimation over the original joint graphical lasso approach using both simulated data and two real datasets in Section 5.7. Finally, in Section 5.8 we discuss future directions for improvements.

5.2 Preliminaries

5.2.1 The joint graphical lasso

We first briefly introduce the notation used throughout this chapter. We let G denote the number of classes in the data, and let $\mathbf{\Omega}_g$ and $\mathbf{\Sigma}_g$ denote the precision and covariance matrix for the g-th class. We let $\omega_{jk}^{(g)}$ denote the (j,k)-th element in $\mathbf{\Omega}_g$ and $\boldsymbol{\omega}_{jk} = \{\omega_{jk}^{(g)}\}_{g=1,...,G}$ denote the vector of all the (j,k)-th elements in $\{\mathbf{\Omega}\}$. Suppose we are given G datasets, $\boldsymbol{X}^{(1)}, ..., \boldsymbol{X}^{(G)}$, where $\boldsymbol{X}^{(g)}$ is a $n_g \times p$ matrix of independent centered observations from the distribution Normal$(\mathbf{0}, \mathbf{\Omega}_g^{-1})$. As maximum likelihood estimates of $\mathbf{\Omega}_g$ can have high variance

and are ill-defined when $p > n_g$, the joint penalized log likelihood for the G dataset is usually considered instead:

$$\ell(\{\boldsymbol{\Omega}\}) = \frac{1}{2}\sum_{g=1}^{G} n_g \log\det \boldsymbol{\Omega}_g - tr(\boldsymbol{S}_g\boldsymbol{\Omega}_g) - pen(\{\boldsymbol{\Omega}\}), \tag{5.1}$$

where $\boldsymbol{S}_g = (\boldsymbol{X}^{(g)})^T\boldsymbol{X}^{(g)}$. The penalty function encourages $\{\boldsymbol{\Omega}\}$ to have zeros on the off-diagonal elements and be similar across groups. In particular, we consider two useful penalty functions studied in Danaher et al. (2014), the group graphical lasso (GGL), and the fused graphical lasso (FGL):

$$pen(\{\boldsymbol{\Omega}\}) \quad = \quad \frac{\lambda_0}{2}\sum_{g}\sum_{j}|\omega_{jj}^{(g)}| + \lambda_1\sum_{g}\sum_{j<k}|\omega_{jk}^{(g)}| + \lambda_2\sum_{j<k}\widetilde{pen}(\boldsymbol{\omega}_{jk}), \tag{5.2}$$

where $\widetilde{pen}(\boldsymbol{\omega}_{jk}) = ||\boldsymbol{\omega}_{jk}||_2$ for GGL and $\sum_{g<g'}|\omega_{jk}^{(g)} - \omega_{jk}^{(g')}|$ for FGL. Both penalties encourage similarity across groups when $\lambda_2 > 0$, and reduce to separate graphical lasso problems when $\lambda_2 = 0$. The group graphical lasso encourages only similar patterns of zero elements across the G precision matrices, while the fused graphical lasso encourages a stronger form of similarity: the values of off-diagonal elements are also encouraged to be similar across the G precision matrices. In practice, λ_0 is typically set to 0 when the diagonal elements are not to be penalized.

5.2.2 Bayesian formulation of Gaussian graphical models

One of the most popular approaches for Bayesian inference with Gaussian graphical models is the G-Wishart prior (Lenkoski and Dobra, 2011; Mohammadi and Wit, 2015a). The G-Wishart prior estimates the precision matrices with exact zeros in the off-diagonal elements and enjoys the conjugacy with the Gaussian likelihood. However, posterior inference under

the G-Wishart prior can be computationally burdensome and has to rely on stochastic search algorithms over the large model space, consisting of all possible graphs. In recent years, several classes of shrinkage priors have been proposed for estimating large precision matrices, including the graphical lasso prior (Wang, 2012; Peterson et al., 2013), the continuous spike-and-slab prior (Wang, 2015; Li et al., 2017b), and the graphical horseshoe prior (Li et al., 2017a). This line of work draws direct connections between penalized likelihood schemes and, as their names suggest, the posterior modes in a Bayesian setting. Unlike the G-Wishart prior, these shrinkage priors do not take point mass at zero for the off-diagonal elements in the precision matrix, and thus usually lead to efficient block sampling algorithms with improved scalability. However, fully Bayesian procedures still need to rely on stochastic search to achieve model selection, making it less appealing for many problems.

To address this issue, deterministic algorithms have been proposed to perform fast posterior exploration and mode searching in Gaussian graphical models (Li and McCormick, 2017; Deshpande et al., 2017). Motivated by the EMVS (Ročková and George, 2014) and spike-and-slab lasso (Ročková and George, 2018) procedures in the linear regression literature, the idea is to use a two-component mixture distribution, i.e., spike-and-slab priors, to parameterize off-diagonal elements in the precision matrix, which allows simultaneous model selection and parameter estimation. We will utilize a similar strategy for model estimation in this chapter.

5.3 Bayesian joint graphical lasso priors

We first provide a Bayesian interpretation of the group and fused graphical lasso estimators. From a probabilistic perspective, it is well understood that estimators that optimize a penalized likelihood can often be seen as the posterior mode estimator under some suitable prior distributions. The Bayesian counterpart to (5.2) can be constructed by puttingr the prior $p(\{\boldsymbol{\Omega}\}) \propto \exp(-pen(\{\boldsymbol{\Omega}\}))$ on the precision matrices. Following directly from the

Bayesian representation of lasso variants demonstrated in Kyung et al. (2010), we can rewrite $p(\{\boldsymbol{\Omega}\})$ as products of scale mixtures of normal distributions on the off-diagonal elements. That is, for the GGL prior, we can let

$$p(\{\boldsymbol{\Omega}\}|\boldsymbol{\tau},\boldsymbol{\rho}) = C^{-1}_{\boldsymbol{\tau},\boldsymbol{\rho}} \prod_{j<k} \text{Normal}(\boldsymbol{\omega}_{jk}; \mathbf{0}, (\boldsymbol{\Theta}^{(G)}_{jk})^{-1}) \prod_g \prod_j \text{Exp}(\omega^{(g)}_{jj}; \frac{\lambda_0}{2}) \mathbf{1}_{\{\boldsymbol{\Omega}\}\in M^+}, \tag{5.3}$$

$$\boldsymbol{\Theta}^{(G)}_{jk} = \text{diag}(\{\frac{1}{\rho_{jk}} + \frac{1}{\tau_{jkg}}\}_{g=1,...,G}), \tag{5.4}$$

$$p(\boldsymbol{\tau},\boldsymbol{\rho}) \propto C_{\boldsymbol{\tau},\boldsymbol{\rho}} \prod_{j<k} \Big(\exp(-\frac{\lambda_1^2}{2}\sum_g \tau_{jkg} - \frac{\lambda_2^2}{2}\rho_{jk}) \rho_{jk}^{-\frac{1}{2}} \prod_g (\tau_{jkg}(\frac{1}{\tau_{jkg}} + \frac{1}{\rho_{jk}}))^{-\frac{1}{2}} \Big), \tag{5.5}$$

where $C_{\boldsymbol{\tau},\rho}$ is a normalizing constant and M^+ denotes the space of symmetric positive definite matrices. The normalizing constant is analytically intractable due to this constraint, but it cancels out in the marginal distribution of $p(\{\boldsymbol{\Omega}\})$. Such cancellation has been studied by several authors (Wang, 2012, 2015; Liu et al., 2014). Similarly, the FGL prior can be defined as

$$p(\{\boldsymbol{\Omega}\}|\boldsymbol{\tau},\boldsymbol{\phi}) = C^{-1}_{\boldsymbol{\tau},\boldsymbol{\phi}} \prod_{j<k} \text{Normal}(\boldsymbol{\omega}_{jk}; \mathbf{0}, (\boldsymbol{\Theta}^{(F)}_{jk})^{-1}) \prod_g \prod_j \text{Exp}(\omega^{(g)}_{jj}; \frac{\lambda_0}{2}) \mathbf{1}_{\{\boldsymbol{\Omega}\}\in M^+}, \tag{5.6}$$

$$\boldsymbol{\Theta}^{(F)}_{jk} = \begin{cases} \theta_{gg} = \frac{1}{\tau_{jkg}} + \sum_{g\neq g'} \frac{1}{\phi_{jkgg'}} & g = 1, ..., G \\ \theta_{gg'} = -\frac{1}{\phi_{jkgg'}} & g' \neq g \end{cases} \tag{5.7}$$

$$p(\boldsymbol{\tau},\boldsymbol{\phi}) \propto C_{\boldsymbol{\tau},\boldsymbol{\phi}} \prod_{j<k} \Big(|\boldsymbol{\Theta}^{(F)}_{jk}|^{-\frac{1}{2}} \exp(-\frac{\lambda_1^2}{2}\sum_g \tau_{jkg} - \frac{\lambda_2^2}{2}\sum_{g<g'} \phi_{jkgg'}) \prod_g \tau_{jkg}^{-\frac{1}{2}} \prod_{g<g'} \phi_{jkgg'}^{-\frac{1}{2}} \Big). \tag{5.8}$$

It is also worth noting that both of the above priors are proper, and we leave the proof of the following proposition in the appendix.

Proposition 1. *The priors defined in (5.3) – (5.5) and (5.6) – (5.8) are proper and the*

posterior mode of $\{\boldsymbol{\Omega}\}$ *is the solution of the group and fused graphical lasso problem with penalty terms defined in (5.2).*

5.4 Bayesian joint spike-and-slab graphical lasso priors

The Bayesian formulation of the joint graphical lasso problems discussed in the previous section provide shrinkage effects at the level of both individual precision matrices and across different classes. However, two issues remain. First, shrinkage priors alone do not produce sparse models since the posterior draws are never exactly 0. Thus, additional thresholding is needed to obtain a sparse representation of the graph structure. Second, the fixed penalty term, λ_1 and λ_2 may be too restrictive, as the non-zero elements in $\{\boldsymbol{\Omega}\}$ are penalized equally to elements close to zero (Li and McCormick, 2017). To reduce the bias from over-penalizing the large elements, different hyper-priors on λ_1 have been proposed to adaptively estimate the penalty term in Bayesian graphical lasso (Wang, 2012; Peterson et al., 2013).

Here we address both challenges simultaneously using the spike-and-slab approaches in Bayesian variable selection (George and McCulloch, 1993). In particular, we employ a set of latent indicators to construct a "selection" prior on both the group level and within-groups for the similarity penalties. We first let binary variables $\boldsymbol{\delta} = \{\delta_{jk}\}_{j<k}$ denote the existence of each edge in the graph, indexing the $2^{p(p-1)/2}$ possible models at the group level, so that $\delta_{jk} = 1$ indicates the (j,k)-th edge is selected for all precision matrices. We then let another set of binary variables $\boldsymbol{\xi} = \{\xi_{jk}\}_{j<k}$ denote the *non-existence* of 'similarities' among the elements in the same cell of different precision matrices, so that $\xi_{jk} = 0$ indicates the (j,k)-th element is expected to be similar. We use the term 'similarity' here as a broad term parameterized by λ_2, since the behavior of the similarity depends on the form of the penalization. Conditional on the two binary indicators, we replace the fixed penalty parameters λ_1 and λ_2 by a mixture of edge-wise penalties that take values from $\{\lambda_1/v_0, \lambda_1/v_1\}$, and $\{\lambda_2/v_0, \lambda_2/v_1\}$ respectively,

with fixed $v_1 > v_0 > 0$. That is, we introduce the following penalties conditional on $\boldsymbol{\delta}$ and $\boldsymbol{\xi}$, and we propose the *doubly spike-and-slab* extensions to GGL and FGL as

$$pen(\{\boldsymbol{\Omega}\}|\boldsymbol{\delta},\boldsymbol{\xi}) = \frac{\lambda_0}{2}\sum_g\sum_j |\omega_{jj}^{(g)}| + \lambda_1\sum_g\sum_{j<k}\frac{|\omega_{jk}^{(g)}|}{v_{\delta_{jk}}} + \lambda_2\sum_{j<k}\frac{\widetilde{pen}(\boldsymbol{\omega}_{jk})}{v_{\xi^*_{jk}}}, \tag{5.9}$$

where $\widetilde{pen}(\boldsymbol{\omega}_{jk})$ is defined as before and $\xi^*_{jk} = \xi_{jk}\delta_{jk}$. The prior defined in (5.9) relate to the unconditional penalties by $pen(\{\boldsymbol{\Omega}\}) = pen(\{\boldsymbol{\Omega}\}|\boldsymbol{\delta},\boldsymbol{\xi}) - \log(p(\boldsymbol{\delta},\boldsymbol{\xi}))$, and we will refer to them as DSS-FGL and DSS-GGL.

In practice, we find it usually reasonable to enforce all elements from the spike distribution to also be similar, since the spike distribution is always chosen to have large penalization and leads to posterior modes at exactly 0. However, other types of element-wise dependence between δ_{jk} and ξ_{jk} are also possible with minor modifications. For example, we can also fix ξ_{jk} to be 1, so that the two penalty terms will always be proportional. We refer to this setting as spike-and-slab group and fused lasso (SS-GGL and SS-FGL) and discuss their behavior in the appendixs.

The original GGL and FGL suffer from the same bias induced by the excessive shrinkage of lasso estimates. With the introduction of v_0 and v_1, we can adaptively estimate which $\boldsymbol{\omega}_{jk}$ to penalize in a data-driven way. As we discuss in more detail in Section 5.6, this adaptive shrinkage property can indeed significantly reduce bias imposed on the lasso penalty. That is, by choosing the hyperparameters so that $\lambda_i/v_0 \gg \lambda_i/v_1$, we impose only minimal shrinkage on values arising from the slab distribution. From now on, in order to avoid confusion from the overparameterization, we always fix $v_1 = 1$, and report results with the effective shrinkage parameters $\lambda_i/v_j, i, j \in \{1, 2\}$. At this point, it may still seem that we need introduced one more hyperparameter that needs to be tuned, but as we show in Section 5.6, model selection can be achieved automatically without cross-validation.

For a fully Bayesian setup, we employ standard priors on the binary indicators to allow the edges to further share information on the sparsity level. The full generative model for $\{\boldsymbol{\Omega}\}$ is:

$$p(\{\boldsymbol{\Omega}\}|\boldsymbol{\delta},\boldsymbol{\xi}) = C^{-1}_{\boldsymbol{\delta},\boldsymbol{\xi}} \exp(-pen(\{\boldsymbol{\Omega}\}|\boldsymbol{\delta},\boldsymbol{\xi}))\mathbf{1}_{\{\boldsymbol{\Omega}\}\in M^+} \tag{5.10}$$

$$p(\boldsymbol{\delta},\boldsymbol{\xi}|\pi_{\boldsymbol{\delta}},\pi_{\boldsymbol{\xi}}) \propto C_{\boldsymbol{\delta},\boldsymbol{\xi}} \prod_{j<k} \pi_{\boldsymbol{\delta}}^{\delta_{jk}}(1-\pi_{\boldsymbol{\delta}})^{1-\delta_{jk}} \pi_{\boldsymbol{\xi}}^{\xi_{jk}}(1-\pi_{\boldsymbol{\xi}})^{1-\xi_{jk}} \tag{5.11}$$

where θ denote $(\boldsymbol{\tau},\boldsymbol{\rho})$ for DSS-GGL, and $(\boldsymbol{\tau},\boldsymbol{\phi})$ for DSS-FGL. $C_{\boldsymbol{\delta},\boldsymbol{\xi}}$ is another intractable normalizing constant. We put standard Beta hyperpriors on the sparsity parameters so that $\pi_{\boldsymbol{\delta}} \sim \text{Beta}(a_1,b_1)$ and $\pi_{\boldsymbol{\xi}} \sim \text{Beta}(a_2,b_2)$. Throughout this chapter, we let $a_1 = a_2 = 1$ and $b_1 = b_2 = p$. Additionally, the (5.10) can be easily reparameterized with scale mixture of normal prior distributions similar as before. For DSS-GGL, it can be shown that the scale mixture of normal representation is

$$p(\{\boldsymbol{\Omega}\}|\boldsymbol{\tau},\boldsymbol{\rho},\boldsymbol{\delta},\boldsymbol{\xi}) = C^{-1}_{\boldsymbol{\tau},\boldsymbol{\rho}} C^{-1}_{\boldsymbol{\delta},\boldsymbol{\xi}} \prod_{j<k} \text{Normal}(\boldsymbol{\omega}_{jk};\mathbf{0},(\boldsymbol{\Theta}^{(G)}_{jk})^{-1}) \prod_{g,j} \text{Exp}(\omega^{(g)}_{jj};\frac{\lambda_0}{2})\mathbf{1}_{\{\boldsymbol{\Omega}\}\in M^+}, \tag{5.12}$$

$$\boldsymbol{\Theta}^{(G)}_{jk} = \text{diag}(\{\frac{v_{\xi^*_{jk}}}{\rho_{jk}} + \frac{v_{\delta_{jk}}}{\tau_{jkg}}\}_{g=1,\ldots,G}), \tag{5.13}$$

and for DSS-FGL, it can be written as

$$p(\{\boldsymbol{\Omega}\}|\boldsymbol{\tau},\boldsymbol{\phi},\boldsymbol{\delta},\boldsymbol{\xi}) = C^{-1}_{\boldsymbol{\tau},\boldsymbol{\phi}} C^{-1}_{\boldsymbol{\delta},\boldsymbol{\xi}} \prod_{j<k} \text{Normal}(\boldsymbol{\omega}_{jk};\mathbf{0},(\boldsymbol{\Theta}^{(F)}_{jk})^{-1}) \prod_{g,j} \text{Exp}(\omega^{(g)}_{jj};\frac{\lambda_0}{2})\mathbf{1}_{\{\boldsymbol{\Omega}\}\in M^+}, \tag{5.14}$$

$$\boldsymbol{\Theta}^{(F)}_{jk} = \begin{cases} \theta_{gg} = \frac{v_{\delta_{jk}}}{\tau_{jkg}} + \sum_{g<g'} \frac{v_{\xi^*_{jk}}}{\phi_{jkgg'}} & g = 1,\ldots,G \\ \theta_{gg'} = -\frac{v_{\xi^*_{jk}}}{\phi_{jkgg'}} & g' \neq g \end{cases} \tag{5.15}$$

Proposition 2. *The DSS-GGL prior defined by (5.12), (5.13), (5.5) and (5.11), and the DSS-FGL prior defined by (5.14), (5.15), (5.8), and (5.11) are proper, and the posterior*

mode of $\{\boldsymbol{\Omega}\}$ *is the solution to the corresponding spike-and-slab version of joint graphical lasso penalties in (5.9).*

Finally, it is straightforward to see that the proposed DSS-GGL and DSS-FGL penalties reduce to their non spike-and-slab counterparts when $\boldsymbol{\delta}$ and $\boldsymbol{\xi}$ are fixed to be 1. Several other spike-and-slab formulations in the literature can be seen as the special case of this prior when $G = 1$ as well. For example, the spike-and-slab mixture of double exponential priors considered in (Deshpande et al., 2017) is a special case with $\lambda_2 = 0$. The spike-and-slab Gaussian mixtures in (Li and McCormick, 2017) can also be considered as a special case where we further fix $\tau_{jkg} = \infty$. This approach is also related to the work on sparse group selection in linear regression, as has been discussed in (Xu and Ghosh, 2015) and (Zhang et al., 2014). As opposed to the point mass priors for the spike distribution commonly in the literature, our doubly spike-and-slab formulation of continuous mixtures allows the spike distribution to absorb small non-zero noises and facilitates fast dynamic explorations, as we will show in Section 5.6.

5.5 Model estimation

Given fixed λ_1, λ_2, and v_0, The representation of $p(\{\boldsymbol{\Omega}\})$ with the scale mixture of normal distributions allows the posterior to be sampled using a block Gibbs algorithm, as described in the appendix. However, choosing the hyperparameters can usually be a nontrivial task. Instead, we focus on faster deterministic methods to detect posterior modes under different choices of hyperparameters (Ročková and George, 2014). We present an EM algorithm that maximizes the complete-data posterior distribution $p(\{\boldsymbol{\Omega}\}, \boldsymbol{\delta}, \boldsymbol{\xi}, \pi_{\boldsymbol{\delta}}, \pi_{\boldsymbol{\xi}} | \boldsymbol{X})$ by treating the binary latent variables as "missing data." Similar ideas have been explored in recent work for linear regression (Ročková and George, 2014, 2018) and single graphical model estimation (Deshpande et al., 2017; Li and McCormick, 2017). Assuming no missing data,

the objective function of the EM algorithm in the t-th iteration is the expectation of the complete data log likelihood, i.e.,

$$
\begin{aligned}
Q(\{\boldsymbol{\Omega}\}, \pi_{\boldsymbol{\delta}}, \pi_{\boldsymbol{\xi}} | \{\boldsymbol{\Omega}\}^{(t)}, \pi_{\boldsymbol{\delta}}^{(t)}, \pi_{\boldsymbol{\xi}}^{(t)}) &= E_{\boldsymbol{\delta},\boldsymbol{\xi}|\{\boldsymbol{\Omega}\}^{(t)},\pi_{\boldsymbol{\delta}}^{(t)},\pi_{\boldsymbol{\xi}}^{(t)},\boldsymbol{X}}(\log p(\{\boldsymbol{\Omega}\}, \pi_{\boldsymbol{\delta}}, \pi_{\boldsymbol{\xi}} | \boldsymbol{X}) | \{\boldsymbol{\Omega}\}^{(t)}, \pi_{\boldsymbol{\delta}}^{(t)}, \pi_{\boldsymbol{\xi}}^{(t)}, \boldsymbol{X}) \\
&= \text{constant} + \sum_g \frac{n_g}{2} \log |\boldsymbol{\Omega}_g| - \frac{1}{2} \sum_g tr(\boldsymbol{S}_g \boldsymbol{\Omega}_g) - \frac{\lambda_0}{2} \sum_j \sum_g |\omega_{jj}^{(g)}| \\
&\quad -\lambda_1 \sum_{j<k} \sum_g |\omega_{jk}^{(g)}| E_{\cdot|\cdot}[\frac{1}{v_0(1-\delta_{jk}) + v_1 \delta_{jk}}] + \sum_{j<k} \log(\frac{\pi_{\boldsymbol{\delta}}}{1-\pi_{\boldsymbol{\delta}}}) E_{\cdot|\cdot}(\delta_{jk}) \\
&\quad -\lambda_2 \sum_{j<k} \widetilde{pen}(\boldsymbol{\omega}_{jk}) E_{\cdot|\cdot}[\frac{1}{v_0(1-\delta_{jk}\xi_{jk}) + v_1 \delta_{jk}\xi_{jk}}] \\
&\quad + \sum_{j<k} \log(\frac{\pi_{\boldsymbol{\xi}}}{1-\pi_{\boldsymbol{\xi}}}) E_{\cdot|\cdot}(\xi_{jk}) \\
&\quad + (a_1 - 1)\log(\pi_{\boldsymbol{\delta}}) + (b_1 + \frac{p(p-1)}{2} - 1)\log(1-\pi_{\boldsymbol{\delta}}) \\
&\quad + (a_2 - 1)\log(\pi_{\boldsymbol{\xi}}) + (b_2 + \frac{p(p-1)}{2} - 1)\log(1-\pi_{\boldsymbol{\xi}}),
\end{aligned}
$$

where $E_{\cdot|\cdot}$ denotes conditional expectation $E_{\boldsymbol{\delta},\boldsymbol{\xi}|\{\boldsymbol{\Omega}\}^{(t)},\pi_{\boldsymbol{\delta}}^{(t)},\pi_{\boldsymbol{\xi}}^{(t)},\boldsymbol{X}}$, and $\widetilde{pen}(\boldsymbol{\omega}_{jk}) = ||\boldsymbol{\omega}_{jk}||_2$ for DSS-GGL and $\widetilde{pen}(\boldsymbol{\omega}_{jk}) = \sum_{g<g'} |\omega_{jk}^{(g)} - \omega_{jk}^{(g')}|$ for DSS-FGL.

In the E-step, we compute the conditional expectation terms in the objective function. It turns out that it suffices to find the conditional distribution of (δ_{jk}, ξ_{jk}). The corresponding cell probabilities are proportional to the following mixture densities:

$$
p^*_{\delta_{jk},\xi_{jk}}(j,k) \propto \begin{cases}
\pi_{\boldsymbol{\delta}}(1-\pi_{\boldsymbol{\xi}}) \frac{\lambda_1\lambda_2}{v_0 v_1} \exp(-\lambda_1 \sum_g |\omega_{jk}^{(g)}|/v_1 - \lambda_2 \widetilde{pen}(\boldsymbol{\omega}_{jk})/v_0) & \delta_{jk}=1, \xi_{jk}=0 \\
\pi_{\boldsymbol{\delta}}\pi_{\boldsymbol{\xi}} \frac{\lambda_1\lambda_2}{v_1^2} \exp(-\lambda_1 \sum_g |\omega_{jk}^{(g)}|/v_1 - \lambda_2 \widetilde{pen}(\boldsymbol{\omega}_{jk})/v_1) & \delta_{jk}=1, \xi_{jk}=1 \\
(1-\pi_{\boldsymbol{\delta}})(1-\pi_{\boldsymbol{\xi}}) \frac{\lambda_1\lambda_2}{v_0^2} \exp(-\lambda_1 \sum_g |\omega_{jk}^{(g)}|/v_0 - \lambda_2 \widetilde{pen}(\boldsymbol{\omega}_{jk})/v_0) & \delta_{jk}=0, \xi_{jk}=0
\end{cases}
$$

It is interesting to note that the three scenarios above represent three types of relationships among $\boldsymbol{\omega}_{jk}$: weak shrinkage but strong similarity, weak shrinkage and weak similarity, and

strong shrinkage across classes. $E_{\cdot|\cdot}(\delta_{jk})$ and $E_{\cdot|\cdot}(\xi_{jk})$ are then simply the marginal probabilities in this 2 by 2 table, i.e., $E_{\cdot|\cdot}(\delta_{jk}) = p^*_{1,0}(j,k) + p^*_{1,1}(j,k)$, and $E_{\cdot|\cdot}(\xi_{jk}) = E_{\cdot|\cdot}(\delta_{jk}\xi_{jk}) = p^*_{1,1}(j,k)$. The EM algorithm also handles missing cells in $\boldsymbol{X}$ naturally. Assuming missing at random, the expectation can also be taken over the space of missing variables, by additionally computing $E_{\cdot|\cdot}(tr(\boldsymbol{S}_g\boldsymbol{\Omega}_g)) = tr(E_{\cdot|\cdot}(\boldsymbol{S}_g\boldsymbol{\Omega}_g))$, using the conditional Gaussian distribution of $\boldsymbol{x}^{(g)}_{i,m}|\boldsymbol{x}^{(g)}_{i,o}$, where $\boldsymbol{x}^{(g)}_{i,o}$ and $\boldsymbol{x}^{(g)}_{i,m}$ denote the observed and missing cells in $\boldsymbol{x}^{(g)}_i$ respectively. That is,

$$E_{\cdot|\cdot}(\boldsymbol{S}_g\boldsymbol{\Omega}_g) = E_{\cdot|\cdot}((\frac{1}{n_g}\sum_i^{n_g}\boldsymbol{x}^{(g)}_i(\boldsymbol{x}^{(g)}_i)^T)\boldsymbol{\Omega}_g) = \frac{1}{n_g}\big(\sum_i^{n_g} E_{\boldsymbol{x}^{(g)}_{i,m}|\boldsymbol{x}^{(g)}_{i,o}}(\boldsymbol{x}^{(g)}_i(\boldsymbol{x}^{(g)}_i)^T)\big)\boldsymbol{\Omega}_g.$$

Since $\boldsymbol{x}^{(g)}_i$ follows a multivariate Gaussian distribution, without loss of generality, if we let $\boldsymbol{x}^{(g)}_i = \begin{pmatrix}\boldsymbol{x}^{(g)}_{i,o}\\ \boldsymbol{x}^{(g)}_{i,m}\end{pmatrix}$, we know

$$\begin{aligned} E_{\boldsymbol{x}^{(g)}_{i,m}|\boldsymbol{x}^{(g)}_{i,o}}(\boldsymbol{x}^{(g)}_{i,m}) &= \boldsymbol{\Sigma}_{mo}\boldsymbol{\Sigma}^{-1}_{oo}\boldsymbol{x}^{(g)}_{i,o} \\ E_{\boldsymbol{x}^{(g)}_{i,m}|\boldsymbol{x}^{(g)}_{i,o}}(\boldsymbol{x}^{(g)}_i(\boldsymbol{x}^{(g)}_i)^T) &= E_{\cdot|\cdot}(\boldsymbol{x}^{(g)}_i)E_{\cdot|\cdot}(\boldsymbol{x}^{(g)}_i)^T + \begin{pmatrix}\boldsymbol{0}_{oo} & \boldsymbol{0}_{om}\\ \boldsymbol{0}_{mo} & \boldsymbol{\Sigma}_{mm} - \boldsymbol{\Sigma}_{mo}\boldsymbol{\Sigma}^{-1}_{oo}\boldsymbol{\Sigma}_{om}\end{pmatrix} \end{aligned}$$

where $\boldsymbol{\Sigma}_{oo}, \boldsymbol{\Sigma}_{om}, \boldsymbol{\Sigma}_{mo}$ and $\boldsymbol{\Sigma}_{mm}$ are the corresponding submatrices of $\boldsymbol{\Sigma}_g$.

Given the expectations calculated in the E-step, the maximization step can be decomposed into three separate steps. First it is straightforward to see that the maximization of $\pi_{\boldsymbol{\delta}}$ and $\pi_{\boldsymbol{\xi}}$ have the close-form solutions:

$$\begin{aligned} \pi^{(k+1)}_{\boldsymbol{\delta}} &= (a_1 + \sum_{j<k}\delta_{jk} - 1)/(a_1 + b_1 + p(p-1)/2 - 2), \\ \pi^{(k+1)}_{\boldsymbol{\xi}} &= (a_2 + \sum_{j<k}\xi_{jk} - 1)/(a_2 + b_2 + p(p-1)/2 - 2). \end{aligned}$$

For the maximization of $\{\boldsymbol{\Omega}\}$, one might proceed with conditional maximization steps using gradient ascent similar to the Gibbs sampler (Li and McCormick, 2017). Alternatively, since the maximization step is equivalent to solving the following joint graphical lasso problem:

$$\{\boldsymbol{\Omega}\} = \text{argmax}_{\{\boldsymbol{\Omega}\}} \sum_g \frac{n_g}{2} \log|\boldsymbol{\Omega}_g| - \frac{1}{2}\sum_g tr(\boldsymbol{S}_g\boldsymbol{\Omega}_g) - \frac{\lambda_0}{2}\sum_j\sum_g |\omega_{jj}^{(g)}|$$
$$- \sum_{j<k} \lambda_1\left(\frac{p^*_{0,0}(j,k)}{v_0} + \frac{1-p^*_{0,0}(j,k)}{v_1}\right)\sum_g |\omega_{jk}^{(g)}| - \sum_{j<k}\lambda_2\left(\frac{1-p^*_{1,1}(j,k)}{v_0} + \frac{p^*_{1,1}(j,k)}{v_1}\right)\widetilde{pen}(\boldsymbol{\omega}_{jk}),$$

meaning we can use the ADMM algorithm described in (Danaher et al., 2014).

5.6 Dynamic posterior exploration

The algorithm proposed in the previous section requires a fixed set of hyperparameters, $(\lambda_0, \lambda_1, \lambda_2, v_0)$. The posterior is relatively insensitive to the choice of λ_0 as long as it is not too large (Wang, 2015). Furthermore, unlike the original joint graphical lasso, where two tunning parameters need to be selected using cross-validation or model selection criterion, it turns out that we can leverage the self-adaptive property from the doubly spike-and-slab mixture setup to achieve automatic tuning using a path-following strategy (Ročková and George, 2018). Specifically, we consider a sequence of decreasing $v_0 = \{v_0^1, ..., v_0^L\}$ and some small λ_1 and λ_2. We initiate $\{\boldsymbol{\Omega}\}_0$ so that $\boldsymbol{\Omega}_{g0} = (\boldsymbol{S}_g/n_g + c\boldsymbol{I})^{-1}$, and iterative estimate $\{\widehat{\boldsymbol{\Omega}}\}_l$ with $v_0 = v_0^l$. After fitting the l-th model, we use the estimated graph structure to warm start the $(l+1)$-th model by initiating $\boldsymbol{\Omega}_g$ to be $\boldsymbol{\Omega}_{g0} \circ \mathbf{1}^l_{\hat{\boldsymbol{\delta}}>0}$, where $\mathbf{1}^l_{\hat{\boldsymbol{\delta}}>0}$ denotes the group level graph structure at the l-th iteration. As v_0 decreases, the shrinkage imposed on the spike elements steadily increases and leads to sparser models. As noted in (Ročková and George, 2018), the solution path from such dynamic reinitialization procedure usually 'stabilizes' as v_0 becomes closer to 0 in linear regression. We found similar behavior in our spike-and-slab joint graphical lasso models too, as illustrated in Figure 5.1.

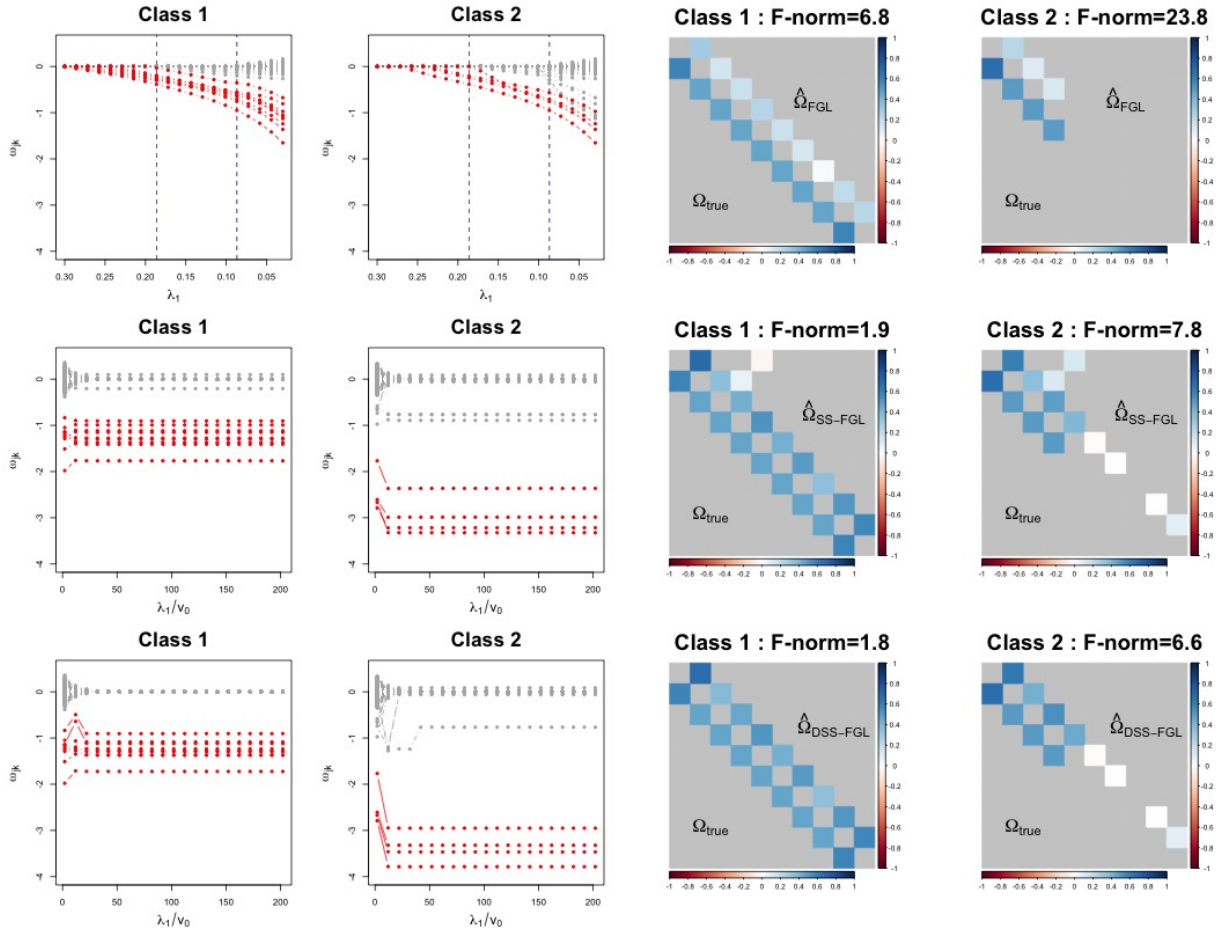


Figure 5.1: The solution paths and estimated precision matrices of FGL (upper row), SS-FGL (middle row) and DSS-FGL (lower row). The red nodes correspond to true edges and the gray nodes correspond to 0's. The two vertical lines in the FGL solution path indicate the model that best matches the true sparsity (left) and the model with the lowest AIC (right). The block containing the edges is plotted for the estimated values (upper triangular) against the truth (lower triangular). The model that best matches the true graphs is plotted for FGL. The off-diagonal values are rescaled and negated to partial correlations, and 0's are colored with light gray background for easier visual comparison. The bias of the estimated precision matrix measured by the Frobenius norm, $||\hat{\mathbf{\Omega}}_g - \mathbf{\Omega}_g||_F$, is also printed in the captions.

To demonstrate the dynamic posterior exploration in action, we simulated a small dataset from two classes, with $n_g = 150$ for $g = 1, 2$, and $p = 100$. The two underlying graphs differ by 5 edges: The first precision matrix contains a 10-node block with an AR(1) precision matrix where $(\Omega^{-1})_{jk} = \rho_1^{|j-k|}$, and $\rho_1 = 0.7$; the second precision matrix in the second class contains a common 5-node AR(1) block with $\rho_2 = 0.9$. The rest of the nodes are all independent. We fit the fused graphical lasso with a sequence of λ_1, and fixed $\lambda_2 = 0.1$, which leads to the best performance in this experiment; and DSS-FGL with $\lambda_1 = 1$, and $\lambda_2 = 1$. Figure 5.1 shows the FGL and DSS-FGL solution path. Unlike the continuous shrinkage of FGL, the zero and non-zero elements under DSS-FGL tend to be separated into two stable clusters as the effective shrinkage λ_1/v_0 increases beyond a critical point. Danaher et al. (2014) noted that graph selection using AIC tends to favor large models. This example also confirms this observation as the likelihood evaluation for smaller models suffers from the overly aggressive shrinkage. In this example, AIC selects 27 edges in both classes, leading to 41 false positives. Assuming we know the true graphs, the best model in terms of edge selection along the FGL solution path contains one false negative edge as shown in Figure 5.1. However, without accurate prior knowledge of graph sparsity, correctly identifying this model is typically difficult, if not impossible. On the other hand, the stable model from the DSS-FGL solution path yields 4 false positive edges in the second graph, but with clear visual separation from the regularization plot: only one false positive edge stabilizing to a larger value away from 0. Thus in practice, the solution path also provides a visual tool to threshold the small values close to 0. Additionally, the bias of the final precision matrices compared to the truth is also much smaller than the best FGL solution.

Figure 5.1 also shows the solution path from SS-FGL. It can be seen that although SS-FGL achieves similar bias as DSS-FGL, it also estimates several more false positive edges. This can be seen from the formulation of the doubly spike-and-slab selection: with only one

spike-and-slab mixture of the penalties, the selected edges from the slab distributions receive also only weak penalization for between-class similarities. Thus it is more likely to pick up spurious edges due to noises that happen to exist in one class. This illustration of the simple example shows the advantage of having the doubly spike-and-slab setup.

We also find that the converged region is insensitive to the choice of λ_1 and λ_2 in all our experiments, as the model allows a flexible combination of shrinkage through the adaptive estimation of p^*. The appendix includes an empirical assessment of sensitivity in the simulation experiments.

5.7 Numerical results

Simulation experiments To assess the performance of the proposed models, we consider a three-class problem similar to the study carried out in (Danaher et al., 2014). We first generate three networks with p features with 10 equal sized unconnected subnetwork. Each of the subnetwork follow a power law degree distribution, which is generally harder to estimate than simpler structures (Peng et al., 2009). The first class contains all ten subnetworks, and the second and third classes each has one and two subnetworks removed. Given the network structure, we generate $\mathbf{\Omega}_g$ from the G-Wishart distribution $W_G(3, I_p)$, and rescale them so that $\mathbf{\Omega}_g^{-1}$ have unit variances. Finally, we generate $n = 150$ independent and identically distributed samples from Normal$(\mathbf{0}, \mathbf{\Omega}_g)$ in each class. The resulted graph for $p = 500$ is shown in Figure B.1. We fit GGL and FGL with various choice of fixed λ_2 and a sequence of λ_1. We fit DSS-GGL and DSS-FGL with $\lambda_1 = 1$, $\lambda_2 = 2/30$ in this case. We explore more choices of λ_2 in the moderate dimensional experiments below and found no substantial changes in the performance of the final models.

The results comparing the proposed model and joint graphical lasso are shown in Figure 5.2. As discussed before, the DSS-FGL and DSS-GGL achieve model selection automatically. Thus

we compare the selected models with the average curve of FGL and GGL under different tuning parameters. Figure 5.2(a) and (c) show that DSS-FGL and DSS-GGL usually achieves better structure learning performance for both identifying edges and *differential edges*. The differential edges are defined as the edges for the (g, g') pair with $|\omega_{jk}^{(g)} - \omega_{jk}^{(g')}| > 0.01$. Figure 5.2(b) and (d) clearly demonstrate the bias-diminishing property of the proposed models. On average, both the sum of bias as measured by the Frobenius norm, $||\hat{\mathbf{\Omega}}_g - \mathbf{\Omega}_g||_F$, and the Kullback-Leibler (KL) divergence achieved by the proposed model is much smaller.

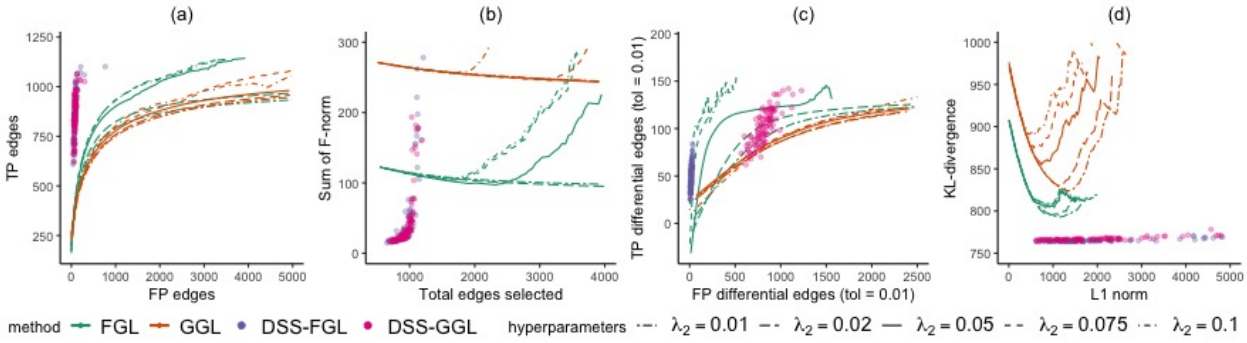


Figure 5.2: Performance of FGL, GGL, DSS-FGL, and DSS-GGL over 100 replications. The dots represent the metrics for the 100 selected models under DSS-FGL and DSS-GGL, and the lines represent the average performance of FGL and GGL over 100 replications under different tuning parameters.

Symptom networks of verbal autopsy data We applied the DSS-FGL and DSS-GGL to a gold-standard dataset of verbal autopsy (VA) surveys Murray et al. (2011a). VA surveys are widely adopted in countries without full-coverage civil registration and vital statistics systems to estimate cause of death. They are conducted by interviewing caregiver of a recently deceased person about the decedent's health history. The standard procedure of preparing the collected data is to dichotomize all continuous variables into binary indicators and many algorithms have been proposed to automatically assign causes of death using the

binary input (Byass et al., 2012; Serina et al., 2015; McCormick et al., 2016). However, more information may be gained by modeling the continuous variables directly (Li et al., 2017b). Here we focus on modeling the joint distribution of the continuous variables. The 27 continuous variables in this dataset contain representations of the duration of symptoms, such as response to the question 'how many days did the fever last', and age of the decedents. It is usually reasonable to assume the response to these questions are jointly distributed in similar ways conditional on each cause of death. We take the raw responses and transform raw duration x_{ij} by $\log(x_{ij}+1)$. We then let $X_{ij}^{(g)}$ denote the j-th transformed variable for observation i due to the cause g. The full dataset contains death assigned to 34 causes. We applied DSS-FGL with $\lambda_1 = \lambda_2 = 1$ to the three largest determined causes of death in this data: Stroke ($n = 630$), Pneumonia ($n = 540$), and AIDS ($n = 542$) in Figure 5.3. The estimated graphs under other models are discussed in the appendix. Both DSS-FGL and DSS-GGL estimated similar graphs and discovered interesting differential symptom pairs, such as the strong conditional dependence between the duration of illness and paralysis in deaths due to stroke. Further incorporating the DSS-FGL and DSS-GGL formulation of multiple precision matrices into a classification framework would likely improve accuracy over existing methods (e.g. McCormick et al. (2016); Byass et al. (2012)) for automatic cause-of-death assignment.

Prediction of missing mortality rates Beyond structure learning, the bias reduction in estimating $\{\mathbf{\Omega}\}$ also makes the proposed method more appealing for prediction tasks involving sparse precision matrices. In this example, we illustrate the potential of using the proposed methods to impute missing mortality rates using a cross-validation study. We construct the data matrices $X_{ij}^{(g)}$ as the log transformed central mortality rate of age group j in year i for subpopulation g (e.g., male and female). Standard approaches in demography, such as the Lee-

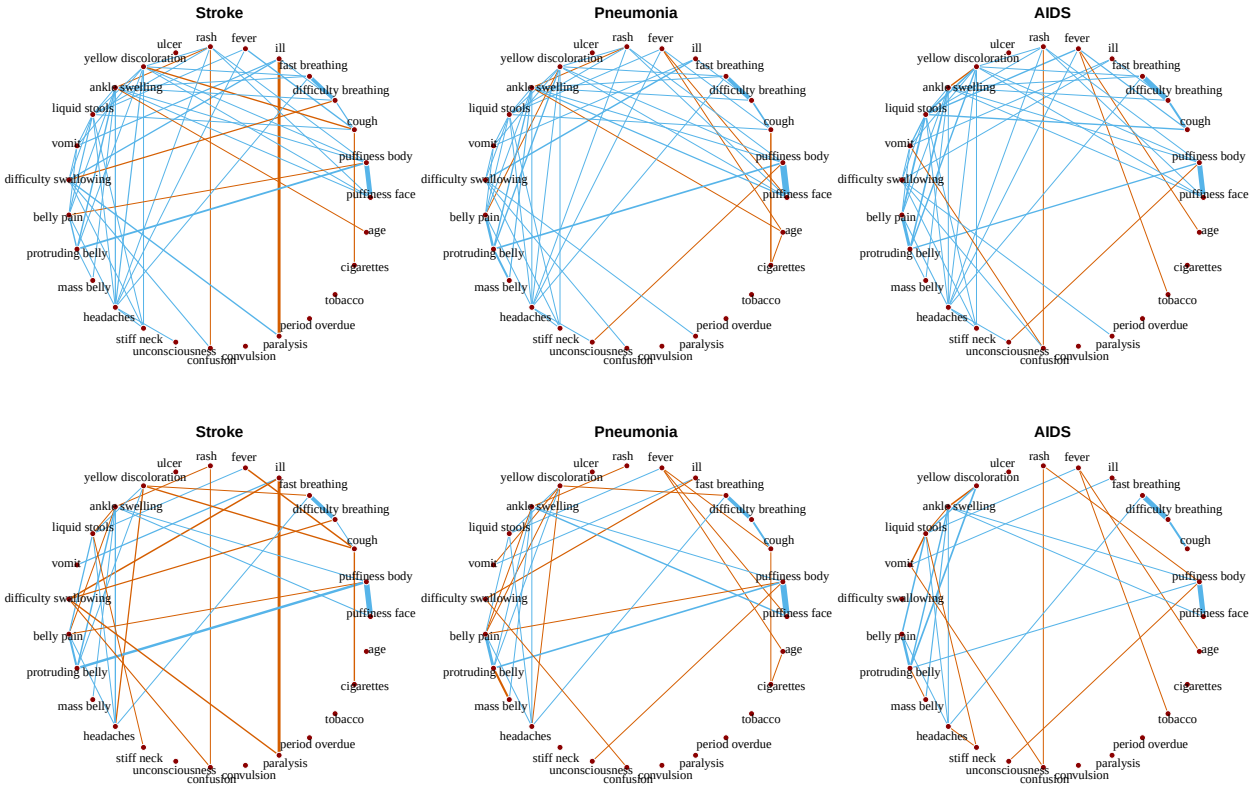


Figure 5.3: Estimated edges between the symptoms under the three causes using DSS-FGL (top row) and DSS-GGL (bottom row). The width of the edges are proportional to the size of $|\omega_{jk}^{(g)}|$. Common edges across all groups are colored in blue, and the differential edges are colored in red.

Carter model (Lee and Carter, 1992), typically use dimension reduction techniques to estimate mean effects due to age and time, and consider the residuals as independent measurement errors. However, residuals from such models are usually still highly correlated (Fosdick and Hoff, 2014). We consider estimating the residual structure with the 1×1 gender-specific mortality table up to age 100 in the US over the period of 1960 to 2010 using data obtained from the Human Mortality Database (HMD) (University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany)). For both the male and female mortality, we first randomly selected 25 years and remove 25 data points in each of those

years. We then fit a Lee-Carter model to estimate the mean model and interpolate the missing rates. Next, we estimate the covariance matrices among the 101 age groups in both genders using FGL and DSS-FGL from the residuals. The estimated residuals for the missing values can then be obtained by the E-step in our EM algorithm, or as the expectation from the conditional Gaussian distributions with covariance matrices estimated by FGL. The average mean squared errors (MSEs) for the prediction of missing log rates are summarized in Table 5.1. Imputation based on DSS-FGL precision matrix reduces the MSE by 27.8% compared to simple interpolation of the mean model (i.e., assuming i.i.d errors), compared to the 6.5% reduction from the FGL precision matrix with the same complexity. The estimated graphs are in the appendix.

	i.i.d	FGL	DSS-FGL
Average MSE	0.00372	0.00348	0.00268
Standard deviation of the MSEs	0.00030	0.00031	0.00028

Table 5.1: Average and standard deviation of the mean squared errors from 50 cross-validation experiments. The FGL model is selected to have the same number of edges as the DSS-FGL.

5.8 Discussion

In this chapter, we introduced a new class of priors for joint estimation of multiple graphical models. The proposed doubly spike-and-slab mixture priors, DSS-FGL and DSS-GGL, provide self-adaptive extensions to the joint graphical lasso penalties, and achieves simultaneous model selection and parameter estimation. Moreover, while taking advantage of the flexible class of penalty functions, the dynamic posterior exploration procedure allows the penalties to be adaptively estimated in a data-driven way, thus freeing practitioners from choosing multiple tunning parameters. This is especially useful in domains where sample sizes are too

small to reliably perform cross-validation. Finally, while not discussed in the main chapter, we note that the posterior uncertainty may be estimated using the Gibbs sampler described in the appendix.

The proposed framework can be extended in a few directions. First, we have assumed all classes to be exchangeable, as reflected in the penalty functions for the between-class similarity. When the classes exhibit hierarchical structures or different strengths of similarities, the indicator $\boldsymbol{\xi}$ may be modeled as functions of the class membership as well. Markov Random Field priors discussed in Saegusa and Shojaie (2016) and Peterson et al. (2015) may also be used to model the between-class similarities. Second, we have considered the estimation of multiple graphical models with missing values in the data matrices, it is also straightforward to extend to data with missing class labels, and latent Gaussian models as discussed in the previous chapters. In this way, the proposed methods can be extended to classification models (Hao et al., 2016) with mixed data. Finally, the proposed model is estimated using an EM algorithm that is iteratively solving the joint graphical lasso problem. It may be interesting to construct coordinate ascent algorithms that optimize on the objective function directly, similar to that described in (Ročková and George, 2018) for linear regression.

Chapter 6

DISCUSSION AND FUTURE WORK

In this dissertation, we considered different approaches to modeling the dependence structures among multivariate data using Gaussian graphical models and Gaussian copula graphical models. A common theme behind all the methods we developed is to estimate graphical models when data potentially contain limited observations, variables of mixed types, and many missing values.

The development of our methods was primarily motivated by the statistical challenges associated with modeling verbal autopsy (VA) data and probabilistic cause-of-death assignments using VA. Since this is the main driving force behind a major part of my research, I first discuss, within the scope of analyzing VA data, a summary of the strengths and limitations of the methods we developed in this dissertation.

The three methodological chapters each correspond to a different scenario researchers may face in practice with data collected through VA. In Chapter 3 we described a latent Gaussian model for mixed data with informative marginal priors. This method is most useful when training data is rare and difficult to obtain, while expert opinions can be collected at a lower cost. This chapter stems directly from the exiting literature of VA analysis where raw data consist of binary and continuous variables, and uses the same form of prior information currently used by other algorithms and software, including our previous work, InSilicoVA (McCormick et al., 2016). Thus it can be readily deployed to analyze current VA data. This is likely going to be useful to practitioners at the moment, as gold-standard training data is still sparse for VAs collected through the WHO questionnaire (World Health

Organization, 2012b; Nichols et al., 2018). However, several limitations exist. First, the heavy computation involved in posterior sampling requires relatively long running time even with high-performance hardware. Second, although it is the only option to perform cause-of-death assignment without training data, the use of expert opinions may still draw critics because of the difficulties in accurately soliciting such prior informations. And third, assuming symptoms share the same conditional dependence structures given any causes of death may be too restrictive in practice.

In light of these potential issues, Chapter 4 and 5 turn to the data-rich scenario and propose fast algorithms for learning cause-specific symptom associations with sufficient labeled data. In Chapter 4, we first laid the groundwork of EM-type algorithms for graphical model determination, and discussed strategies of learning the underlying dependence structures from a more general situation where the variables can be any of the binary, ordinal, or continuous types. Although not explicitly discussed in this chapter, the method can be extended to mixture of Gaussian copula models to perform cause-of-death assignment by estimating cause-specific precision matrices separately. We then further explored the joint estimation of multiple related graphical models in Chapter 5. This eventually allows us to achieve the several goals in modeling symptom associations simultaneously: acknowledging the heterogeneity among observations due to different causes of death, borrowing information across different causes of death, as well as relatively fast computation of the posterior modes. With our ongoing work of building a global VA data repository, we expect the methods we discussed in these two chapters will become more relevant in the short future. My ongoing work focuses on extending this framework to mixed data, which, I believe, is the most promising direction in the next stage of developing probabilistic cause-of-death assignment algorithm.

Outside of the scope of VA analysis, these methods can also be useful in other related

problems as we demonstrated in these chapters. Possible extensions for specific models have been discussed in the previous chapters. To conclude this dissertation, I would like to draw attention to four open questions that may have high impacts in this area:

1. We have focused on the methodological and computational aspects of developing spike-and-slab framework for sparse precision matrix estimation. Theoretical properties of these models need to be studied to better understand their behaviors, and provide more theoretical justification of such priors. Properties of the spike-and-slab lasso has been studied in Ročková (2018) for linear regression, but little do we know about precision matrix estimation.

2. All of the methods we discussed focus on obtaining a single summary of the precision matrices, either through posterior means or modes. It is most likely that the posterior surface contains multiple modes, and it will be interesting to explore multi-modal representations (Rocková, 2018) of the underlying precision matrices, which can potentially improve estimation and interpretability in many problems. This is also important for properly quantify uncertainty of the posterior estimates as well.

3. Variable selection in clustering and classification context have been studied extensively. For VA analysis, symptom reduction is highly needed in the practice (Kunihama et al., 2018), not only because it may improve cause-of-death assignments, but also since it can help public health officials to design questionnaires, as well as interviewers to prioritize on what questions to ask with limited resources.

4. Finally, in all our discussions of VA, we restricted our task to determining causes of death from a pre-defined list and using a single dataset. In the near future, however, researchers may have access to more training datasets collected at different locations,

under different conditions, and using different sampling schemes. It is also likely that these training datasets are to be coded by physicians at different levels of causes of death. This will present both challenges and opportunities for new methodologies to determine the optimal resolution of cause-of-death assignments given multiple sources of information.

BIBLIOGRAPHY

Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, 1998.

Jeffrey L Andrews and Paul D McNicholas. Variable selection for clustering and classification. *Journal of Classification*, 31(2):136–153, 2014.

Sayantan Banerjee and Subhashis Ghosal. Bayesian structure learning in graphical models. *Journal of Multivariate Analysis*, 136:147–162, 2015.

John Barnard, Robert McCulloch, and Xiao-Li Meng. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10 (4):1281–1311, 2000.

Laurent Briollais, Adrian Dobra, Jinnan Liu, Matt Friedlander, Hilmi Ozcelik, and Hélène Massam. A Bayesian graphical model for genome-wide association studies (GWAS). *The Annals of Applied Statistics*, 10(2):786–811, 2016.

Yunqi Bu and Johannes Lederer. Integrating additional knowledge into estimation of graphical models. *arXiv preprint arXiv:1704.02739*, 2017.

Z Butt, S Haberman, and HL Shang. *ilc: Lee-Carter Mortality Models using Iterative Fitting Algorithms*, 2014.

Peter Byass, Dao Lan Huong, and Hoang Van Minh. A probabilistic approach to interpreting verbal autopsies: methodology and preliminary validation in vietnam. *Scandinavian Journal of Public Health*, 31(62 suppl):32–37, 2003.

Peter Byass, Daniel Chandramohan, Samuel J Clark, Lucia D'Ambruoso, Edward Fottrell, Wendy J Graham, Abraham J Herbst, Abraham Hodgson, Sennen Hounton, and Kathleen Kahn. Strengthening standardised interpretation of verbal autopsy data: The new InterVA-4 tool. *Global Health Action*, 5, 2012.

T. Tony Cai, Cun H. Zhang, and Harrison H. Zhou. Optimal rates of convergence for covariance matrix estimation. *Annals of Statistics*, 38(4):2118–2144, 2010.

Amelia C Crampin, Albert Dube, Sebastian Mboma, Alison Price, Menard Chihana, Andreas Jahn, Angela Baschieri, Anna Molesworth, Elnaeus Mwaiyeghele, and Keith Branson. Profile: the Karonga health and demographic surveillance system. *International Journal of Epidemiology*, 41(3):676–685, 2012.

Patrick Danaher, Pei Wang, and Daniela M Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):373–397, 2014.

A. P. Dawid and S. L. Lauritzen. Hyper Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, 21(3):1272–1317, 09 1993.

Bernard Delyon, Marc Lavielle, and Eric Moulines. Convergence of a stochastic approximation version of the EM algorithm. *Annals of Statistics*, 27(1):94–128, 03 1999.

Sameer K Deshpande, Veronika Ročková, and Edward I George. Simultaneous variable and covariance selection with the multivariate spike-and-slab lasso. *arXiv preprint arXiv:1708.08911*, 2017.

Adrian Dobra. Graphical modeling of spatial health data. *arXiv preprint arXiv:1411.6512*, 2014.

Adrian Dobra and Alex Lenkoski. Copula Gaussian graphical models and their application to modeling functional disability data. *The Annals of Applied Statistics*, 5(2A):969–993, 2011.

Adrian Dobra, Theo S Eicher, and Alex Lenkoski. Modeling uncertainty in macroeconomic growth determinants using Gaussian graphical models. *Statistical Methodology*, 7(3): 292–306, 2010.

Adrian Dobra, Alex Lenkoski, and Abel Rodriguez. Bayesian inference for general Gaussian graphical models with application to multivariate lattice data. *Journal of the American Statistical Association*, 106(496):1418–1433, 2011.

Jianqing Fan, Yang Feng, and Yichao Wu. Network exploration via the adaptive lasso and scad penalties. *The Annals of Applied Statistics*, 3(2):521, 2009.

Jianqing Fan, Han Liu, Yang Ning, and Hui Zou. High dimensional semiparametric latent graphical model for mixed data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016.

Bailey K Fosdick and Peter D Hoff. Separable factor analysis with applications to mortality data. *The Annals of Applied Statistics*, 8(1):120, 2014.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008. ISSN 14654644.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Applications of the lasso and grouped lasso to the estimation of sparse graphical models. *Technical Report*, 2010.

Michel Garenne. Prospects for automated diagnosis of verbal autopsies. *BMC Medicine*, 12 (1):18, 2014.

Andrew Gelman. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515–534, 2006.

Edward I George and Robert E McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

Paolo Giudici and Peter J Green. Decomposable graphical Gaussian model determination. *Biometrika*, 86(4):785–801, 1999.

Jian Guo, Elizaveta Levina, George Michailidis, and Ji Zhu. Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15, 2011.

P Richard Hahn and Carlos M Carvalho. Decoupling shrinkage and selection in Bayesian linear models: a posterior summary perspective. *Journal of the American Statistical Association*, 110(509):435–448, 2015.

Chris Hans, Adrian Dobra, and Mike West. Shotgun Stochastic Search for "large p" regression. *Journal of the American Statistical Association*, 102(478):507–516, 2007.

Botao Hao, Will Wei Sun, Yufeng Liu, and Guang Cheng. Simultaneous clustering and estimation of heterogeneous graphical models. *arXiv preprint arXiv:1611.09391*, 2016.

Peter D Hoff. Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics*, pages 265–283, 2007.

Richard Horton. Counting for health. *Lancet*, 370(9598):1526–1526, 2007.

Hemant Ishwaran and J Sunil Rao. Detecting differentially expressed genes in microarrays using bayesian model selection. *Journal of the American Statistical Association*, 98(462): 438–455, 2003.

Hemant Ishwaran and J Sunil Rao. Spike and slab variable selection: frequentist and Bayesian strategies. *Annals of statistics*, pages 730–773, 2005.

Hemant Ishwaran and J Sunil Rao. Consistency of spike and slab regression. *Statistics & Probability Letters*, 81(12):1920–1928, 2011.

S. L. James, A. D. Flaxman, C. J. Murray, and Consortium Population Health Metrics Research. Performance of the tariff method: validation of a simple additive algorithm for analysis of verbal autopsies. *Population Health Metrics*, 9(31), 2011.

Prabhat Jha. Reliable direct measurement of causes of death in low-and middle-income countries. *BMC medicine*, 12(1):19, 2014.

Beatrix Jones, Carlos Carvalho, Adrian Dobra, Chris Hans, Chris Carter, and Mike West. Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science*, pages 388–400, 2005.

Kathleen Kahn, Mark A Collinson, F Xavier Gómez-Olivé, Obed Mokoena, Rhian Twine, Paul Mee, Sulaimon A Afolabi, Benjamin D Clark, Chodziwadziwa W Kabudula, and Audrey Khosa. Profile: Agincourt health and socio-demographic surveillance system. *International Journal of Epidemiology*, 41(4):988–1001, 2012.

G. King and Y. Lu. Verbal autopsy methods with multiple causes of death. *Statistical Science*, 100(469), 2008.

Chris AJ Klaassen and Jon A Wellner. Efficient estimation in the bivariate normal copula model: normal margins are least favourable. *Bernoulli*, 3(1):55–77, 1997.

Tsuyoshi Kunihama, Zehang R Li, Samuel J Clark, and Tyler H McCormick. Bayesian factor

models for probabilistic cause of death assessment with verbal autopsies. *arXiv preprint arXiv:1803.01327*, 2018.

Minjung Kyung, Jeff Gilly, Malay Ghoshz, and George Casellax. Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, 5(2):369–412, 2010.

Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

Ronald D Lee and Lawrence R Carter. Modeling and forecasting us mortality. *Journal of the American statistical association*, 87(419):659–671, 1992.

Alex Lenkoski and Adrian Dobra. Computational aspects related to inference in Gaussian graphical models with the G-Wishart prior. *Journal of Computational and Graphical Statistics*, 20(1):140–157, 2011.

Richard A Levine and George Casella. Implementations of the Monte Carlo EM algorithm. *Journal of Computational and Graphical Statistics*, 10(3):422–439, 2001.

Yunfan Li, Bruce A Craig, and Anindya Bhadra. The graphical horseshoe estimator for inverse covariance matrices. *arXiv preprint arXiv:1707.06661*, 2017a.

Zehang R Li and Tyler H McCormick. An Expectation Conditional Maximization approach for Gaussian graphical models. *arXiv preprint arXiv:1709.06970*, 2017.

Zehang R Li, Tyler H McCormick, and Samuel J Clark. Bayesian inference of latent Gaussian graphical models for mixed data. *arXiv preprint arXiv:1711.00877*, 2017b.

Zehang R Li, Jason Thomas, Tyler H McCormick, and Samuel J Clark. The openVA toolkit for Verbal Autopsies. Technical report, 2017c. URL http://openva.net/openVA_files/opeVA-vignette.pdf.

Zehang R Li, Tyler H McCormick, and Samuel J Clark. Bayesian joint spike-and-slab graphical lasso. *arXiv preprint arXiv:1805.07051*, 2018.

Zhixiang Lin, Tao Wang, Can Yang, and Hongyu Zhao. On joint estimation of Gaussian graphical models for spatial and temporal data. *Biometrics*, 73(3):769–779, 2017.

Fei Liu, Sounak Chakraborty, Fan Li, Yan Liu, and Aurelie C Lozano. Bayesian regularization via graph laplacian. *Bayesian Analysis*, 9(2):449–474, 2014.

Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10:2295–2328, 2009.

Han Liu, Kathryn Roeder, and Larry Wasserman. Stability approach to regularization selection (StARS) for high dimensional graphical models. In *Advances in Neural Information Processing Systems*, pages 1432–1440, 2010.

Han Liu, Fang Han, Ming Yuan, John Lafferty, and Larry Wasserman. High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics*, 40(4): 2293–2326, 2012.

Jun S. Liu and Ying Nian Wu. Parameter expansion for data augmentation. *Journal of the American Statistical Association*, 94(448):1264–1274, 1999.

Jing Ma and George Michailidis. Joint structural estimation of multiple graphical models. *Journal of Machine Learning Research*, 17(166):1–48, 2016.

Rahul Mazumder and Trevor Hastie. Exact covariance thresholding into connected components for large-scale graphical lasso. *Journal of Machine Learning Research*, 13 (Mar):781–794, 2012a.

Rahul Mazumder and Trevor Hastie. The graphical lasso: New insights and alternatives. *Electronic Journal of Statistics*, 6:2125, 2012b.

Tyler H McCormick, Zehang Richard Li, Clara Calvert, Amelia C Crampin, Kathleen Kahn, and Samuel J Clark. Probabilistic cause-of-death assignment using verbal autopsies. *Journal of the American Statistical Association*, 111(515):1036–1049, 2016.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, pages 1436–1462, 2006.

Xiao-Li Meng and Donald B Rubin. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2):267–278, 1993.

Xiao-Li Meng and David A Van Dyk. Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika*, 86(2):301–320, 1999.

Pierre Miasnikof, Vasily Giannakeas, Mireille Gomes, Lukasz Aleksandrowicz, Alexander Y Shestopaloff, Dewan Alam, Stephen Tollman, Akram Samarikhalaj, and Prabhat Jha. Naive bayes classifiers for verbal autopsies: comparison to physician-based classification for 21,000 child and adult deaths. *BMC medicine*, 13(1):1, 2015.

Abdolreza Mohammadi and Ernst C Wit. Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Analysis*, 10(1):109–138, 2015a.

Abdolreza Mohammadi and Ernst C Wit. BDgraph: An R package for Bayesian structure learning in graphical models. *arXiv preprint arXiv:1501.05108*, 2015b.

Abdolreza Mohammadi, Fentaw Abegaz, Edwin van den Heuvel, and Ernst C. Wit. Bayesian modelling of Dupuytren disease by using Gaussian copula graphical models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(3):629–645, 2017.

Christopher JL Murray, Alan D Lopez, Robert Black, Ramesh Ahuja, Said M Ali, Abdullah Baqui, Lalit Dandona, Emily Dantzer, Vinita Das, and Usha Dhingra. Population health metrics research consortium gold standard verbal autopsy validation study: design, implementation, and development of analysis datasets. *Population Health Metrics*, 9(1):27, 2011a.

Christopher JL Murray, Rafael Lozano, Abraham D Flaxman, Alireza Vahdatpour, and Alan D Lopez. Robust metrics for assessing the performance of different verbal autopsy cause assignment methods in validation studies. *Population Health Metrics*, 9(1):28, 2011b.

Iain Murray, Ryan Prescott Adams, and David JC MacKay. Elliptical slice sampling. In *AISTATS*, volume 13, pages 541–548, 2010.

Roger B Nelsen. An introduction to copulas, volume 139 of Lecture Notes in Statistics, 1999.

Erin K Nichols, Peter Byass, Daniel Chandramohan, Samuel J Clark, Abraham D Flaxman, Robert Jakob, Jordana Leitao, Nicolas Maire, Chalapati Rao, Ian Riley, and Philip W Setel. The who 2016 verbal autopsy instrument: An international standard suitable for automated analysis by interva, insilicova, and tariff 2.0. *PLoS medicine*, 15(1):e1002486, 2018.

Søren Feodor Nielsen. The stochastic EM algorithm: estimation and asymptotic results. *Bernoulli*, 6(3):457–489, 2000.

Jie Peng, Pei Wang, Nengfeng Zhou, and Ji Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486): 735–746, 2009.

Christine Peterson, Marina Vannucci, Cemal Karakas, William Choi, Lihua Ma, and Mirjana

Meletić-Savatić. Inferring metabolic networks using the Bayesian adaptive graphical lasso with informative priors. *Statistics and its Interface*, 6(4):547, 2013.

Christine Peterson, Francesco Stingo, and Marina Vannucci. Bayesian inference of multiple Gaussian graphical models. *Journal of the American Statistical Association*, 110(509): 159–174, 2015.

Michael Pitt, David Chan, and Robert Kohn. Efficient Bayesian inference for Gaussian copula regression models. *Biometrika*, 93(3):537–554, 2006.

Veronika Rocková. Particle EM for variable selection. *Journal of the American Statistical Association*, 0(0):1–14, 2018. doi: 10.1080/01621459.2017.1360778.

Veronika Ročková. Bayesian estimation of sparse signals with a continuous spike-and-slab prior. *The Annals of Statistics*, 46(1):401–437, 02 2018.

Veronika Ročková and Edward I George. EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846, 2014.

Veronika Ročková and Edward I George. The spike-and-slab lasso. *Journal of the American Statistical Association*, 113(521):431–444, 2018.

Adam J Rothman, Peter J Bickel, Elizaveta Levina, and Ji Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

Alberto Roverato. Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for gaussian graphical models. *Scandinavian Journal of Statistics*, 29(3):391–411, 2002.

Takumi Saegusa and Ali Shojaie. Joint estimation of precision matrices in heterogeneous populations. *Electronic Journal of Statistics*, 10(1):1341, 2016.

Tracy A Schifeling and Jerome P Reiter. Incorporating marginal prior information in latent class models. *Bayesian Analysis*, 11(2):499–518, 2016.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464, 1978.

Peter Serina, Ian Riley, Andrea Stewart, Abraham D Flaxman, Rafael Lozano, Meghan D Mooney, Richard Luning, Bernardo Hernandez, Robert Black, and Ramesh Ahuja. A shortened verbal autopsy instrument for use in routine mortality surveillance systems. *BMC medicine*, 13(1):1, 2015.

Aline Talhouk, Arnaud Doucet, and Kevin Murphy. Efficient Bayesian inference for multivariate probit models with sparse inverse correlation matrices. *Journal of Computational and Graphical Statistics*, 21(February 2015):739–757, 2012.

Rajesh Talluri, Veerabhadran Baladandayuthapani, and Bani K Mallick. Bayesian sparse graphical models and their mixtures. *Stat*, 3(1):109–125, mar 2014.

Linda SL Tan, Ajay Jasra, Maria De Iorio, and Timothy MD Ebbels. Bayesian inference for multiple Gaussian graphical models with application to metabolic association networks. *The Annals of Applied Statistics*, 11(4):2222–2251, 2017.

Carl E Taylor, RL Parker, WA Reinke, and R Faruquee. *Child and maternal health services in rural India: The Narangwal experiment*. Johns Hopkins University Press, 1983.

University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Human Mortality Database.

Jon Wakefield, Frank De Vocht, and Rayjean J Hung. Bayesian mixture modeling of gene-environment and gene-gene interactions. *Genetic Epidemiology*, 34(1):16–25, 2010.

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

Hao Wang. Scaling it up: Stochastic search structure learning in graphical models. *Bayesian Analysis*, 10(2):351–377, 2015.

Hao Wang and Sophia Zhengzi Li. Efficient Gaussian graphical model determination under G-wishart prior distributions. *Electronic Journal of Statistics*, 6:168–198, 2012.

Greg CG Wei and Martin A Tanner. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American statistical Association*, 85(411):699–704, 1990.

Daniela M Witten, Jerome H Friedman, and Noah Simon. New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics*, 20(4):892–900, 2011.

World Health Organization. Verbal autopsy standards: ascertaining and attributing causes of death. http://www.who.int/healthinfo/statistics/verbalautopsystandards/en/, 2012a.

World Health Organization. *Verbal Autopsy Standards: The 2012 WHO verbal autopsy instrument*, accessed 2018-02 2012b.

Xiaofan Xu and Malay Ghosh. Bayesian variable selection and estimation for group lasso. *Bayesian Analysis*, 10(4):909–936, 2015.

Lingzhou Xue and Hui Zou. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *The Annals of Statistics*, 40(5):2541–2571, 2012.

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

Lin Zhang, Veerabhadran Baladandayuthapani, Bani K Mallick, Ganiraju C Manyam, Patricia A Thompson, Melissa L Bondy, and Kim-Anh Do. Bayesian hierarchical structured variable selection methods with application to molecular inversion probe studies in breast cancer. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63 (4):595–620, 2014.

Tuo Zhao, Han Liu, Kathryn Roeder, John Lafferty, and Larry Wasserman. The huge package for high-dimensional undirected graph estimation in R. *Journal of Machine Learning Research*, 13(Apr):1059–1062, 2012.

Appendix A

APPENDIX FOR CHAPTER 3

A.1 Derivation of the spike-and-slab prior

The proposed prior distribution for $\boldsymbol{R}$ is

$$p(\boldsymbol{R}|\delta) \propto C_{\boldsymbol{\delta}}^{-1}|\boldsymbol{R}|^{-(p+1)}\prod_{j<k}\exp(-(r^{jk})^2/2v_{\delta_{jk}}^2)\prod_{j}\exp(-\lambda r^{jj}/2)$$

First we show that $C_{\boldsymbol{\delta}} < \infty$ so that the prior distribution is proper. We note

$$\begin{aligned} C_{\boldsymbol{\delta}} &= C\int_{R^+}|\boldsymbol{R}|^{-(p+1)}\prod_{j<k}\exp(-(r^{jk})^2/2v_{\delta_{jk}}^2)\prod_{j}\exp(-\lambda r^{jj}/2)d\boldsymbol{R} \\ &\leq C\int_{R^+}|\boldsymbol{R}|^{-(p+1)}\prod_{j}\exp(-\lambda r^{jj}/2)d\boldsymbol{R} \\ &= C\int_{R^+}|\boldsymbol{R}|^{-(p+1)}\prod_{j}(r^{jj})^{-\frac{p+1}{2}}\prod_{j}\exp(-\lambda r^{jj}/2+\frac{p+1}{2}\log(r^{jj}))d\boldsymbol{R} \end{aligned}$$

Since $\exp(-\lambda r^{jj}/2+\frac{p+1}{2}\log(r^{jj}))$ is a non-negative function of r^{jj}, and has a global maximum at $r^{jj} = (p+1)/\lambda$, and C is a positive constant, we have

$$C_{\boldsymbol{\delta}} \leq C'\int_{R^+}|\boldsymbol{R}|^{-(p+1)}\prod_{j}(r^{jj})^{-\frac{p+1}{2}}d\boldsymbol{R},$$

where the constant $C' < \infty$, and $\int_{R^+}|\boldsymbol{R}|^{-(p+1)}\prod_j(r^{jj})^{-\frac{p+1}{2}}d\boldsymbol{R} < \infty$ as well since it is proportional to the marginally uniform prior of $\boldsymbol{R}$ derived from the Wishart distribution. Therefore the normalizing constant $C_\delta < \infty$, and the prior is proper.

In order to obtain the prior distribution on the expanded precision matrix $\boldsymbol{\Omega} = (\boldsymbol{DRD})^{-1}$, we put prior on the marginal expansion parameter $\boldsymbol{D}$ with a prior distribution so that $p(d_j^2|\boldsymbol{R})$ is an inverse Gamma distribution with shape and rate parameter being $((p+1)/2, 1/2)$, we have

$$p(\boldsymbol{D}|\boldsymbol{R}) \propto \prod_j d_j^{-(p+2)} \exp(\frac{1}{2d_j^2})$$

Since $r^{jk} = \omega_{jk} d_j d_k$, we can derive

$$\begin{aligned} p(\boldsymbol{\Omega}|\boldsymbol{\delta}) &= p(\boldsymbol{R}|\delta) p(\boldsymbol{D}|\boldsymbol{R}) |\mathcal{J}| \\ &\propto C_{\boldsymbol{\delta}}^{-1} |\boldsymbol{R}|^{-(p+1)} \prod_{j<k} \exp(-(r^{jk})^2/2v_{\delta_{jk}}^2) \prod_j \exp(-\lambda r^{jj}/2) \prod_j d_j^{-(p+2)} \exp(\frac{1}{2d_j^2}) |\boldsymbol{R}|^{p+1} \prod d_j^{p+2} \\ &= C_{\boldsymbol{\delta}}^{-1} \prod_{j<k} \exp(-\frac{\omega_{jk}^2}{2v_{\delta_{jk}}^2/d_j^2 d_k^2}) \prod_j \exp(-\frac{\lambda d_j^2}{2}\omega_{jj}) \prod_j \exp(\frac{1}{2d_j^2}), \end{aligned}$$

where $d_j = \sigma_j$ is the square root of the k-th diagonal element of $\boldsymbol{\Sigma} = \boldsymbol{\Omega}^{-1}$, i.e.,

$$p(\boldsymbol{\Omega}|\boldsymbol{\delta}) \propto C_{\boldsymbol{\delta}}^{-1} \prod_{j<k} \exp(-\frac{\omega_{jk}^2}{2v_{\delta_{jk}}^2/\sigma_j^2\sigma_k^2}) \prod_j \exp(-\frac{\lambda\sigma_j^2}{2}\omega_{jj}) \prod_j \exp(\frac{1}{2\sigma_j^2})$$

A.2 Implied prior sparsity with different hyperparameters

In this section, we provide more prior simulation results to facilitate the choice of λ, v_0, v_1, and $\pi_{\boldsymbol{\delta}}$. Figure A.1 illustrates our approach in understanding how these 4 parameters jointly imply the prior sparsity. It can be seen that small λ and extremely small v_0 usually leads to denser prior graph unless v_1 is also small, which defeats the purpose of using the continuous mixture prior. We choose to use $\lambda = 10$, $v_0 = 0.01$, $v_1/v_0 = 100$, and $\pi_{\boldsymbol{\delta}} = 0.0001$ in our experiments. In general, for the prior edge probability to be calibrated between 0.05 to 0.2, we believe the model is not very sensitive to parameters in the close range to our choices.

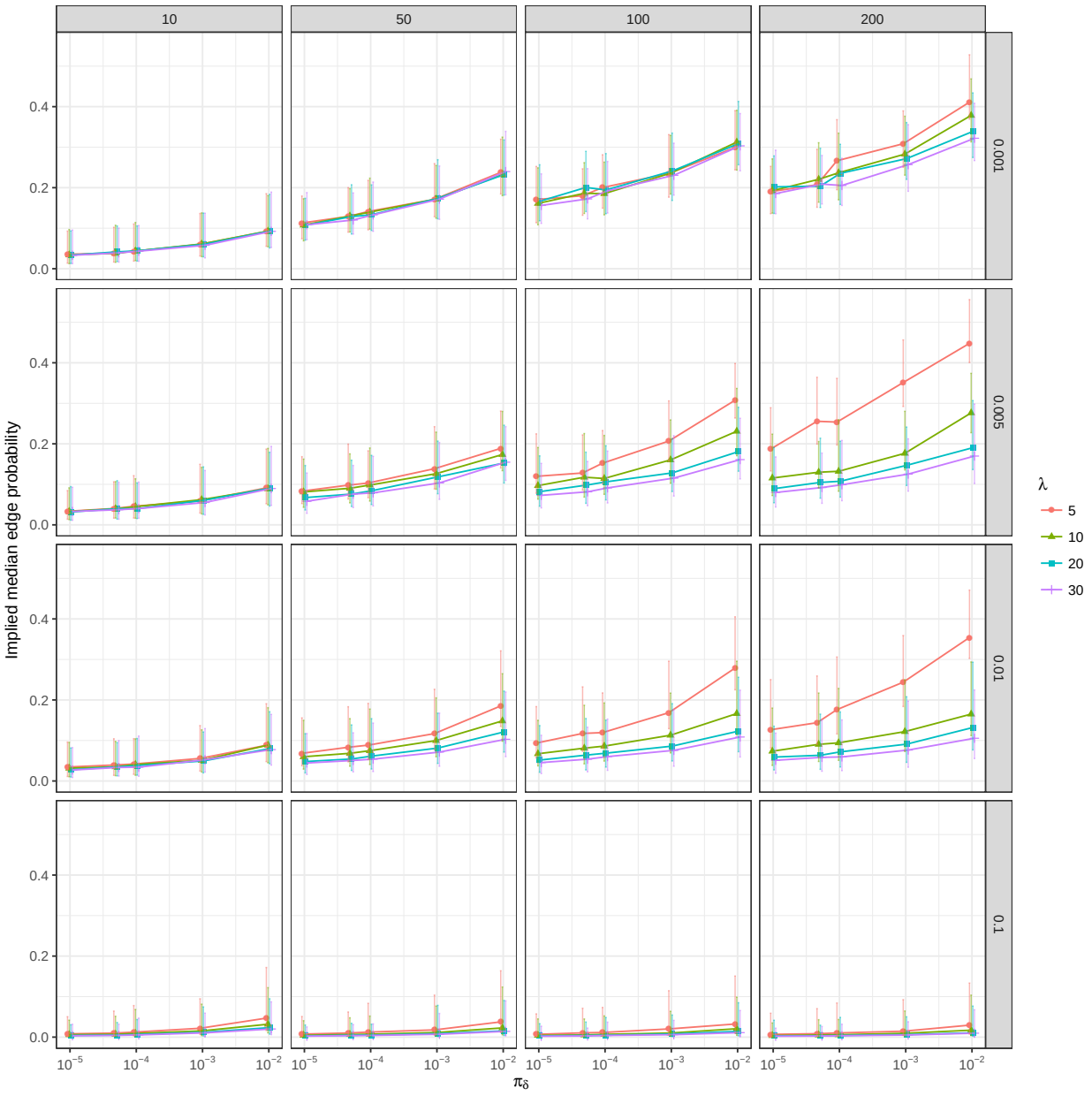


Figure A.1: Implied prior edge probability for $p = 100$ graph. The dots represent the median prior probabilities and the error bars represent the 0.025 and 0.975 quantiles The rows in the panel represent the value of v_0, and the columns represent the choice of v_1/v_0. For each combination of v_0 and v_1, the edge probabilities induced by different λ and $\pi_{\boldsymbol{\delta}}$ are plotted. The densities are derived from sampling $1,000$ draws using MCMC from the prior distribution after $1,000$ iterations of burn-in.

A.3 Posterior sampling for the classification model

This section describes the inference procedure for the model presented in Section 3.2 of the main paper. The steps are mostly similar to Section 3.3.2 of the paper.

Update $\boldsymbol{Z}$ *and* $\boldsymbol{\Lambda}$. This first two steps are the same as in Section 3.3.2 of the main paper, except replacing $\boldsymbol{\mu}$ to the corresponding $\boldsymbol{\mu}_c$.

Update $\boldsymbol{\mu}$. The conditional posterior distribution for the mean parameters is also multivariate normal,

$$\boldsymbol{\mu}_c|\boldsymbol{Y}, \tilde{\boldsymbol{R}}, \boldsymbol{X} \sim \text{Normal}\left((\frac{1}{\sigma^2}\boldsymbol{I}_p + n_c\tilde{\boldsymbol{R}}^{-1})^{-1}(\frac{1}{\sigma^2}\boldsymbol{\mu}_{0c} + n_c\tilde{\boldsymbol{R}}^{-1}\bar{z}_c), (\frac{1}{\sigma^2}\boldsymbol{I}_p + n_c\tilde{\boldsymbol{R}}^{-1})^{-1}\right)$$

where $n_c = \sum_i \mathbf{1}_{y_i=c}$ and $\bar{z}_c = \sum_{i:y_i=c} \boldsymbol{Z}_i$.

Update $\boldsymbol{R}$. To update the latent correlation matrix, we first draw the working expansion and expand the observations in the same way as Section 3.3.2 of the main paper. The rescaled sample covariance matrix is $\boldsymbol{S} = \sum_{i=1}^{n}(W_i - \boldsymbol{D}\boldsymbol{\mu}_{y_i})'\boldsymbol{\Lambda}^{-2}(W_i - \boldsymbol{D}\boldsymbol{\mu}_{y_i})$. The rest of the sampling steps are the same.

Update π. The conditional distribution of π is still Dirichlet:

$$\pi \sim \text{Dirichlet}(\alpha_1 + n_1, ..., \alpha_c + n_c) \ .$$

Update $\boldsymbol{Y}$. This step can be performed with Equation (3.3) in the main paper, or by integrating out π, so that

$$p(Y_i = c|\boldsymbol{Z}_i, \boldsymbol{Y}_{-i}, \boldsymbol{\mu}, \tilde{\boldsymbol{R}}) \propto \frac{n_{-i,c} + \alpha_c}{n - 1 + \sum_c \alpha_c}\phi(\boldsymbol{Z}_i; \boldsymbol{\mu}_c, \tilde{\boldsymbol{R}})$$

where $n_{-i,c} = \sum_{i' \neq i} \mathbf{1}_{Y_{i'}=c}$.

(optional) Update σ_c^2. When σ_c^2 is not fixed in the model, we can sample them from the conjugate posterior distribution

$$\sigma_c^2 \sim \text{InvGamma}(0.001 + \frac{p}{2}, 0.001 + \frac{\sum_{j=1}^{p}(\mu_{cj} - \mu_{0cj})^2}{2}) \ .$$

Appendix B

APPENDIX FOR CHAPTER 5

sectionProof of Proposition 1 and 2 Here we provide a proof of Proposition 2. The same arguments generalize to Proposition 1 directly by fixing all binary indicators to 1 and thus are not repeated.

Proof of DSS-GGL prior We first consider the GGL and DSS-GGL penalties. The joint distribution of all the parameters under the parameterization of the scale mixture of Normal distributions is

$$\begin{aligned} p(\{\boldsymbol{\Omega}\}, \boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\delta}, \boldsymbol{\xi}, \pi_{\boldsymbol{\delta}}, \pi_{\boldsymbol{\xi}}) &= \prod_g \prod_{j<k} \exp(-\frac{1}{2}(\omega_{jk}^{(g)})^2(\frac{v_{\delta_{jk}}}{\tau_{jkg}} + \frac{v_{\xi^*_{jk}}}{\rho_{jk}})) \prod_g \prod_j \exp(-\frac{\lambda_0}{2}\omega_{jj}^{(g)}) \\ &\times \prod_g \prod_{j<k} \tau_{jkg}^{-\frac{1}{2}} \exp(-\frac{\lambda_1^2}{2}\tau_{jkg}) \prod_{j<k} \rho_{jk}^{-\frac{1}{2}} \exp(-\frac{\lambda_2^2}{2}\rho_{jk}) \\ &\times \prod_{j<k} \pi_{\boldsymbol{\delta}}^{\delta_{jk}}(1-\pi_{\boldsymbol{\delta}})^{1-\delta_{jk}} \prod_{j<k} \pi_{\boldsymbol{\xi}}^{\xi_{jk}}(1-\pi_{\boldsymbol{\xi}})^{1-\xi_{jk}} \\ &\times \prod_{j<k} \pi_{\boldsymbol{\delta}}^{a_1-1}(1-\pi_{\boldsymbol{\delta}})^{b_1-1} \prod_{j<k} \pi_{\boldsymbol{\xi}}^{a_2-1}(1-\pi_{\boldsymbol{\xi}})^{b_2-1} \end{aligned}$$

The following two identities provide the key steps to connect the scale mixture of normal distributions to the Laplace representation in the penalty function:

$$\int_0^\infty \frac{1}{\sqrt{2\pi s}} \exp(-\frac{z^2}{2s})\frac{\lambda^2}{2}\exp(-\frac{\lambda^2 s}{2})ds = \frac{\lambda}{2}\exp(-\lambda|z|) \quad \text{(B.1)}$$

$$\int_0^\infty \frac{1}{\sqrt{2\pi s}} \exp(-\frac{||\boldsymbol{z}||_2}{2s})(\frac{\lambda^2}{2})^{\frac{p+1}{2}}\exp(-\frac{\lambda^2 s}{2})s^{\frac{p+1}{2}-1}\frac{1}{\Gamma(\frac{p+1}{2})}ds = \frac{\lambda}{2}\exp(-\lambda||\boldsymbol{z}||_2) \quad \text{(B.2)}$$

where $\boldsymbol{z}$ is a vector of length p and $||z||_2 = \sqrt{\sum_{i=1}^{p} z_i^2}$. By rearranging the terms and plugging in the two identities above, it can be seen that

$$p(\{\boldsymbol{\Omega}\}|\boldsymbol{\delta},\boldsymbol{\xi}) \;\propto\; \exp(-\sum_{j<k}\frac{\lambda_1}{v_{\delta_{jk}}}|\omega_{jk}^{(g)}| - \sum_{j<k}\frac{\lambda_2}{v_{\xi_{jk}^*}}||\boldsymbol{\omega}_{jk}||_2 - \sum_{g}\sum_{j}\frac{\lambda_0}{2}\omega_{jj}^{(g)}).$$

The conditional distribution of $\{\boldsymbol{\Omega}\}|\boldsymbol{\delta},\boldsymbol{\xi}$ takes the form of $\exp(-pen(\{\boldsymbol{\Omega}\}|\boldsymbol{\delta},\boldsymbol{\xi}))$ for DSS-GGL, and thus the mode of the posterior is equivalent to the DSS-GGL solution. It still remains to be seen that the intractable constant terms are all finite, so that each of the three conditional distributions are proper. This can be seen as follows:

$$\begin{aligned}
C_{\boldsymbol{\tau},\boldsymbol{\rho}}C_{\boldsymbol{\delta},\boldsymbol{\xi}} &= \int\prod_{j<k}\text{Normal}(\boldsymbol{\omega}_{jk};0,\boldsymbol{\Theta})\prod_{g}\prod_{j}\text{Exp}(\omega_{jj}^{(g)};\frac{\lambda_0}{2})\mathbf{1}_{\{\boldsymbol{\Omega}\}\in M^+}d\{\boldsymbol{\Omega}\}\\
&< \int\prod_{j<k}\text{Normal}(\boldsymbol{\omega}_{jk};0,\boldsymbol{\Theta})\prod_{g}\prod_{j}\text{Exp}(\omega_{jj}^{(g)};\frac{\lambda_0}{2})d\{\boldsymbol{\Omega}\} = 1\\
C_{\boldsymbol{\tau},\boldsymbol{\rho}}^{-1} &\propto \int\prod_{j<k}\big(\exp(-\frac{\lambda_1^2}{2}\sum_{g}\tau_{jkg} - \frac{\lambda_2^2}{2}\rho_{jk})\rho_{jk}^{-\frac{1}{2}}\prod_{g}(\tau_{jkg}(\frac{1}{\tau_{jkg}}+\frac{1}{\rho_{jk}}))^{-\frac{1}{2}}\big)d\{\boldsymbol{\tau}_{jk},\rho_{jk}\}\\
&< \int\prod_{j<k}\big(\exp(-\frac{\lambda_1^2}{2}\sum_{g}\tau_{jkg} - \frac{\lambda_2^2}{2}\rho_{jk})\rho_{jk}^{-\frac{1}{2}}\big)d\rho_{jk}\\
&< \int\prod_{j<k}\big(\exp(-\frac{\lambda_2^2}{2}\rho_{jk})\rho_{jk}^{-\frac{1}{2}}\big)d\rho_{jk} = (2\pi/\lambda_2^2)^{\frac{p(p-1)}{4}}
\end{aligned}$$

The above inequalities completes the proof that the conditional prior distributions are all proper for DSS-GGL. For GGL prior, the proof is essentially the same by fixing $\boldsymbol{\delta}$ and $\boldsymbol{\xi}$ to be 1.

Proof of DSS-FGL prior For DSS-FGL, the joint distribution of all the parameters using the scale Normal mixture representation is

$$\begin{aligned} p(\{\boldsymbol{\Omega}\}, \boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\delta}, \boldsymbol{\xi}, \pi_{\boldsymbol{\delta}}, \pi_{\boldsymbol{\xi}}) &= \prod_g \prod_{j<k} \exp(-\frac{1}{2}(\boldsymbol{\omega}_{jk}^{(g)})^T \boldsymbol{\Theta}_{jk} \boldsymbol{\omega}_{jk}^{(g)}) \prod_g \prod_j \exp(-\frac{\lambda_0}{2}\omega_{jj}^{(g)}) \\ &\times \prod_g \prod_{j<k} \tau_{jkg}^{-\frac{1}{2}} \exp(-\frac{\lambda_1^2}{2}\tau_{jkg}) \prod_{j<k} \rho_{jk}^{-\frac{1}{2}} \exp(-\frac{\lambda_2^2}{2}\rho_{jk}) \\ &\times \prod_{j<k} \pi_{\boldsymbol{\delta}}^{\delta_{jk}} (1-\pi_{\boldsymbol{\delta}})^{1-\delta_{jk}} \prod_{j<k} \pi_{\boldsymbol{\xi}}^{\xi_{jk}} (1-\pi_{\boldsymbol{\xi}})^{1-\xi_{jk}} \\ &\times \prod_{j<k} \pi_{\boldsymbol{\delta}}^{a_1-1} (1-\pi_{\boldsymbol{\delta}})^{b_1-1} \prod_{j<k} \pi_{\boldsymbol{\xi}}^{a_2-1} (1-\pi_{\boldsymbol{\xi}})^{b_2-1} \end{aligned}$$

where the first term can be rewritten as the same form as $\exp(-pen(\{\boldsymbol{\Omega}\}|\boldsymbol{\delta}, \boldsymbol{\xi}))$ for DSS-FGL:

$$(\boldsymbol{\omega}_{jk}^{(g)})^T \boldsymbol{\Theta}_{jk} \boldsymbol{\omega}_{jk}^{(g)} = \sum_g \frac{v_{\delta_{jk}}}{\tau_{jkg}} (\omega_{jk}^{(g)})^2 + \sum_{g<g'} \frac{v_{\xi_{jk}^*}}{\phi_{jkgg'}} (\omega_{jk}^{(g)} - \omega_{jk}^{(g')})^2.$$

Using the same identity as before, we can rewrite the conditional distribution below into the form of the DSS-FGL penalty:

$$p(\{\boldsymbol{\Omega}\}|\boldsymbol{\delta}, \boldsymbol{\xi}) \propto \exp(-\sum_{j<k} \frac{\lambda_1}{v_{\delta_{jk}}} |\omega_{jk}^{(g)}| - \sum_{j<k} \sum_{g<g'} \frac{\lambda_2}{v_{\xi_{jk}^*}} |\omega_{jk}^{(g)} - \omega_{jk}^{(g')}| - \sum_g \sum_j \frac{\lambda_0}{2} \omega_{jj}^{(g)}).$$

The proof of the DSS-FGL conditional distributions being proper is similar to the previous case. We first note that

$$\boldsymbol{\Theta}_{jk} = \text{diag}(\{\frac{1}{\tau_{jkg}}\}_{g=1,\ldots,G}) + \boldsymbol{L}_{jk}$$

where $\boldsymbol{L}_{jk}$ is a graph Laplacian matrix and is positive semi-definite. Then by Minkowski inequality, $\det(\boldsymbol{\Theta}_{jk}) \geq \det(\text{diag}(\{\frac{1}{\tau_{jkg}}\}_{g=1,\ldots,G}))$. Then we have

$$
\begin{aligned}
C^{-1}_{\boldsymbol{\tau},\boldsymbol{\phi}} &\propto \int \prod_{j<k} \big(\det(\boldsymbol{\Theta}_{jk})^{-\frac{1}{2}} \exp(-\frac{\lambda_1^2}{2}\sum_g \tau_{jkg} - \frac{\lambda_2^2}{2}\sum_{g<g'}\phi_{jkgg'}) \prod_g \tau_{jkg}^{-\frac{1}{2}} \prod_{g<g'} \phi_{jkgg'}^{-\frac{1}{2}}\big) d\{\boldsymbol{\tau},\boldsymbol{\phi}\} \\
&\leq \int \prod_{j<k} \big(\exp(-\frac{\lambda_1^2}{2}\sum_g \tau_{jkg} - \frac{\lambda_2^2}{2}\sum_{g<g'}\phi_{jkgg'}) \prod_{g<g'} \phi_{jkgg'}^{-\frac{1}{2}}\big) d\{\boldsymbol{\tau},\boldsymbol{\phi}\} \\
&\leq \int \prod_{j<k} \big(\exp(-\frac{\lambda_2^2}{2}\sum_{g<g'}\phi_{jkgg'}) \prod_{g<g'} \phi_{jkgg'}^{-\frac{1}{2}}\big) d\boldsymbol{\phi},
\end{aligned}
$$

which is again finite since the integral consists of products of Gamma densities. The rest of the argument follows in the same way as the DSS-GGL case.

B.1 Gibbs sampler of the proposed models

The EM algorithm introduced in Section 5.5 maximizes the complete data likelihood by looking at the Laplace representation after integrating out all the latent parameters. In this section, we show that these latent parameters, $\boldsymbol{\tau}, \boldsymbol{\phi}$ and $\boldsymbol{\rho}$, facilitates efficient block Gibbs sampling algorithms for fully Bayesian inference.

We start by describing the posterior sampling of $\{\boldsymbol{\Omega}\}$. The basic idea is to sample each column and row for all the precision matrices jointly. To similify notation, we separate out the last column and row in $\boldsymbol{\Omega}_g$ and $\boldsymbol{S}_g$ and define

$$
\boldsymbol{\Omega}_g = \begin{pmatrix} \boldsymbol{\Omega}_{11}^{(g)} & \boldsymbol{\omega}_{12}^{(g)} \\ \boldsymbol{\omega}_{21}^{(g)} & \omega_{22}^{(g)} \end{pmatrix}, \quad \boldsymbol{S}_g = \begin{pmatrix} \boldsymbol{S}_{11}^{(g)} & \boldsymbol{s}_{12}^{(g)} \\ \boldsymbol{s}_{21}^{(g)} & s_{22}^{(g)} \end{pmatrix}.
$$

We further let $\boldsymbol{\omega}_{12} = [(\boldsymbol{\omega}_{12}^{(1)})^T, (\boldsymbol{\omega}_{12}^{(2)})^T, ..., (\boldsymbol{\omega}_{12}^{(G)})^T]^T$, and $\boldsymbol{s}_{12} = [(\boldsymbol{s}_{12}^{(1)})^T, (\boldsymbol{s}_{12}^{(2)})^T, ..., (\boldsymbol{s}_{12}^{(G)})^T]^T$, each denoting a vector of length $(p-1)G$, and $\boldsymbol{\omega}_{22} = [\omega_{22}^{(1)}, \omega_{22}^{(2)}, ..., \omega_{22}^{(G)}]$.

The conditional distribution of $(\boldsymbol{\omega}_{12}, \boldsymbol{\omega}_{22})$ given the rest of the elements in $\{\boldsymbol{\Omega}\}$ does not seem to take any standard form. However, if we perform a change of variables and let

$\theta_g = \omega_{22}^{(g)} - \boldsymbol{\omega}_{21}^{(g)}(\boldsymbol{\Omega}_{11}^{(g)})^{-1}\boldsymbol{\omega}_{12}^{(g)}$, the conditional distribution of $(\boldsymbol{\omega}_{12}, \boldsymbol{\theta})$ becomes

$$\begin{aligned} p(\boldsymbol{\omega}_{12}, \boldsymbol{\theta}) &\propto \sum_g \theta_g^{\frac{n_g}{2}} \exp(-\frac{s_{22}^{(g)} + \lambda_0}{2}\theta_g - \boldsymbol{s}_{12}^T\boldsymbol{\omega}_{12} - \frac{1}{2}\boldsymbol{\omega}_{12}^T\boldsymbol{A}\boldsymbol{\omega}_{12}) \\ &= \prod_g \text{Gamma}(\theta_g; \frac{n_g}{2} + 1, \frac{s_{22}^{(g)} + \lambda_0}{2}) \times \text{Normal}(\boldsymbol{\omega}_{12}; -\boldsymbol{A}^{-1}\boldsymbol{s}_{12}, \boldsymbol{A}^{-1}). \end{aligned}$$

The $\boldsymbol{A}$ matrix can be calculated by $\boldsymbol{A} = \boldsymbol{U} + \boldsymbol{V}$, where $\boldsymbol{U}$ is a matrix by rearrangeing the precision matrices so that its $((g-1)(p-1)+k, (g'-1)(p-1)+k)$-th element is the (g, g')-element in $\boldsymbol{\Theta}_{jk}$ defined in (5.13) and (5.15), and

$$\boldsymbol{V} = \begin{pmatrix} (\lambda_0 + s_{22}^{(1)})(\boldsymbol{\Omega}_{11}^{(1)})^{-1} & & \\ & \ddots & \\ & & (\lambda_0 + s_{22}^{(G)})(\boldsymbol{\Omega}_{11}^{(G)})^{-1} \end{pmatrix},$$

For DSS-GGL, we notice that $\boldsymbol{A}$ is block diagonal, thus we can alternatively sample $\boldsymbol{\omega}_{12}^{(g)}$ independently by

$$\boldsymbol{\omega}_{12}^{(g)}|\cdot \sim \text{Normal}(-\boldsymbol{A}_g^{-1}\boldsymbol{s}_{12}^{(g)}, \boldsymbol{A}_g^{-1})$$

where $\boldsymbol{A}_g = (\lambda_0 + s_{22}^{(g)})(\boldsymbol{\Omega}_{11}^{(g)})^{-1} + \boldsymbol{\Theta}_{11}^{(g)}$.

Given $\{\boldsymbol{\Omega}\}$, the latent parameters in DSS-GGL have simple conditional distribution as

follows:

$$\begin{aligned}
\tau_{jkg}^{-1}|\cdot &\sim \text{InvGaussian}(\frac{\lambda_1}{v_{\delta_{jk}}^{\frac{1}{2}}|\omega_{jk}^{(g)}|}, \lambda_1^2), \quad j,k = 1,...,p, g = 1,..,G \\
\rho_{jk}^{-1}|\cdot &\sim \text{InvGaussian}(\frac{\lambda_2}{v_{\xi_{jk}^*}\sum_g(\omega_{jk}^{(g)})^2}, \lambda_2^2), \quad j,k = 1,...,p \\
\delta_{jk},\xi_{jk}|\cdot &\sim p^*(\boldsymbol{\delta}_{jk},\boldsymbol{\xi}_{jk}), \quad j,k = 1,...,p \\
\pi_{\boldsymbol{\delta}} &\sim \text{Beta}(\sum \delta_{jk} + a_1, \sum(1-\delta_{jk}) + b_1) \\
\pi_{\boldsymbol{\xi}} &\sim \text{Beta}(\sum \xi_{jk} + a_2, \sum(1-\xi_{jk}) + b_2)
\end{aligned}$$

where $p^*(\boldsymbol{\delta},\boldsymbol{\xi})$ is defined in Section 5.5.

For DSS-FGL, the conditional distribution of $\boldsymbol{\tau},\boldsymbol{\delta},\boldsymbol{\xi},\pi_{\boldsymbol{\delta}}$, and $\pi_{\boldsymbol{\xi}}$ are the same as DSS-GGL. The conditional distribution of $\boldsymbol{\phi}$ is

$$\phi_{jkgg}^{-1} \sim \text{InvGaussian}(\frac{\lambda_2}{v_{\xi_{jk}^*}^{\frac{1}{2}}|\omega_{jk}^{(g)} - \omega_{jk}^{(g')}|}, \lambda_2^2), \quad j,k = 1,...,p, g, g' = 1,..,G$$

The Gibbs sampler is then complete by circling through and sampling each blocks of $\{\boldsymbol{\Omega}\}$ and the latent parameters with the above posterior conditional distributions.

B.2 Additional simulation evidence

Additional results for $p = 100$ and $p = 200$ are shown in Figure B.2 and B.3. The dots correspond to DSS-GGL and DSS-FGL with $\lambda_1 = 1$, $\lambda_2 = 0.1$. We also examined different choices of λ_2 are found no substantial differences in performance. An exploratory sensitivity analysis is presented in the next subsection.

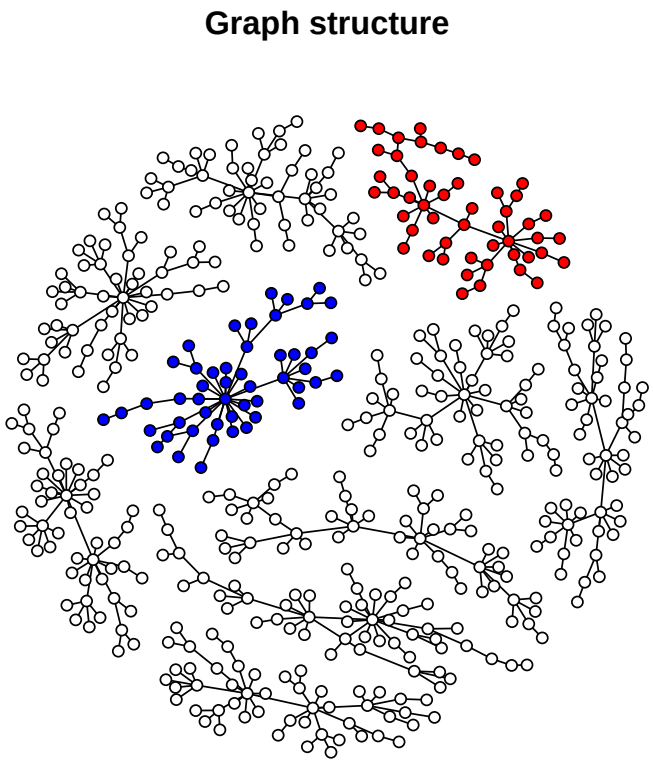


Figure B.1: Graph structure of the simulated dataset. The edges between the red nodes are removed from the second class, and edges between both the red and blue nodes are removed from the third class.

B.2.1 Sensitivity to hyperparameters

Figure B.4 and B.5 shows the converged regions over 100 replications on space of the true positive against false positive discoveries for edges and differential edges respectively when $p = 200$. It can be seen that the performance is relatively stable under different choices of λ_2.

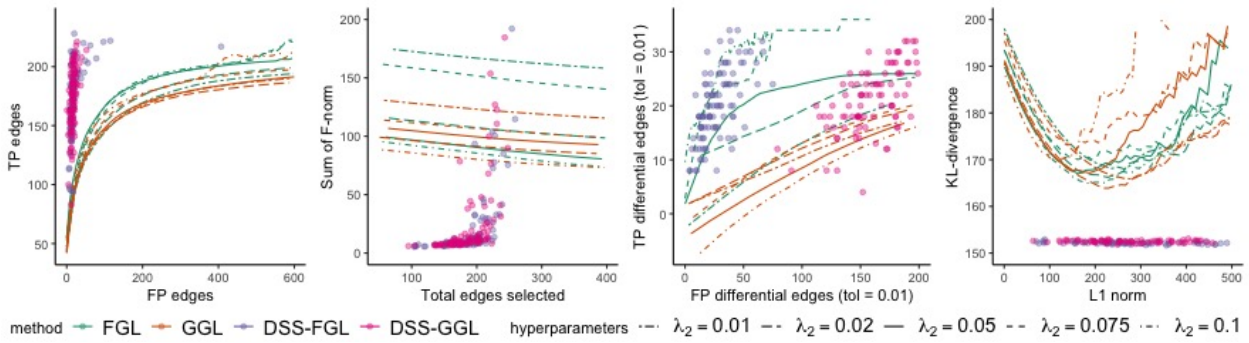


Figure B.2: Performance of FGL, GGL, DSS-FGL, and DSS-GGL over 100 replications, $p = 100$. The dots represent the metrics for the 100 selected models under DSS-FGL and DSS-GGL, and the lines represent the average performance of FGL and GGL over 100 replications under different tuning parameters.

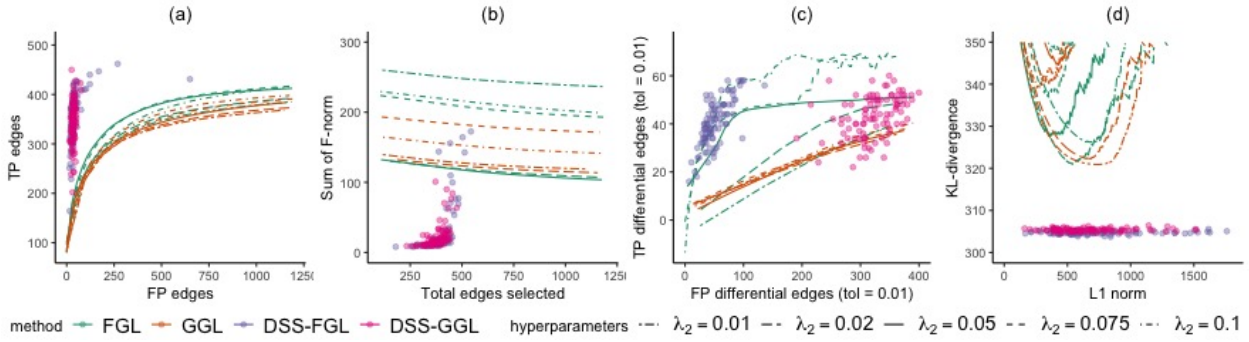


Figure B.3: Performance of FGL, GGL, DSS-FGL, and DSS-GGL over 100 replications, $p = 200$. The dots represent the metrics for the 100 selected models under DSS-FGL and DSS-GGL, and the lines represent the average performance of FGL and GGL over 100 replications under different tuning parameters.

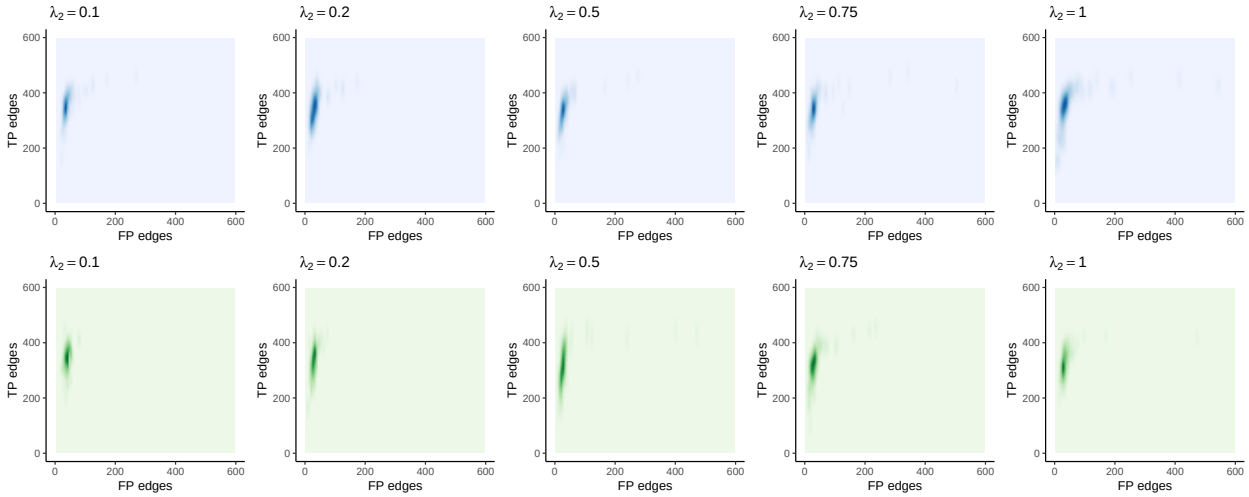


Figure B.4: The density plot of true positive edges against false positive edges for DSS-FGL (top row), and DSS-GGL (bottom row) under different choices of λ_2. λ_1 is set to 1.

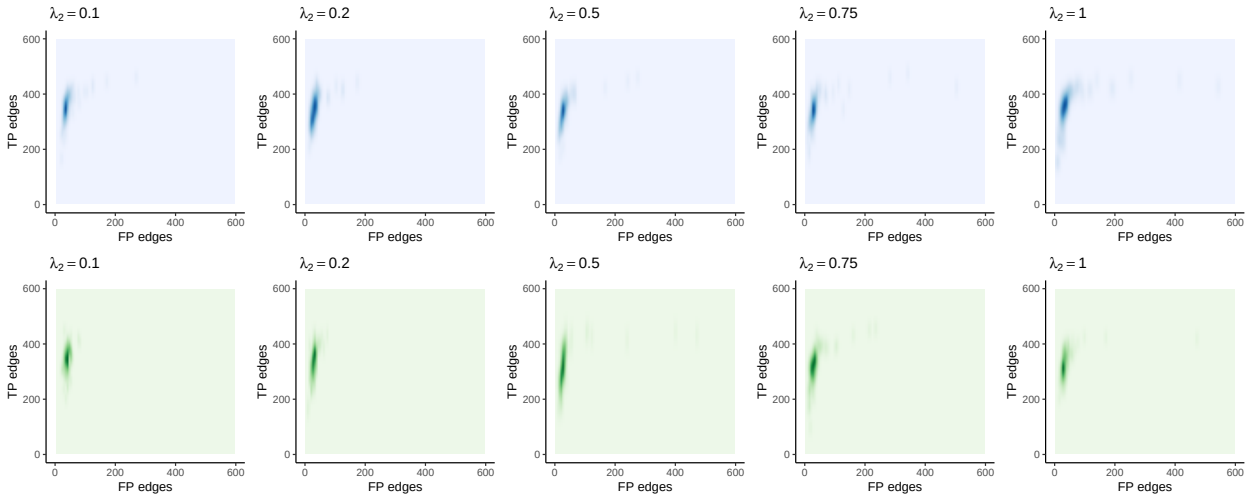


Figure B.5: The density plot of true positive differential edges against false positive positive edges for DSS-FGL (top row), and DSS-GGL (bottom row) under different choices of λ_2. λ_1 is set to 1.

B.3 Details on the verbal autopsy data analysis

B.3.1 List of symptoms

Table B.1 shows the questions with continuous responses used in the analysis in Section 5.7.

Abbreviation	Questionnaire item
ill	For how long was [name] ill before s/he died? [days]
fever	How many days did the fever last? [days]
rash	How many days did [name] have the rash? [days]
ulcer	For how many days did the ulcer ooze pus? [days]
yellow discoloration	For how long did [name] have the yellow discoloration? [days]
ankle swelling	For how long did [name] have ankle swelling? [days]
puffiness face	For how long did [name] have puffiness of the face? [days]
puffiness body	For how long did [name] have puffiness all over his/her body? [days]
cough	For how long did [name] have a cough? [days]
difficulty breathing	For how long did [name] have difficulty breathing? [days]
fast breathing	For how long did [name] have fast breathing? [days]
liquid stool	For how long before death did [name] have loose or liquid stools? [days]
vomit	For how long before death did [name] vomit? [days]
difficulty swallowing	For how long before death did [name] have difficulty swallowing? [days]
belly pain	For how long before death did [name] have belly pain? [days]
protruding belly	For how long before death did [name] have a protruding belly? [days]
mass belly	For how long before death did [name] have a mass in the belly [days]
headaches	For how long before death did [name] have headaches? [days]
stiff neck	For how long before death did [name] have stiff neck? [days]
unconsciousness	For how long did the period of loss of consciousness last? [days]
confusion	For how long did the period of confusion last? [days]
convulsion	For how long before death did the convulsions last? [days]
paralysis	For how long before death did [name] have paralysis? [days]
period overdue	For how many weeks was her period overdue? [days]
tobacco	How much pipe/chewing tobacco did [name] use daily?
cigarettes	How many cigarettes did [name] smoke daily?
age	Age [years]

Table B.1: List of symptoms considered in this analysis.

B.3.2 Comparing with JGL

The estimated symptom network from the FGL and GGL are summarized in Figure B.6. We fit both models under a 2-dimensional grids over λ_1 and λ_2. As expected, AIC selects very dense graphs (first two rows of Figure B.6) and are difficult to interpret. We also compare the FGL and GGL graph with the closest number of edges as those from DSS-FGL and DSS-GGL in the third and forth row of Figure B.6. The number of differential edges is typically smaller compared to DSS-FGL and DSS-GGL, which is likely due to over penalization of similarities, i.e., edges become too similar using FGL, and too sparse among half of the nodes using GGL.

B.4 Details on prediction of missing mortality rates

The data we consider in this example consist of log mortality rates over $n = 51$ years for $p = 101$ age groups, and 2 classes representing female and male series respectively. The estimated graph structure from one of the cross-validation dataset using FGL and DSS-FGL are shown in Figure B.7. The Lee-Carter model are estimated using the R package `ilc` (Butt et al., 2014) for each gender separately.

The DSS-FGL is able to pick up more conditional dependence structures along the diagonal among several age groups, while the FGL estimates mostly within adults only. It is interesting that both approaches identifies positive partial correlations between age $14 - 17$ and $30 - 40$ between male and female. This is likely due to the fact that male mortality around age 20 typically shows a hump of increase due to young adult accident mortality, which leads to the mean model more likely to underestimations for mortality during age $18 - 30$ and overestimations both before and after that period. This relationship of the age curve, however, is not seen in female mortality.

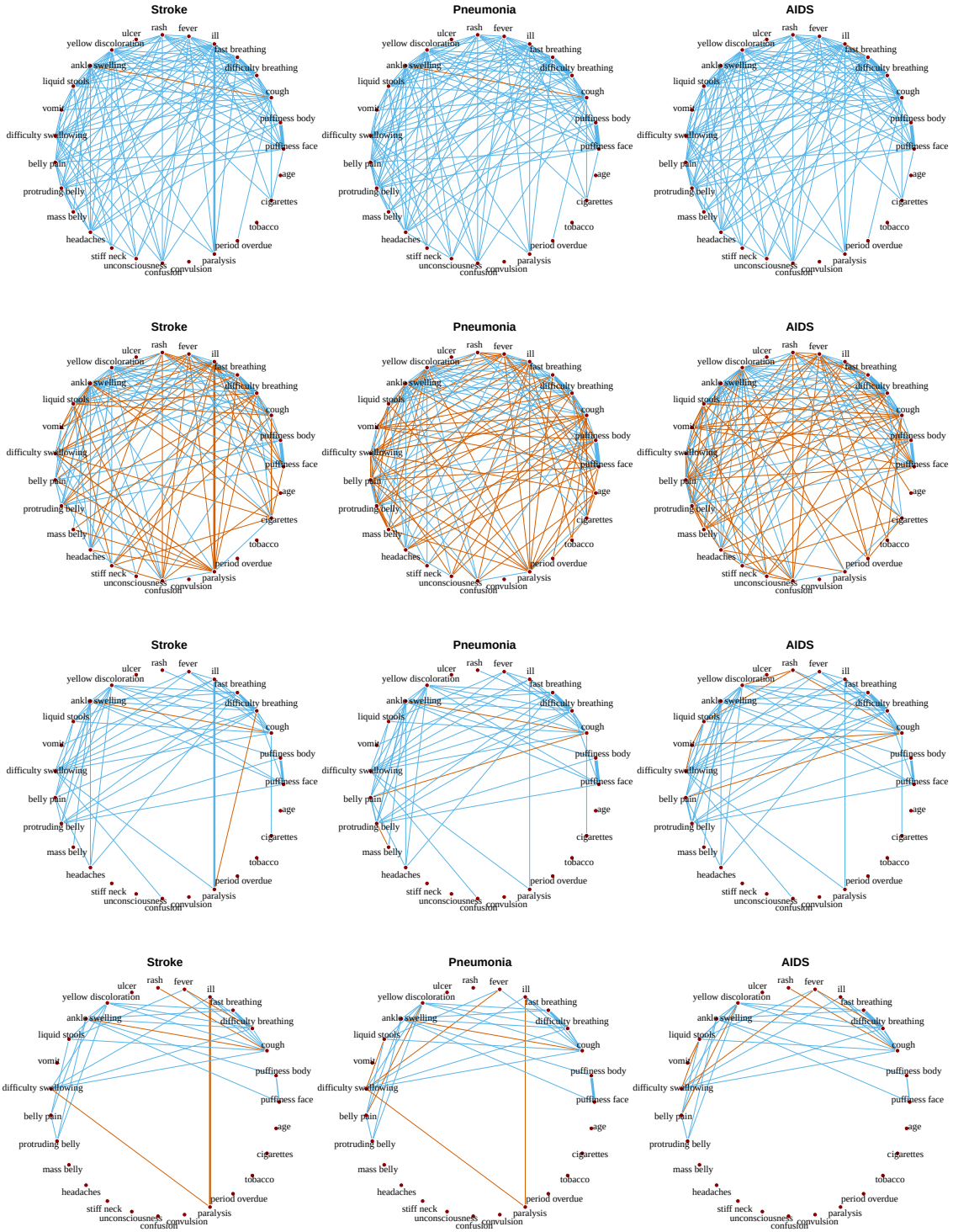


Figure B.6: Estimated edges between the symptoms under the three causes using FGL using AIC (first row), GGL using AIC (second row), FGL with the same number of edges as selected by DSS-FGL (third row), and GGL with the same number of edges as selected by DSS-GGL (last row). The width of the edges are proportional to the size of $|\omega_{jk}^{(g)}|$. Common edges across all groups are colored in blue, and the differential edges are colored in red.

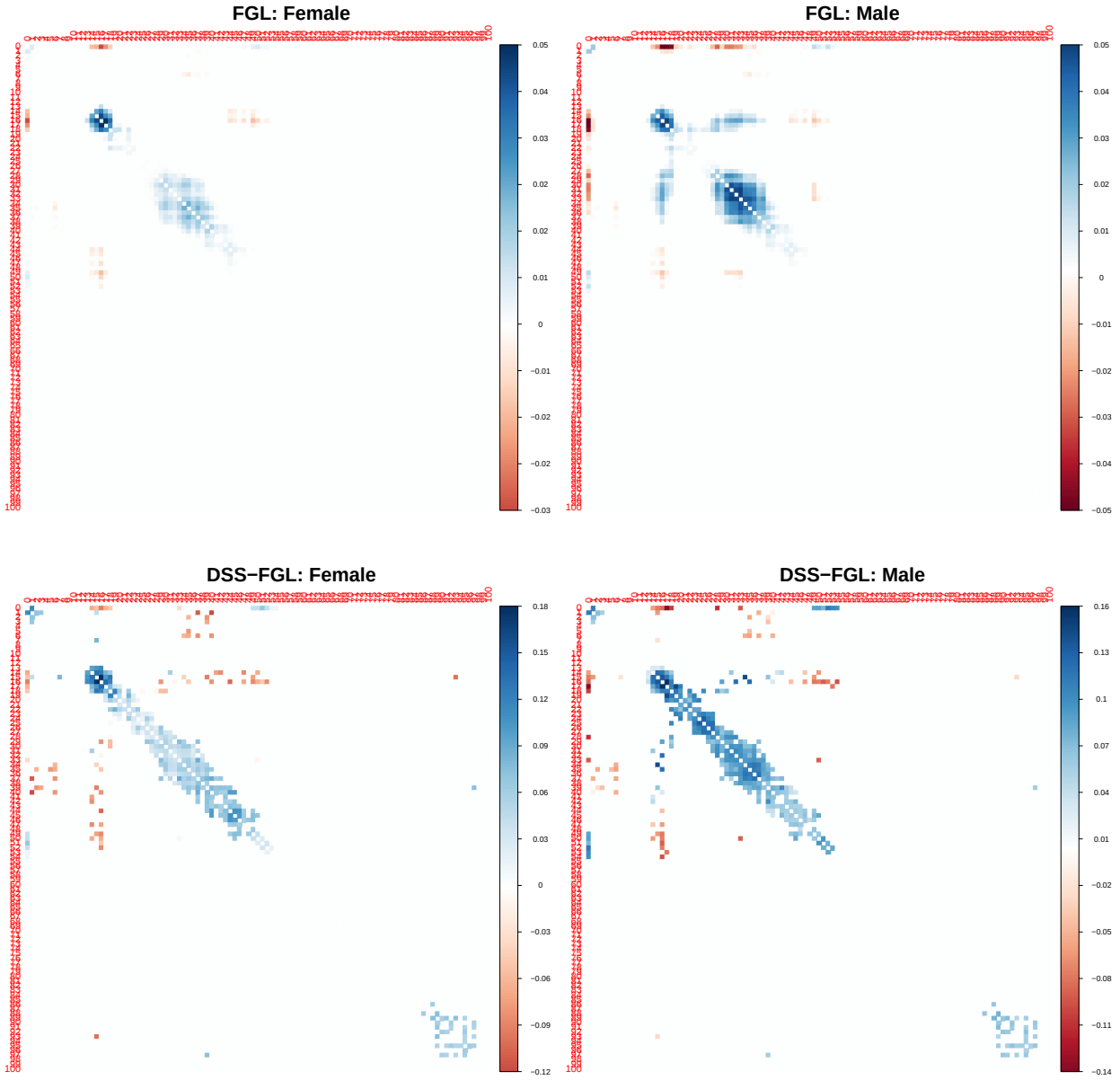


Figure B.7: Estimated partial correlation matrix using one cross-validation dataset. The partial correlations among the 101 age groups are estimated using FGL with the same number of edges as selected by DSS-FGL (top row), and DSS-FGL (bottom row). DSS-FGL estimates 197 and 199 edges respectively for female and male. The closet configuration of FGL estimates 157 and 241 edges respectively. The precision matrices are rescaled and negated to partial correlations for easier interpretation.