Topics in High-dimensional Unsupervised Learning

by

Jian Guo

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2011

Doctoral Committee:

Associate Professor Elizaveta Levina, Co-Chair
Associate Professor Ji Zhu, Co-Chair
Professor George Michailidis
Professor Susan A. Murphy
Assistant Professor Qiaozhu Mei

THE UNIVERSITY OF MICHIGAN

ARTES · SCIENTIA · VERITAS

1817

© Jian Guo 2011

All Rights Reserved

ACKNOWLEDGEMENTS

First of all, I would particularly like to thank my advisers, Professor Elizaveta Levina, Professor George Michailidis and Professor Ji Zhu, for their guidance and help throughout my research. Without their valuable suggestions and support, this dissertation could not be completed. I also thank the dissertation committee member Professor Susan Murphy for her helpful comments. Professor Qiaozhu Mei deserves special thanks for leading me to an exciting field about internet research. Finally, I would express my gratitude to my wife and my parents for their constant support and encouragement.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS ii

LIST OF FIGURES vi

LIST OF APPENDICES x

ABSTRACT xi

CHAPTER

I. Introduction and Literature Review 1

1.1 Structure Estimation of High-dimensional Graphical Models . 1
1.2 Grouped Variable Selection for High-dimensional Data Analysis 4
1.2.1 Pairwise Variable Selection for High-dimensional Model-based Clustering 5
1.2.2 Sparse Principal Component Analysis 7
1.3 Organization of the Chapters 8

II. Joint Estimation of Multiple Graphical Models 9

2.1 Introduction 9
2.2 Estimation of Single Graphical Models 10
2.3 Methodology 11
2.3.1 The Joint Estimation Method 11
2.3.2 The Algorithm 13
2.3.3 Model Selection 14
2.4 Asymptotic Properties 15
2.5 Numerical Evaluation 16
2.5.1 Simulation Settings 16
2.5.2 Simulation Results 19
2.6 University Webpages Example 20

III. Asymptotic Properties of the Joint Neighborhood Selection Method for Estimating Categorical Markov Networks 26

3.1 Introduction 26
3.2 Methodology 27
3.3 Theoretical Properties 30
3.4 Application to the Senate Voting Record 32
3.5 Extension to General Markov Networks 35

IV. Estimating Heterogeneous Graphical Models for Discrete Data with an Application to Roll Call Voting 37

4.1 Introduction 37
4.2 Model and Estimation Algorithm 38
4.2.1 Problem Setup and Separate Estimation Method . . 39
4.2.2 Joint Estimation of Heterogeneous Networks 40
4.2.3 Algorithm and Model Selection 42
4.3 Simulation Study 43
4.4 Analysis of the U.S. Senate voting records 47
4.5 Asymptotic Properties 49
4.6 Concluding Remarks 52

V. Graphical Models with Ordinal Variables 54

5.1 Introduction 54
5.2 Methodology 55
5.2.1 Probit Graphical Model 55
5.2.2 Algorithm for Probit Graphical Model 57
5.2.3 Approximation of the Conditional Expectation . . . 59
5.2.4 Model Selection 63
5.3 Simulated Examples 63
5.3.1 Computational Cost and Performance 64
5.3.2 High-dimensional Experiments 65
5.4 Application to Movie Rating Records 70

VI. Pairwise Variable Selection for High-dimensional Model-based Clustering 74

6.1 Introduction 74
6.2 Problem Formulation and Pairwise Fusion 75
6.2.1 The Pairwise Fusion Penalty 77
6.2.2 The Adaptive Pairwise Fusion Penalty 78
6.2.3 Model Selection 80
6.3 The Optimization Algorithm 80
6.4 Numerical Results 83

6.5 Applications to Gene Expression Data 88
6.5.1 The SRBCT Data 89
6.5.2 PALL Data Set 90
6.6 Conclusions 91

VII. Sparse Fused Principal Component Analysis 99

7.1 Introduction 99
7.2 The Model and its Estimation 100
7.2.1 Preliminaries and Sparse Variants of PCA 100
7.2.2 Sparse Fused Loadings 102
7.2.3 Optimization of the Objective Function 104
7.2.4 Estimation of $\boldsymbol{B}$ Given $\boldsymbol{A}$ 105
7.2.5 Selection of Tuning Parameters 106
7.2.6 Computational Complexity and Convergence 107
7.3 Numerical illustration of SFPCA 108
7.4 Application of SFPCA to Real Datasets 111
7.5 Concluding Remarks 114

APPENDICES 120

BIBLIOGRAPHY 147

LIST OF FIGURES

Figure

2.1 The common links present in all categories in the three simulated networks. . . . 18

2.2 Receiver operating characteristic curves. The horizontal and vertical axes in each panel are false positive rate and sensitivity, respectively. The solid line corresponds to the joint estimation method, and the dashed line corresponds to the separate estimation method. ρ is the ratio of the number of individual links to the number of common links. 23

2.3 Common structure in the webpages data. Panel (A) shows the estimated common structure for the four categories. The nodes represent 100 terms with the highest log-entropy weights. The area of the circle representing a node is proportional to its log-entropy weight. The width of an edge is proportional to the magnitude of the associated partial correlation. Panels (B)–(D) show subgraphs extracted from the graph in panel (A). . . . 24

2.4 "Student" and "Faculty" graphs. The light lines are the links appearing in both categories, and the dark lines are the links only appearing in one category. . . . 25

2.5 "Course" and "Project" graphs. The light lines are the links appearing in both categories, and the dark lines are the links only appearing in one category. . . . 25

3.1 Voting dependencies between senators estimated by the joint neighborhood selection method. Each red (blue) circle represents a Republican (Democratic) senator, the circle size is proportional to the degree of the node. Senator Jeffords (the purple circle) is an independent senator. A solid green (dashed red) link represents a positive (negative) dependence between two senators. The width of each link is proportional to its associated $|\widehat{\theta}_{j,j'}|$. For clarity, all links with $|\widehat{\theta}_{j,j'}| \leq 0.1$ have the same width. 34

4.1 The networks used in three simulated examples. The black lines represent the common structure, whereas the red, blue and green lines represent the individual links in the three categories. ρ is the ratio of the number of individual links to the number of common links. 45

4.2 Results for the balanced scenario ($n_1 = n_2 = n_3 = 200$). The ROC curves are averaged over 50 replications. ρ is the ratio between the number of individual links and the number of common links. 46

4.3 Results for the unbalanced scenario ($n_1 = 150$, $n_2 = 300$, $n_3 = 450$). The ROC curves are averaged over 50 replications. ρ is the ratio between the number of individual links and the number of common links. 47

4.4 The common and individual structures for the Senate voting data. The nodes represent the 100 senators, with red, blue and purple node colors corresponding to Republican, Democrat, or Independent (Senator Jeffords), respectively. A solid line corresponds to a positive interaction effect and a dashed line to a negative interaction effect. The width of a link is proportional to the magnitude of the corresponding overall interaction effect. For each individual network, the links that only appear in this category are highlighted in purple. . 53

5.1 The comparison of Frobenius loss and Entropy loss over different tuning parameters. The direct estimation, the Gibbs sampling estimation and the approximation estimation are represented by blue dotted, red dashed and black solid lines. 66

5.2 Illustration of the networks used in four simulated examples: scale-free graph, hub graph, nearest-neighbor graph and block graph. . . 67

5.3 The ROC curves estimated by probit graphical model (solid dark line), the oracle model (dotted blue line) and the Gaussian model (dashed red line). The oracle model and the Gaussian model applies the graphical lasso algorithm to the latent data $\boldsymbol{Z}$ and the observed data $\boldsymbol{X}$, respectively. 72

5.4 The network estimated by the probit graphical model. The nodes represent the movies labeled by their titles. The area of a node is proportional to its degree and the width of a link is proportional to the magnitude of the corresponding partial correlations. 73

6.1 A toy example. Variable 1 is informative for separating clusters 2 and 3, and variable 2 is informative for separating clusters 1 and 2. 75

6.2 The distribution of informative variables in Simulation 1 (left) and Simulation 2 (right). The red star indicates the position of the overall sample mean. 85

6.3 Simulation 3. The sample sizes of clusters 1, 2, 3 and 4 are 20, 20, 200, and 200, respectively. The red star indicates the position of the overall sample mean, and the plot is shifted to show centered data. 86

6.4 Plots of the negative logarithm p-values vs variance for SRBCT data. The left column is the top 100 genes (largest overall variances), and the right column is the bottom 100 genes. The upper row is negative logarithm p-values corresponding to an F-statistics comparing four tumor subtypes, and the lower row is the negative logarithm p-values for the six identified clusters. Triangles denote the genes that are not selected by the APFP method. 95

6.5 Pairwise variable selection results for the APFP method on the SRBCT data with top 100 genes. Each row corresponds to a gene. Each column corresponds to a cluster pair; for example, “1/2” indicates clusters 1 and 2. A black (white) spot indicates that the estimated means of the corresponding gene for the two clusters are different (the same). For example, gene “435953” is non-informative for separating clusters 1 and 3, 2 and 5, and 4 and 6. 96

6.6 Plots of the negative logarithm p-values vs variance for PALL data. The left column is the top 100 genes (largest overall variances), and the right column is the bottom 100 genes. The upper row is negative logarithm p-values corresponding to an F-statistics comparing four tumor subtypes, and the lower row is the negative logarithm p-values for the six identified clusters. Triangles denote the genes that are not selected by the APFP method. 97

6.7 Pairwise variable selection results for the APFP method on the PALL data with top 100 genes. Each row corresponds to a gene. Each column corresponds to a cluster pair; for example, “1/2” indicates clusters 1 and 2. A black (white) spot indicates that the estimated means of the corresponding gene for the two clusters are different (the same). 98

7.1 Factor loadings of the first (left column) and second (right column) PC vectors estimated by ordinary PCA from the true covariance (first row), ordinary PCA from the sample covariance (second row), sparse PCA (third row) and SFPCA (fourth row). The horizontal axis is the variables and the vertical axis is the value of the loadings. Each colored curve represents the PC vector in one replication. The median loadings over 50 repetitions are represented by the black bold lines. 116

7.2 The histogram of the pairwise correlations between the height/length variables: weight, height in shoes, height in bare feet, seated height, lower arm length, thigh length, and lower leg length. 117

7.3 Comparison of the first (left panel) and second (right panel) PC vectors from ordinary PCA (dashed line), sparse PCA (dotted line) and SFPCA (solid line). 117

7.4 The first two eigen-images of digit “3” estimated by PCA and SFPCA, respectively. 119

LIST OF APPENDICES

Appendix

A. Joint Estimation of Multiple Graphical Models 121

B. Asymptotic Properties of the Joint Neighborhood Selection Method for Estimating Categorical Markov Networks 125

C. Estimating Heterogeneous Graphical Models for Discrete Data with an Application to Roll Call Voting 139

ABSTRACT

Topics in High-Dimensional Unsupervised Learning

by

Jian Guo

Co-Chairs: Elizaveta Levina and Ji Zhu

The first part of the dissertation introduces several new methods for estimating the structure of graphical models. Firstly, we consider estimating graphical models with discrete variables, including nominal variables and ordinal variables. For the nominal variables, we prove the asymptotic properties of the joint neighborhood selection method proposed by Hoefling and Tibshirani (2009) and Wang et al. (2009), which is used to fit high-dimensional graphical models with binary random variables. We show that this method is consistent in terms of both parameter estimation and structure estimation and extend it to general nominal variables. For ordinal variables, we introduce a new graphical model, which assumes that the ordinal variables are generated by discretizing marginal distributions of a latent multivariate Gaussian distribution and the relationships of these ordinal variables are described by the underlying Gaussian graphical model. We develop an EM-like algorithm to estimate the underlying latent network and apply the mean field theory to improve computational efficiency.

We also consider the problem of jointly estimating multiple graphical models which share the variables but come from different categories. Compared with separate estimation for each category, the proposed joint estimation method significantly improves performance when graphical models in different categories have some similarities. We develop joint estimation methods both for Gaussian graphical models and for graphical models for categorical variables.

In the second part of the dissertation, we develop two methods to improve interpretability of high-dimensional unsupervised learning methods. First, we introduce a pairwise variable selection method for high-dimensional model-based clustering. Unlike existing variable selection methods for clustering problems, the proposed method not only selects the informative variables, but also identifies which pairs of clusters are separable by each informative variable. We also propose a new method to identify both sparse structures and "block" structures in factor loadings in principal component analysis. This is achieved by forcing highly correlated variables to have identical factor loadings via a regularization penalty.

CHAPTER I

Introduction and Literature Review

1.1 Structure Estimation of High-dimensional Graphical Models

Undirected graphical models have proved useful in a number of application areas, including bioinformatics (Airoldi, 2007), natural language processing (Jung et al., 1996), image analysis (Li, 2001), and many others, due to their ability to succinctly represent dependence relationships among a set of random variables. Such models represent the relationships between p variables $X_1, \cdots, X_p$ through an undirected graph $G = (V, E)$, whose node set V corresponds to the variables and the edge set E characterizes their pairwise relationships. Specifically, variables X_j and $X_{j'}$ are conditionally independent given all other variables if their associated nodes are not linked by an edge.

Two important types of graphical models are the Gaussian graphical model, where the p variables are assumed to follow a joint Gaussian distribution, and the Markov network, which captures relationships between categorical variables. In the former, the structure of the underlying graph can be recovered by estimating the corresponding inverse covariance (precision) matrix, whose off-diagonal elements are proportional to the partial correlations between the variables. A large body of literature has emerged over the past few years addressing this issue, especially for sparse networks. A number of methods focus on estimating a sparse inverse covariance matrix and inferring the network from estimated zeros (Banerjee et al., 2008; Yuan and Lin, 2007; Rothman et al., 2008; Friedman et al., 2008; Lam and Fan, 2009; Rocha et al., 2008; Ravikumar et al., 2008; Peng et al., 2009). Another class of methods focuses on estimating the network directly without first estimating the precision matrix (Drton and Perlman, 2004; Meinshausen and Buhlmann, 2006). There is also some recent literature on directed acyclic graphical models (see, for example, Shojaie and Michailidis (2010) and references therein).

For the Markov network, the estimation problem is significantly harder, since it is computationally infeasible for any realistic size network to directly evaluate the likelihood, due to the intractable constant (the log-partition function). Several methods in the literature overcome this difficulty by employing computationally tractable approximations. For example, d'Aspremont et al. (2008) proposed estimating the network structure using an ℓ_1-penalized surrogate likelihood, where the log-partition function is approximated by a log-determinant relaxation. Kolar and Xing (2008) improved on this method by incorporating a cutting-plane algorithm to obtain a tighter outer bound on the marginal polytope. Alternatively, Ravikumar et al. (2009) proposed a neighborhood selection method that approximates the likelihood by a pseudo-likelihood function, in analogy to the Meinshausen and Buhlmann (2006) method for Gaussian graphical models, where p individual ℓ_1-penalized regressions were fitted, regressing each variable on all others, and the network structure was recovered from the regression coefficients. Ravikumar et al. (2009) separately fit p individual penalized logistic regressions, whose coefficients are used to recover the Markov network structure. They also showed that the neighborhood selection method satisfies both estimation consistency and model selection consistency. However, estimating pairwise interactions by fitting p separate logistic regression leads to lack of symmetry; the estimate of interaction between X_i and X_j may have a different value and even a different sign from the interaction between X_j and X_i.

In this dissertation, we consider two general problems about graphical models. First, we propose a new method to jointly estimate multiple graphical models. This work was motivated by the fact that, in many applications, the data consist of several categories that share the same variables but differ in their dependence structure. The underlying networks have some edges in common but also have others unique to each category. Consider, for example, the gene regulatory networks describing different subtypes of the same cancer: there are some shared pathways across different subtypes, and there are also links that are unique to a particular subtype. To our best knowledge, existing graphical models are only concerned with estimating a single network. In this work, we constructed an estimator that jointly estimates such Gaussian graphical models through a hierarchical penalty function. Compared with separate estimation, the proposed joint estimator is more effective in discovering the common structure and in reducing the estimation variance by borrowing strength across different categories. In addition, we also extended this idea to joint estimation of multiple graphical models with categorical variables. In these two papers, we established the consistency of both parameter and structure estimation in high-dimension settings (allowing p to grow faster than n).

The second problem consider the graphical models with categorical variables, which is a challenging task compared to fitting Gaussian graphical models both analytically and computationally. My research studies the network structure estimation problems for two types of categorical variables: nominal variables and ordinal variables. The two types have intrinsic differences and thus we need to build different graphical models to characterize the association structure for each specific variable type. Markov network is a graphical model which captures associations among nominal variables and the underlying network can be estimated by solving an ℓ_1-regularized log-linear model likelihood. However, the estimation problem is computationally infeasible for large networks due to the intractable partition function in the log-linear model likelihood. Here, we prove the asymptotic properties of an efficient approximate optimization algorithm for estimating large-scale Markov networks. Another important type of discrete variables is the ordinal variables, which have a number of ordered levels. The ordered nature of the ordinal variable means that neither the Markov network nor the Gaussian graphical model is appropriate for characterizing the associations between ordinal variables. In Chapter V, we proposed a latent graphical model where the observed ordinal variables are assumed to be discretized latent continuous variables jointly following a Gaussian graphical model. It is computationally infeasible to directly estimate the proposed latent graphical model using the Expectation-Maximization (EM) algorithm, even for modest-sized networks. In this thesis, we overcome this limitation by developing an approximate algorithm which can efficiently estimate large-scale graphical models with ordinal variables.

1.2 Grouped Variable Selection for High-dimensional Data Analysis

With the accumulation of large amount of high-dimensional data, it is becoming increasingly important to identify informative variables and improve interpretability of high-dimensional statistical models. Most existing high-dimensional models achieve these goals by imposing sparsity in parameters. In addition to sparsity, this thesis seeks to improve interpretability from several other angles by introducing the group variable selection for high-dimensional clustering and sparse principal component analysis, respectively.

1.2.1 Pairwise Variable Selection for High-dimensional Model-based Clustering

The goal of clustering is to organize data into a small number of homogeneous groups, thus aiding interpretation. Clustering techniques have been employed in a wide range of scientific fields, including biology, physics, chemistry and psychology. These techniques can broadly be classified into two categories: hierarchical methods and partition methods (see Gordon (2008), Kaufman and Rousseeuw (1990), and references therein). The former typically start from a dissimilarity matrix that captures differences between the objects to be clustered and produce a family of cluster solutions, whose main property is that any two clusters in the family are either disjoint or one is a superset of the other. Various popular agglomerative algorithms, such as single, complete and average linkage belong to this class. Partition algorithms produce non-overlapping clusters, whose defining characteristic is that distances between objects belonging to the same cluster are in some sense smaller than distances between objects in different clusters. The popular K-means algorithm (MacQueen, 1967) and its variants are members of this class. A statistically motivated partition method is model-based clustering, which models the data as a sample from a Gaussian mixture distribution, with each component corresponding to a cluster (McLachlan and Basford, 1988). A number of extensions addressing various aspects of this approach have recently appeared in the literature. For example, Banfield and Raftery (1993) generalized model-based clustering to the non-Gaussian case, while Fraley (1993) extended it to incorporate hierarchical clustering techniques.

The issue of variable selection in clustering, also known as subspace clustering, has started receiving increased attention in the literature recently (for a review of some early algorithms see Parsons et al. (2004)). For example, Friedman and Meulman (2004) proposed a hierarchical clustering method which uncovers cluster structure on separate subsets of variables; Tadesse et al. (2005) formulated the clustering problem in Bayesian terms and developed an MCMC sampler that searches for models comprised of different clusters and subsets of variables; Hoff (2006) also employed a Bayesian formulation based on a Polya urn model; and Raftery and Dean (2006) introduced a method to sequentially compare two nested models to determine whether a subset of variables should be included or excluded from the current model. Some recent approaches addressing variable selection are based on a regularization framework. Specifically, Pan and Shen (2006) proposed to maximize the Gaussian mixture likelihood while imposing an ℓ_1 penalty on the cluster means. In addition, the means of all clusters were required to sum up to zero for each variable. This method removes variables for which all cluster means are shrunk to zero and hence regarded as uninformative. Wang and Zhu (2007) treated the cluster mean parameters associated with the same variable as a natural "group" and proposed an adaptive ℓ_∞ penalty and an adaptive hierarchical penalty to make use of the available group information. Finally, Jornsten and Keles (2008) introduced mixture models that lead to sparse cluster representations in complex multifactor experiments.

Existing variable selection methods for multi-category clustering select informative variables in a "one-in-all-out" manner; that is, a variable is selected if at least one pair of categories is separable by this variable and removed if it fails to separate any of them. In many applications, however, it is useful to further explore which categories can be separated by each informative variable. We refer to this task as category-specific variable selection. In Chapter VI, we proposed a penalty function for high-dimensional model-based clustering. For each variable, this penalty shrinks the difference between all pairs of cluster centroids for each variable and identifies clusters as nonseparable if their centroids are fused to an identical value.

1.2.2 Sparse Principal Component Analysis

Principal component analysis (PCA) is a widely used data analytic technique that aims to reduce the dimensionality of the data for simplifying further analysis and visualization. It achieves its goal by constructing a sequence of *orthogonal linear combinations* of the original variables, called the principal components (PC), that have maximum variance. The technique is often used in exploratory mode and hence good interpretability of the resulting principal components is an important goal. However, it is often hard to achieve this in practice, since PCA tends to produce principal components that involve *all* the variables. Further, the orthogonality requirement often determines the signs of the variable loadings (coefficients) beyond the first few components, which makes meaningful interpretation challenging.

Various alternatives to ordinary PCA have been proposed in the literature to aid interpretation, including rotations of the components (Jollife, 1995), restrictions for their loadings to take values in the set $\{-1, 0, 1\}$ (Vines, 2000), and construction of components based on a subset of the original variables (McCabe, 1984). More recently, variants of PCA that attempt to select different variables for different components have been proposed and are based on a regularization framework that penalizes some norm of the PC vectors. Such variants include SCoTLASS (Jollife et al., 2003) that imposes an ℓ_1 penalty on the ordinary PCA loadings and a recent sparse PCA technique (Zou et al., 2006) that extends the elastic net (Zou and Hastie, 2005) procedure by relaxing the PCs orthogonality requirement.

While existing research addressed this problem by imposing sparsity in the factor loadings, in Chapter VII, we explore a different way to improve interpretability of PCA. The new method aims to capture natural "block" structures in highly correlated variables. For example, the spectra exhibit high correlations within the high and low frequency regions, thus giving rise to such a block structure. Something analogous occurs in image data, where the background forms one natural block, and the foreground one or more such blocks. In such cases, the factor loadings within the same block tend to be of similar magnitude. The proposed method is geared towards exploring such block structures and producing sparse loadings which are further fused to the same value with a block, thus significantly aiding interpretation of the results.

1.3 Organization of the Chapters

The dissertation is organized as follows. Chapters II, III, IV and V study the estimation problems in graphical models. Specifically, Chapter II introduces the joint structure estimation method for learning multiple graphical models, Chapter III shows the asymptotic properties of the joint neighborhood selection method for estimating large-scale binary Markov networks, Chapter IV extends the estimator in Chapter III to multiple graphical models, and Chapter V develops a new graphical model for modeling the conditional dependence between ordinal variables.

The last two chapters consider group variable selection problems in unsupervised learning. Specifically, Chapter VI develops a new model-based clustering method that simultaneously selects the important variables and identifies the separability of the clusters with respect to each selected variable, while Chapter VII introduces a new sparse principal component analysis capturing the “blocking” structures in highly correlated variables.

CHAPTER II

Joint Estimation of Multiple Graphical Models

2.1 Introduction

The focus so far in the literature about graphical models has been on estimating a single Gaussian graphical model. However, in many applications it is more realistic to fit a collection of such models, due to the heterogeneity of the data involved. By heterogeneous data we mean data from several categories that share the same variables but differ in their dependence structure, with some edges common across all categories and other edges unique to each category. For example, consider gene networks describing different subtypes of the same cancer: there are some shared pathways across different subtypes, and there are also links that are unique to a particular subtype. Another example from text mining, which is discussed in detail in Section 2.6, is word relationships inferred from webpages. In our example, the webpages are collected from university computer science departments, and the different categories correspond to faculty, student, course, etc. In such cases, borrowing strength across different categories by jointly estimating these models could reveal a common structure and reduce the variance of the estimates, especially when the number of samples is relatively small. To accomplish this joint estimation, we propose a method that links the estimation of separate graphical models through a hierarchical penalty. Its main advantage is the ability to discover a common structure and jointly estimate common links across graphs, which leads to improvements over fitting separate models, since it borrows information from other related graphs. While in this paper we focus on continuous data, this methodology can be extended to graphical models with categorical variables; fitting such models to a single graph has been considered by Kolar and Xing (2008), Hoefling and Tibshirani (2009) and Ravikumar et al. (2009).

2.2 Estimation of Single Graphical Models

Suppose we have a heterogeneous data set with p variables and K categories. The kth category contains n_k observations $(\boldsymbol{x}_1^{(k)}, \ldots, \boldsymbol{x}_{n_k}^{(k)})^\top$, where each $\boldsymbol{x}_i^{(k)} = (x_{i,1}^{(k)}, \ldots, x_{i,p}^{(k)})$ is a p-dimensional row vector. Without loss of generality, we assume the observations in the same category are centered along each variable, i.e., $\sum_{i=1}^{n_k} x_{i,j}^{(k)} = 0$ for all $j = 1, \ldots, p$ and $k = 1, \ldots, K$. We further assume that $\boldsymbol{x}_1^{(k)}, \ldots, \boldsymbol{x}_{n_k}^{(k)}$ are an independent and identically distributed sample from a p-variate Gaussian distribution with mean zero, without loss of generality since the data are centered, and covariance matrix $\boldsymbol{\Sigma}^{(k)}$. Let $\boldsymbol{\Omega}^{(k)} = (\boldsymbol{\Sigma}^{(k)})^{-1} = (\omega_{j,j'}^{(k)})_{p\times p}$. The log-likelihood of the observations in the kth category is

$$l(\boldsymbol{\Omega}^{(k)}) = -\frac{n_k}{2}\log(2Pi) + \frac{n_k}{2}\Big[\log\{\det(\boldsymbol{\Omega}^{(k)})\} - \mathrm{trace}(\widehat{\boldsymbol{\Sigma}}^{(k)}\boldsymbol{\Omega}^{(k)})\Big],$$

where $\widehat{\boldsymbol{\Sigma}}^{(k)}$ is the sample covariance matrix for the kth category, and $\det(\cdot)$ and $\mathrm{trace}(\cdot)$ are the determinant and the trace of a matrix, respectively.

The most direct way to deal with such heterogeneous data is to estimate K individual graphical models. We can compute a separate ℓ_1-regularized estimator for each category k $(k = 1, \ldots, K)$, by solving

$$\min_{\boldsymbol{\Omega}^{(k)}} \ \mathrm{trace}(\widehat{\boldsymbol{\Sigma}}^{(k)}\boldsymbol{\Omega}^{(k)}) - \log\{\det(\boldsymbol{\Omega}^{(k)})\} + \lambda_k \sum_{j\neq j'} |\omega_{j,j'}^{(k)}|, \qquad (2.1)$$

where the minimum is taken over symmetric positive definite matrices. The ℓ_1 penalty shrinks some of the off-diagonal elements in $\mathbf{\Omega}^{(k)}$ to zero and the tuning parameter λ_k controls the degree of the sparsity in the estimated inverse covariance matrix. Problem (2.1) can be efficiently solved by existing algorithms such as graphical lasso (Friedman et al., 2008). We will refer to this approach as the separate estimation method and use it as a benchmark to compare with the joint estimation method we propose next.

2.3 Methodology

2.3.1 The Joint Estimation Method

To improve estimation in cases where graphical models for different categories may share some common structure, we propose a joint estimation method. First, we reparametrize each off-diagonal element $\omega_{j,j'}^{(k)}$ as $\omega_{j,j'}^{(k)} = \theta_{j,j'}\gamma_{j,j'}^{(k)}$ $(1 \leq j \neq j' \leq p;\ k = 1, \ldots, K)$. An analogous parametrization in a dimension reduction setting was used in Michailidis and de Leeuw (2001). To avoid sign ambiguity between θ and γ, we restrict $\theta_{j,j'} \geq 0$, $1 \leq j \neq j' \leq p$. To preserve symmetry, we require that $\theta_{j,j'} = \theta_{j',j}$ and $\gamma_{j,j'}^{(k)} = \gamma_{j',j}^{(k)}$ $(1 \leq j \neq j' \leq p;\ k = 1, \ldots, K)$. For all diagonal elements, we also require $\theta_{j,j} = 1$ and $\gamma_{j,j}^{(k)} = \omega_{j,j}^{(k)}$ $(j = 1, \ldots, p;\ k = 1, \ldots, K)$. This decomposition treats $(\omega_{j,j'}^{(1)}, \ldots, \omega_{j,j'}^{(K)})$ as a group, with the common factor $\theta_{j,j'}$ controlling the presence of the link between nodes j and j' in any of the categories, and $\gamma_{j,j'}^{(k)}$ reflects the differences between categories. Let $\mathbf{\Theta} = (\theta_{j,j'})_{p\times p}$ and $\mathbf{\Gamma}^{(k)} = (\gamma_{j,j'}^{(k)})_{p\times p}$. To estimate this model, we propose the following penalized criterion subject to all constraints mentioned above:

$$\min_{\mathbf{\Theta},(\mathbf{\Gamma}^{(k)})_{k=1}^{K}} \sum_{k=1}^{K} \Big[\mathrm{trace}(\widehat{\mathbf{\Sigma}}^{(k)}\mathbf{\Omega}^{(k)}) - \log\{\det(\mathbf{\Omega}^{(k)})\}\Big] + \eta_1 \sum_{j\neq j'} \theta_{j,j'} + \eta_2 \sum_{j\neq j'}\sum_{k=1}^{K} |\gamma_{j,j'}^{(k)}| \quad (2.2)$$

where η_1 and η_2 are two tuning parameters. The first one, η_1, controls the sparsity of the common factors $\theta_{j,j'}$'s and can effectively identify the common zero elements across $\mathbf{\Omega}^{(1)}, \ldots, \mathbf{\Omega}^{(K)}$; i.e., if $\theta_{j,j'}$ is shrunk to zero, there will be no link between nodes j and j' in any of the K graphs. If $\theta_{j,j'}$ is not zero, some of the $\gamma_{j,j'}^{(k)}$'s, and hence some of the $\omega_{j,j'}^{(k)}$'s, can still be set to zero by the second penalty. This allows graphs belonging to different categories to have different structures. This decomposition has also been used by Zhou and Zhu (2010) for group variable selection in regression problems.

Criterion (2.2) involves two tuning parameters η_1 and η_2; it turns out that this could be reduced to an equivalent problem with a single tuning parameter. Specifically, consider

$$\min_{\mathbf{\Theta},(\mathbf{\Gamma}^{(k)})_{k=1}^{K}} \sum_{k=1}^{K} \left[\operatorname{trace}(\widehat{\mathbf{\Sigma}}^{(k)}\mathbf{\Omega}^{(k)}) - \log\{\det(\mathbf{\Omega}^{(k)})\}\right] + \sum_{j\neq j'} \theta_{j,j'} + \eta \sum_{j\neq j'} \sum_{k=1}^{K} |\gamma_{j,j'}^{(k)}|, \quad (2.3)$$

where $\eta = \eta_1\eta_2$. For two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of the same size, we denote their Schur–Hadamard product by $\boldsymbol{A} \cdot \boldsymbol{B}$. Then, criteria (2.2) and (2.3) are equivalent in the following sense:

Lemma II.1. *Let* $\{\widehat{\mathbf{\Theta}}^{*}, (\widehat{\mathbf{\Gamma}}^{(k)*})_{k=1}^{K}\}$ *be a local minimizer of criterion* (2.3)*. Then, there exists a local minimizer of criterion* (2.2)*, denoted as* $\{\widehat{\mathbf{\Theta}}^{**}, (\widehat{\mathbf{\Gamma}}^{(k)**})_{k=1}^{K}\}$*, such that* $\widehat{\mathbf{\Theta}}^{**} \cdot \widehat{\mathbf{\Gamma}}^{(k)**} = \widehat{\mathbf{\Theta}}^{*} \cdot \widehat{\mathbf{\Gamma}}^{(k)*}$ *for all* $k = 1, \ldots, K$*. Similarly, if* $\{\widehat{\mathbf{\Theta}}^{**}, (\widehat{\mathbf{\Gamma}}^{(k)**})_{k=1}^{K}\}$ *is a local minimizer of criterion* (2.2)*, then there exists a local minimizer of criterion* (2.3)*, denoted as* $\{\widehat{\mathbf{\Theta}}^{*}, (\widehat{\mathbf{\Gamma}}^{(k)*})_{k=1}^{K}\}$*, such that* $\widehat{\mathbf{\Theta}}^{**} \cdot \widehat{\mathbf{\Gamma}}^{(k)**} = \widehat{\mathbf{\Theta}}^{*} \cdot \widehat{\mathbf{\Gamma}}^{(k)*}$ *for all* $k = 1, \ldots, K$*.*

The proof follows closely the proof of the Lemma in Zhou and Zhu (2010) and is omitted. This result implies that in practice, instead of tuning two parameters η_1 and η_2, we only need to tune one parameter η, which reduces the overall computational cost.

2.3.2 The Algorithm

First we reformulate the problem (2.3) in a more convenient form for computational purposes.

Lemma II.2. *Let* $(\widehat{\mathbf{\Omega}}^{(k)})_{k=1}^K$ *be a local minimizer of*

$$\min_{(\mathbf{\Omega}^{(k)})_{k=1}^K} \sum_{k=1}^K \Big[\text{trace}(\widehat{\mathbf{\Sigma}}^{(k)}\mathbf{\Omega}^{(k)}) - \log\{\det(\mathbf{\Omega}^{(k)})\}\Big] + \lambda \sum_{j\neq j'} (\sum_{k=1}^K |\omega_{j,j'}^{(k)}|)^{1/2}, \tag{2.4}$$

where $\lambda = 2\eta^{1/2}$. *Then, there exists a local minimizer of* (2.3), $\{\widehat{\mathbf{\Theta}}, (\widehat{\mathbf{\Gamma}}^{(k)})_{k=1}^K\}$, *such that* $\widehat{\mathbf{\Omega}}^{(k)} = \widehat{\mathbf{\Theta}} \cdot \widehat{\mathbf{\Gamma}}^{(k)}$, *for all* $k = 1, \ldots, K$. *On the other hand, if* $\{\widehat{\mathbf{\Theta}}, (\widehat{\mathbf{\Gamma}}^{(k)})_{k=1}^K\}$ *is a local minimizer of* (2.3), *then there also exists a local minimizer of* (2.4), $(\widehat{\mathbf{\Omega}}^{(k)})_{k=1}^K$, *such that* $\widehat{\mathbf{\Omega}}^{(k)} = \widehat{\mathbf{\Theta}} \cdot \widehat{\mathbf{\Gamma}}^{(k)}$, *for all* $k = 1, \ldots, K$.

The proof follows closely the proof of the Lemma in Zhou and Zhu (2010) and is omitted. To optimize (2.4) we use an iterative approach based on Local Linear Approximation (Zou and Li, 2008). Specifically, letting $(\omega_{j,j'}^{(k)})^{(t)}$ denote the estimates from the previous iteration t, we approximate $(\sum_{k=1}^K |\omega_{j,j'}^{(k)}|)^{1/2} \approx \sum_{k=1}^K |\omega_{j,j'}^{(k)}|/\{\sum_{k=1}^K |(\omega_{j,j'}^{(k)})^{(t)}|\}^{1/2}$ Thus, at the $(t+1)$th iteration, problem (2.4) is decomposed into K individual optimization problems:

$$(\mathbf{\Omega}^{(k)})^{(t+1)} = \arg\min_{\mathbf{\Omega}^{(k)}} \Big[\text{trace}(\widehat{\mathbf{\Sigma}}^{(k)}\mathbf{\Omega}^{(k)}) - \log\{\det(\mathbf{\Omega}^{(k)})\}\Big] + \lambda \sum_{j\neq j'} \tau_{j,j'}^{(k)} |\omega_{j,j'}^{(k)}| \tag{2.5}$$

where $\tau_{j,j'}^{(k)} = \{\sum_{k=1}^K |(\omega_{j,j'}^{(k)})^{(t)}|\}^{-1/2}$. Criterion (2.5) is exactly the sparse inverse covariance matrix estimation problem with weighted ℓ_1 penalty; the solution can be efficiently computed using the graphical lasso algorithm of Friedman et al. (2008). For numerical stability, we threshold $\big\{\sum_{k=1}^K |(\omega_{j,j'}^{(k)})^{(t)}|\big\}^{1/2}$ at 10^{-10}. In summary, the proposed algorithm for solving (2.4) is:

Step 0. Initialize $\widehat{\mathbf{\Omega}}^{(k)} = (\widehat{\mathbf{\Sigma}}^{(k)} + \nu \boldsymbol{I}_p)^{-1}$ for all $k = 1, \ldots, K$, where $\boldsymbol{I}_p$ is the identity matrix and the constant ν is chosen to guarantee $\widehat{\mathbf{\Sigma}}^{(k)} + \nu \boldsymbol{I}_p$ is positive definite;

Step 1. Update $\widehat{\mathbf{\Omega}}^{(k)}$ by (2.5) for all $k = 1, \dots, K$ using graphical lasso;

Step 2. Repeat Step 1 until convergence is achieved.

2.3.3 Model Selection

The tuning parameter λ in (2.4) controls the sparsity of the resulting estimator. It can be selected either by some type of Bayesian information criterion or through cross-validation. The former balances the goodness of fit of the model and its complexity, while the latter seeks to optimize its predictive power. Specifically, we define the Bayesian information criterion for the proposed joint estimation method as

$$\mathrm{BIC}(\lambda) = \sum_{k=1}^{K} \Big[\mathrm{trace}(\widehat{\mathbf{\Sigma}}^{(k)} \widehat{\mathbf{\Omega}}_{\lambda}^{(k)}) - \log\{\det(\widehat{\mathbf{\Omega}}_{\lambda}^{(k)})\} + \log(n_k) df_k \Big],$$

where $\widehat{\mathbf{\Omega}}_{\lambda}^{(1)}, \dots, \widehat{\mathbf{\Omega}}_{\lambda}^{(K)}$ are the estimates from (2.4) with tuning parameter λ and the degrees of freedom are defined as $df_k = \#\{(j, j') : j < j', \widehat{\omega}_{j,j'}^{(k)} \neq 0\}$. The cross-validation method randomly splits the data set into D segments with equal sizes. For the kth category, we denote the sample covariance matrix using the data in the dth segment ($d = 1, \dots, D$) by $\widehat{\mathbf{\Sigma}}^{(k,d)}$ and the inverse covariance matrix estimated using all the data excluding those in the dth segment and the tuning parameter λ by $\widehat{\mathbf{\Omega}}_{\lambda}^{(k,-d)}$. Then we choose λ that minimizes the average predictive negative log-likelihood as follows:

$$\mathrm{CV}(\lambda) = \sum_{d=1}^{D} \sum_{k=1}^{K} \Big[\mathrm{trace}(\widehat{\mathbf{\Sigma}}^{(k,d)} \widehat{\mathbf{\Omega}}_{\lambda}^{(k,-d)}) - \log\{\det(\widehat{\mathbf{\Omega}}_{\lambda}^{(k,-d)})\} \Big].$$

Cross-validation can in general be expected to be more accurate than the heuristic Bayesian information criterion, but it is much more computationally intensive, which is why we consider both options. We provide some comparisons between the two tuning parameter selection methods in Section 2.5.

2.4 Asymptotic Properties

Next, we derive the asymptotic properties of the joint estimation method, including consistency, as well as sparsistency, when both p and n go to infinity and the tuning parameter goes to 0 at a certain rate. First, we introduce the necessary notation and state certain regularity conditions on the true precision matrices $(\mathbf{\Omega}_0^{(1)}, \ldots, \mathbf{\Omega}_0^{(K)})$, where $\mathbf{\Omega}_0^{(k)} = (\omega_{0,j,j'}^{(k)})_{p\times p}$ $(k = 1, \ldots, K)$.

Let $T_k = \{(j, j') : j \neq j', \omega_{j,j'}^{(k)} \neq 0\}$ be the set of indices of all nonzero off-diagonal elements in $\mathbf{\Omega}^{(k)}$, and let $T = T_1 \cup \cdots \cup T_K$. Let $q_k = |T_k|$ and $q = |T|$ be the cardinalities of T_k and T, respectively. In general, T_k and q_k depend on p. In addition, let $\|\cdot\|_F$ and $\|\cdot\|$ be the Frobenius norm and the 2-norm of matrices, respectively. We assume that the following regularity conditions hold:

(A) there exist constants τ_1, τ_2 such that for all $p \geq 1$ and $k = 1, \ldots, K$,

$$0 < \tau_1 < Phi_{\min}(\mathbf{\Omega}_0^{(k)}) \leq Phi_{\max}(\mathbf{\Omega}_0^{(k)}) < \tau_2 < \infty$$

where $Phi_{\min}$ and $Phi_{\max}$ indicate the minimal and maximal eigenvalues;

(B) there exists a constant $\tau_3 > 0$ such that

$$\min_{k=1,\ldots,K} \min_{(j,j')\in T_k} |\omega_{0,j,j'}^{(k)}| \geq \tau_3 \ .$$

Condition (A) is a standard one, also used in Bickel and Levina (2008) and Rothman et al. (2008), that guarantees that the inverse exists and is well conditioned. Condition (B) ensures that non-zero elements are bounded away from 0.

Theorem II.3. *(Consistency) Suppose conditions (A) and (B) hold,* $(p+q)(\log p)/n = o(1)$ *and* $\Lambda_1\{(\log p)/n\}^{1/2} \leq \lambda \leq \Lambda_2\{(1+p/q)(\log p)/n\}^{1/2}$ *for some positive constants*

Λ_1 and Λ_2. Then there exists a local minimizer $(\widehat{\boldsymbol{\Omega}}^{(k)})_{k=1}^K$ of (2.4), such that

$$\sum_{k=1}^{K} \|\widehat{\boldsymbol{\Omega}}^{(k)} - \boldsymbol{\Omega}_0^{(k)}\|_F = O_P\Big[\Big\{\frac{(p+q)\log p}{n}\Big\}^{1/2}\Big].$$

Theorem II.4. *(Sparsistency) Suppose all conditions in Theorem II.3 hold. We further assume $\sum_{k=1}^K \|\widehat{\boldsymbol{\Omega}}^{(k)} - \boldsymbol{\Omega}_0^{(k)}\|^2 = O_P(\eta_n)$, where $\eta_n \to 0$ and $\{(\log p)/n\}^{1/2} + \eta_n^{1/2} = O(\lambda)$. Then with probability tending to 1, the local minimizer $(\widehat{\boldsymbol{\Omega}}^{(k)})_{k=1}^K$ in Theorem II.3 satisfies $\widehat{\omega}_{j,j'}^{(k)} = 0$ for all $(j, j') \in T_k^c$, $k = 1, \ldots, K$.*

This theorem is analogous to Theorem 2 in Lam and Fan (2009). The consistency requires both an upper and a lower bound on λ, whereas sparsistency requires consistency and an additional lower bound on λ. To make the bounds compatible, we require $\{(\log p)/n\}^{1/2} + \eta_n^{1/2} = O(\{(1+p/q)(\log p)/n\}^{1/2})$. Since η_n is the rate of convergence in the operator norm, we can bound it using the fact that $\|M\|_F^2/p \le \|M\|^2 \le \|M\|_F^2$. This leads to two extreme cases. In the worst-case scenario, $\sum_k \|\widehat{\boldsymbol{\Omega}}^{(k)} - \boldsymbol{\Omega}_0^{(k)}\|$ has the same rate as $\sum_k \|\widehat{\boldsymbol{\Omega}}^{(k)} - \boldsymbol{\Omega}_0^{(k)}\|_F$ and thus $\eta_n = O\{(p+q)(\log p)/n\}$. The two bounds are compatible only when $q = O(1)$. In best-case scenario, $\sum_k \|\widehat{\boldsymbol{\Omega}}^{(k)} - \boldsymbol{\Omega}_0^{(k)}\|$ has the same rate as $\sum_k \|\widehat{\boldsymbol{\Omega}}^{(k)} - \boldsymbol{\Omega}_0^{(k)}\|_F/p^{1/2}$. Then, $\eta_n = O\{(1+q/p)(\log p)/n\}$ and we have both consistency and sparsistency as long as $q = O(p)$.

2.5 Numerical Evaluation

2.5.1 Simulation Settings

In this section, we assess the performance of the joint estimation method on three types of simulated networks: a chain, a nearest-neighbor, and a scale-free network. In all cases, we set $p = 100$ and $K = 3$. For each $k = 1, \ldots, K$, we generate $n_k = 100$ independently and identically distributed observations from a multivariate normal distribution $N\{\mathbf{0}, (\boldsymbol{\Omega}^{(k)})^{-1}\}$, where $\boldsymbol{\Omega}^{(k)}$ is the inverse covariance matrix of the kth category. The details of the three simulated examples are described as follows.

In the first example, we follow the simulation setup in Fan et al. (2009) to generate a chain network, which corresponds to a tridiagonal inverse covariance matrix. The covariance matrices $\mathbf{\Sigma}^{(k)}$ are constructed as follows: let the (j, j')th element $\sigma_{j,j'}^{(k)} = \exp(-|s_j - s_{j'}|/2)$, where $s_1 < s_2 < \cdots < s_p$ and

$$s_j - s_{j-1} \sim \text{Uniform}(0{\cdot}5, 1),\ j = 2, \ldots, p$$

Further, let $\mathbf{\Omega}^{(k)} = \left(\mathbf{\Sigma}^{(k)}\right)^{-1}$. The K precision matrices generated by this procedure share the same pattern of zeros, i.e., the common structure, but the values of their non-zero off-diagonal elements may be different. The left panel of Fig. 2.1 shows the common link structure across the K categories. Further, we add heterogeneity to the common structure by creating additional individual links as follows: for each $\mathbf{\Omega}^{(k)}$ $(k = 1, \ldots, K)$, we randomly pick a pair of symmetric zero elements and replace them with a value uniformly generated from the $[-1, -0{\cdot}5] \cup [0{\cdot}5, 1]$ interval. This procedure is repeated ρM times, where M is the number of off-diagonal non-zero elements in the lower triangular part of $\mathbf{\Omega}^{(k)}$ and ρ is the ratio of the number of individual links to the number of common links. In the simulations, we considered values of ρ=0, 1/4, 1 and 4, thus gradually increasing the proportion of individual links.

In the second example, the nearest-neighbor networks are generated by modifying the data generating mechanism described in Li and Gui (2006). Specifically, we generate p points randomly on a unit square, calculate all $p(p-1)/2$ pairwise distances, and find m nearest neighbors of each point in terms of this distance. The nearest neighbor network is obtained by linking any two points that are m-nearest neighbors of each other. The integer m controls the degree of sparsity of the network and the value $m = 5$ was chosen in our study. The middle panel of Fig. 2.1 illustrates a realization of the common structure of a nearest-neighbor network. Subsequently, K individual graphs were generated, by adding some individual links to the common graph with $\rho = 0, 1/4, 1, 4$ by the same method as described in Example 1, with values for the individual links $\omega_{j,j'}^{(k)}$ generated from a uniform distribution on $[-1, -0{\cdot}5] \cup [0{\cdot}5, 1]$.

In the last example, we generate the common structure of a scale-free network using the Barabasi–Albert algorithm (Barabasi and Albert, 1999); a realization is depicted in the right panel of Fig. 2.1. The individual links in the kth network ($k = 1, \ldots, K$), are randomly added as before, with $\rho = 0, 1/4, 1, 4$ and the associated elements in $\mathbf{\Omega}^{(k)}$ are generated uniformly on $[-1, -0{\cdot}5] \cup [0{\cdot}5, 1]$.

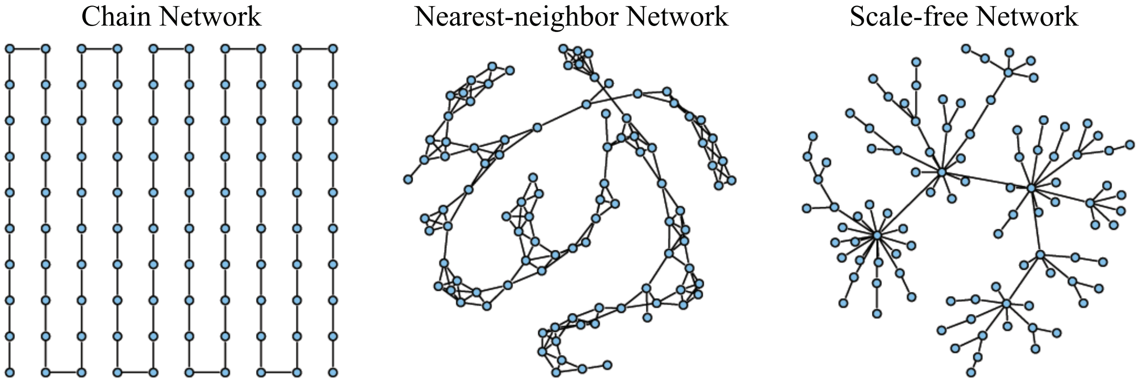


Figure 2.1: The common links present in all categories in the three simulated networks.

We compare the joint estimation method to the method that estimates each category separately via (2.1). A number of metrics are used to assess performance, including receiver operating characteristic curves, average entropy loss, average Frobenius loss, average false positive and average false negative rates, and the average rate of mis-identified common zeros among the categories. For the receiver operating characteristic curve, we plot sensitivity, the average proportion of correctly detected links, against the average false positive rate over a range of values of the tuning parameter

λ. The average entropy loss and average Frobenius loss are defined as:

$$\begin{aligned} EL &= \frac{1}{K}\sum_{k=1}^{K} \mathrm{trace}\{(\mathbf{\Omega}^{(k)})^{-1}\widehat{\mathbf{\Omega}}^{(k)}\} - \log[\det\{(\mathbf{\Omega}^{(k)})^{-1}\widehat{\mathbf{\Omega}}^{(k)}\}] - p \ , \\ FL &= \frac{1}{K}\sum_{k=1}^{K} \|\mathbf{\Omega}^{(k)} - \widehat{\mathbf{\Omega}}^{(k)}\|_F^2 / \|\mathbf{\Omega}^{(k)}\|_F^2 \ . \end{aligned} \tag{2.6}$$

The average false positive rate gives the proportion of false discoveries, that is, true zeros estimated as non-zero; the average false negative rate gives the proportion of off-diagonal non-zero elements estimated as zero; and the common zeros error rate gives the proportion of common zeros across $\mathbf{\Omega}^{(1)}, \ldots, \mathbf{\Omega}^{(K)}$ estimated as non-zero. The respective formal definitions are:

$$\begin{aligned} FP &= \frac{1}{K}\sum_{k=1}^{K} \frac{\sum_{1\le j<j'\le p} \mathrm{I}(\omega_{j,j'}^{(k)} = 0, \widehat{\omega}_{j,j'}^{(k)} \neq 0)}{\sum_{1\le j<j'\le p} \mathrm{I}(\omega_{j,j'}^{(k)} = 0)} \ , \\ FN &= \frac{1}{K}\sum_{k=1}^{K} \frac{\sum_{1\le j<j'\le p} \mathrm{I}(\omega_{j,j'}^{(k)} \neq 0, \widehat{\omega}_{j,j'}^{(k)} = 0)}{\sum_{1\le j<j'\le p} \mathrm{I}(\omega_{j,j'}^{(k)} \neq 0)} \ , \\ CZ &= \frac{\sum_{1\le j<j'\le p} \mathrm{I}(\sum_{k=1}^{K} |\omega_{j,j'}^{(k)}| = 0, \sum_{k=1}^{K} |\widehat{\omega}_{j,j'}^{(k)}| \neq 0)}{\sum_{1\le j<j'\le p} \mathrm{I}(\sum_{k=1}^{K} |\omega_{j,j'}^{(k)}| = 0)} \ . \end{aligned} \tag{2.7}$$

2.5.2 Simulation Results

Figure 2.2 shows the estimated receiver operating characteristic curves averaged over 50 replications for all three simulated examples, obtained by varying the tuning parameter. It can be seen that the curves estimated by the joint estimation method dominate those of the separate estimation method when the proportion of individual links is low. As ρ increases, the structures become more and more different, and the joint and separate methods move closer together, with the separate method eventually slightly outperforming the joint method at $\rho = 4$, although the results are still fairly similar. This is precisely as it should be, since the joint estimation method has the biggest advantage with the most overlap in structure. In order to assess the variability of the two methods, we drawn the boxplots of the sensitivity of the two models with the false positive rate controlled at 5% and the results indicate that as long as there is a sustantial common structure, the joint method is superior to the separate method and the difference is statistically significant.

Table 2.1 summarizes the results based on 50 replications with the tuning parameter selected by the Bayesian information criterion and cross-validation as described in Section 2.3.3. In general, the joint estimation method produces lower entropy and Frobenius norm losses for both model selection criteria, with the difference most pronounced at low values of ρ. For the joint method, the two model selection criteria exhibit closer agreement in false positive and negative rates and the proportion of misidentified common zeros. For the separate method, however, cross-validation tends to select more false positive links which result in more misidentified common zeros.

2.6 University Webpages Example

The data set was collected in 1997 and includes webpages from computer science departments at Cornell, University of Texas, University of Washington, and University of Wisconsin. The original data has been preprocessed using standard text processing procedures, such as removing stop-words and stemming the words. The preprocessed data set can be downloaded from `http://web.ist.utl.pt/~acardoso/datasets/`. The webpages were manually classified into seven categories, from which we selected the four largest ones for our analysis: student, faculty, course and project, with 544, 374, 310 and 168 webpages, respectively. The log-entropy weighting method (Dumais, 1991) was used to calculate the term-document matrix $\boldsymbol{X} = (x_{i,j})_{n \times p}$, with n and p denoting the number of webpages and distinct terms, respectively. Let $f_{i,j}$ $(i = 1, \ldots, n;\; j = 1, \ldots, p)$ be the number of times the jth term appears in the ith webpage and let $p_{i,j} = f_{i,j} / \sum_{i=1}^{n} f_{i,j}$. Then, the log-entropy weight of the jth term

Table 2.1: Results from the three simulated examples. S and J stand for the separate and the joint methods, respectively. In each cell, the numbers before and after the slash correspond to the results from selected by Bayesian information criterion and cross-validation, respectively. EL, FL, FN, FP and CZ are defined in equation (2.6) and (2.7). ρ: ratio of the number of individual links to the number of common links.

Example	ρ	Method	EL	FL	FN (%)	FP (%)	CZ (%)
Chain	0	S	20.7 / 21.9	0.5 / 0.5	0.8 / 0.1	5.7 / 21.8	14.5 / 51.0
		J	12.8 / 6.6	0.3 / 0.3	0.0 / 0.0	4.3 / 0.5	7.0 / 1.2
	1/4	S	21.3 / 16.6	0.5 / 0.5	41.3 / 9.0	1.3 / 18.7	3.8 / 46.0
		J	9.5 / 8.7	0.3 / 0.3	15.6 / 17.6	1.7 / 0.7	3.2 / 1.4
	1	S	23.0 / 17.1	0.5 / 0.5	73.7 / 24.4	0.7 / 18.8	1.9 / 46.4
		J	12.5 / 12.4	0.4 / 0.4	44.2 / 45.8	1.6 / 1.1	3.0 / 2.0
	4	S	29.8 / 20.2	0.6 / 0.5	97.3 / 47.5	0.1 / 19.5	0.3 / 47.8
		J	20.0 / 20.7	0.5 / 0.5	75.5 / 76.2	1.9 / 1.8	3.2 / 3.0
NN	0	S	11.9 / 15.9	0.4 / 0.5	40.1 / 33.5	2.2 / 16.1	6.1 / 40.5
		J	6.1 / 11.3	0.3 / 0.4	18.5 / 52.7	1.6 / 0.6	3.2 / 1.3
	1/4	S	13.9 / 17.1	0.4 / 0.5	44.0 / 32.5	2.4 / 17.6	6.9 / 43.9
		J	8.1 / 14.5	0.3 / 0.4	27.4 / 57.5	1.7 / 1.0	2.9 / 1.7
	1	S	18.5 / 18.0	0.5 / 0.5	48.5 / 45.3	4.0 / 17.8	11.2 / 44.3
		J	13.0 / 19.0	0.4 / 0.5	40.0 / 77.3	2.8 / 1.2	3.8 / 2.0
	4	S	24.8 / 20.1	0.5 / 0.5	98.7 / 65.5	0.1 / 18.1	0.3 / 44.9
		J	19.3 / 23.8	0.7 / 0.5	80.8 / 95.0	3.2 / 1.0	4.8 / 1.6
Scale-free	0	S	16.9 / 15.5	0.5 / 0.5	20.7 / 6.4	1.9 / 17.1	5.3 / 42.1
		J	8.1 / 7.0	0.3 / 0.3	9.4 / 11.2	1.5 / 0.5	2.8 / 1.0
	1/4	S	17.1 / 14.5	0.5 / 0.4	49.6 / 17.5	1.2 / 16.6	3.7 / 41.8
		J	9.4 / 9.1	0.3 / 0.3	29.3 / 32.2	1.3 / 0.8	2.4 / 1.4
	1	S	22.3 / 18.1	0.5 / 0.5	51.8 / 22.5	2.8 / 19.3	8.2 / 47.4
		J	15.2 / 15.3	0.4 / 0.4	42.5 / 43.1	2.2 / 2.0	3.2 / 2.9
	4	S	27.9 / 20.0	0.6 / 0.5	99.6 / 49.6	0.0 / 19.1	0.0 / 47.0
		J	23.0 / 23.8	0.5 / 0.5	82.5 / 84.1	2.1 / 1.8	3.2 / 2.7

is defined as $e_j = 1+\sum_{i=1}^{n} p_{i,j}(\log p_{i,j})/\log n$. Finally, the term-document matrix $\boldsymbol{X}$ is defined as $x_{i,j} = e_j \log(1+f_{i,j})$ $(i = 1,\ldots,n;\ j = 1,\ldots,p)$. and it is normalized along each column. We applied the proposed joint estimation method to $n = 1396$ documents in the four largest categories and $p = 100$ terms with the highest log-entropy weights out of a total of 4800 terms. The resulting common network structure is shown in panel (A) of Fig. 2.3. The area of the circle representing a node is proportional to its log-entropy weight, while the thickness of an edge is proportional to the magnitude of the associated partial correlation. The plot reveals the existence of some high degree nodes, such as research, data, system, perform, that are part of the computer science vocabulary. Further, some standard phrases in computer science, such as home-page, comput-scienc, program-languag, data-structur, distribut-system and high-perform, have high partial correlations among their constituent words in all four categories. A few subgraphs extracted from the common network are shown in panels (B)–(D) of Fig. 2.3; each graph clearly has its own semantic meaning, which we loosely label as webpage generic, research area/lab and parallel programming.

The model also allows us to explore the heterogeneity between different categories. As an example, we show the graphs for the student and faculty categories in Fig. 2.4. It can be seen that terms teach and assist are only linked in the student category, since many graduate students are employed as teaching assistants. On the other hand, some term pairs only have links in the faculty category, such as select-public, faculti-student, assist-professor and associ-professor. Similarly, we illustrate the differences between the course and project categories in Fig. 2.5. Some teaching-related terms are linked only in the course category, such as office-hour, office-instructor and teach-assist, while pairs in the project category are connected to research, such as technolog-center, technolog-institut, research-scienc and research-inform. Overall, the model captures the basic common semantic structure of the websites, but also identifies meaningful differences across the various categories. When each category is estimated separately, individual links dominate, and the results are not as easy to interpret. The graphical models obtained by separate estimation are not shown for lack of space.

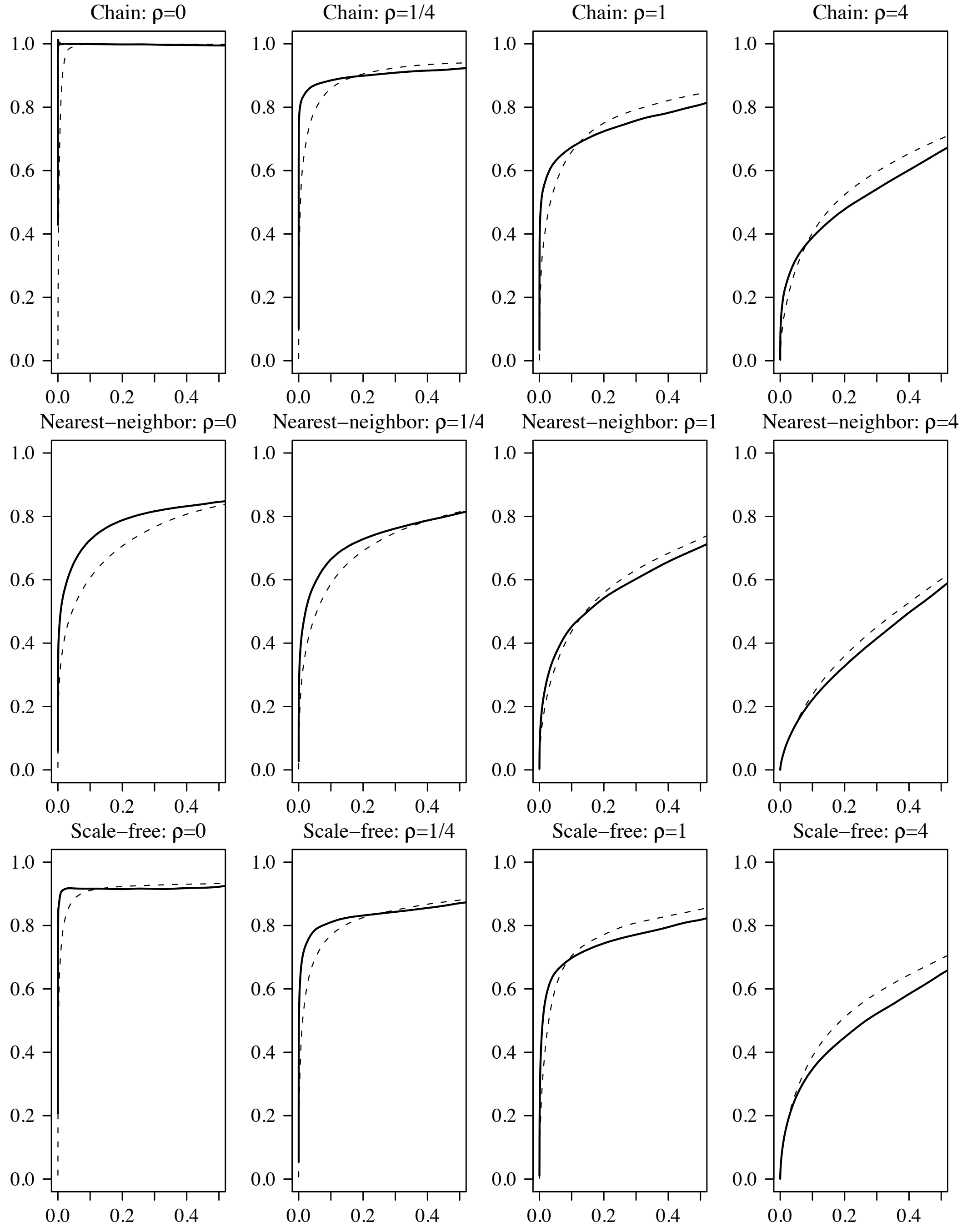


Figure 2.2: Receiver operating characteristic curves. The horizontal and vertical axes in each panel are false positive rate and sensitivity, respectively. The solid line corresponds to the joint estimation method, and the dashed line corresponds to the separate estimation method. ρ is the ratio of the number of individual links to the number of common links.

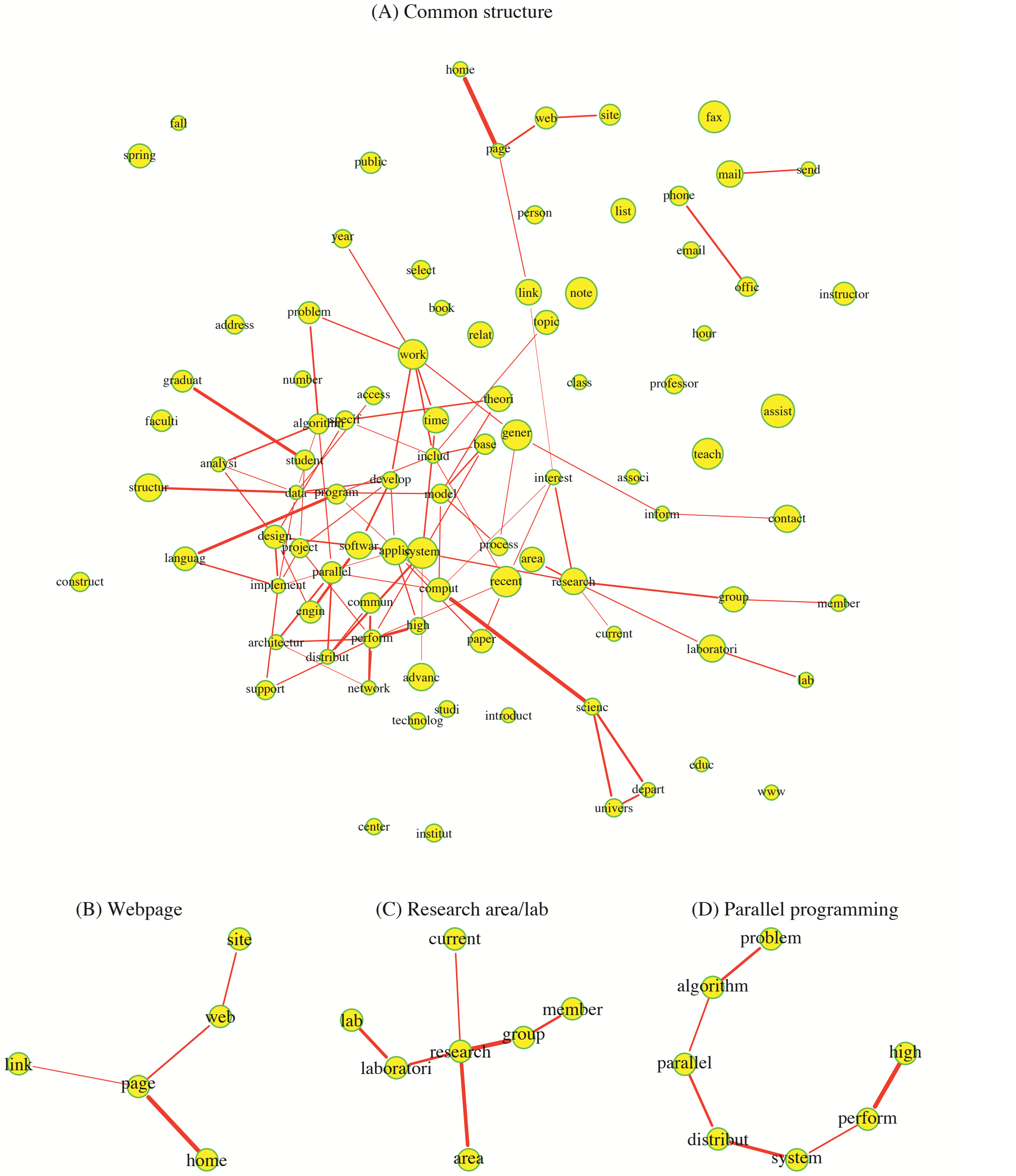


Figure 2.3: Common structure in the webpages data. Panel (A) shows the estimated common structure for the four categories. The nodes represent 100 terms with the highest log-entropy weights. The area of the circle representing a node is proportional to its log-entropy weight. The width of an edge is proportional to the magnitude of the associated partial correlation. Panels (B)–(D) show subgraphs extracted from the graph in panel (A).

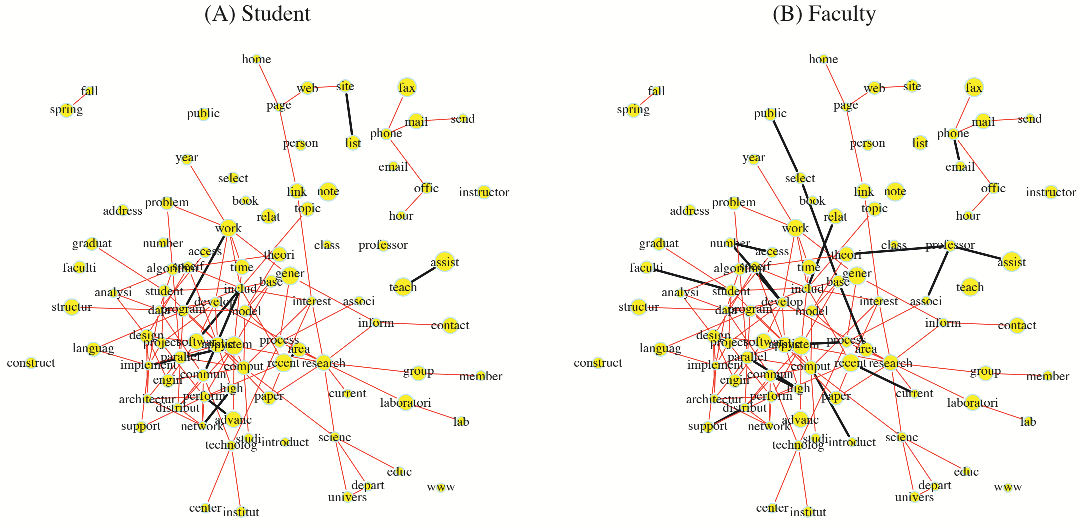


Figure 2.4: "Student" and "Faculty" graphs. The light lines are the links appearing in both categories, and the dark lines are the links only appearing in one category.

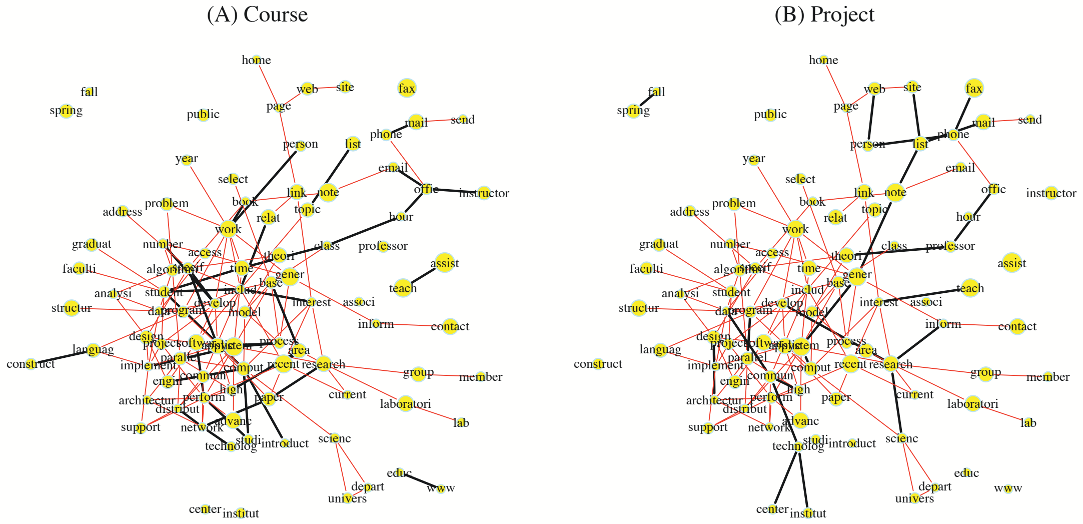


Figure 2.5: "Course" and "Project" graphs. The light lines are the links appearing in both categories, and the dark lines are the links only appearing in one category.

CHAPTER III

Asymptotic Properties of the Joint Neighborhood Selection Method for Estimating Categorical Markov Networks

3.1 Introduction

Hoefling and Tibshirani (2009) and Wang et al. (2009) proposed a joint neighborhood selection method to estimate high-dimensional Market network. This method provides a direct solution for parameter symmetrization by estimating all the regressions jointly. They simultaneously solve the p logistic regression problems and encourages the sparsity of the interaction parameters, thus automatically ensuring symmetry. The joint application of the ℓ_1 penalty allows for a more flexible degree distribution in the estimated graph, as explained in Section 2.3.

In this chapter, we show that the joint neighborhood selection algorithm in Hoefling and Tibshirani (2009) and Wang et al. (2009) leads to consistent parameter estimation and model selection under high-dimensional asymptotics. Moreover, we also apply the algorithm to a very different application and finally extend the algorithm to estimate graphical models with general categorical variables.

The remainder of the chapter is organized as follows. Section 3.2 reviews the structure estimation problem of Markov networks and introduce the joint neighborhood selection method. Section 2.4 establishes the theoretical properties of the method, including consistency of parameter estimation and network recovery. Section 3.4 applies the method to explore voting dependencies between senators in the 109th Congress. An extension to Markov networks with general categorical variables is discussed in Section 3.5.

3.2 Methodology

We focus initially on a Markov network for binary variables and discuss the extension to general categorical variables in Section 3.5. We start by setting up the problem and also discuss the joint neighborhood selection criterion proposed by Hoefling and Tibshirani (2009) and Wang et al. (2009).

Suppose we have p binary random variables $X_1, \ldots, X_p$, with $X_j \in \{1, 0\}$, $1 \leq j \leq p$, whose joint distribution has the following probability mass function:

$$\mathrm{f}(X_1, \ldots, X_p) = \frac{1}{\mathrm{Z}(\mathbf{\Theta})} \exp\Big(\sum_{j=1}^{p} \theta_{j,j} X_j + \sum_{1 \leq j < j' \leq p} \theta_{j,j'} X_j X_{j'} \Big), \tag{3.1}$$

where $\mathbf{\Theta} = (\theta_{j,j'})_{p \times p}$ is a symmetric matrix specifying the network structure.

Note that $\theta_{j,j}$, $1 \leq j \leq p$, corresponds to the main effect for variable X_j, whereas $\theta_{j,j'}$, $1 \leq j < j' \leq p$, corresponds to the interaction effect between variables X_j and $X_{j'}$. These $\theta_{j,j'}$'s reflect the structure of the underlying network. Specifically, if $\theta_{j,j'} = 0$, then X_j and $X_{j'}$ are conditionally independent given other variables and hence their corresponding nodes are *not* connected. Ravikumar et al. (2009) pointed out that one could consider *only* the pairwise interaction effects, since higher order interactions can be approximately converted to pairwise ones through the introduction of additional variables (Wainwright and Jordan, 2008). The partition function $\mathrm{Z}(\mathbf{\Theta}) = \sum_{X_j \in \{0,1\}, 1 \leq j \leq p} \exp(\sum_{j=1}^{p} \theta_{j,j} X_j + \sum_{1 \leq j < j' \leq p} \theta_{j,j'} X_j X_{j'})$ ensures that the probability mass function in (3.1) is a proper one, integrating to one.

The structure of the partition function with its 2^p terms renders optimizing (3.1) infeasible, except in toy problems. A strategy to overcome this difficulty is to use the pseudo-likelihood function to approximate the joint likelihood function associated with mass (3.1). Specifically, let $x_{i,j}$ be the i-th realization of variable X_j, then the pseudo-likelihood function can be written as follows:

$$\prod_{j=1}^{p}\prod_{i=1}^{n}\phi_{i,j}^{x_{i,j}}(1-\phi_{i,j})^{1-x_{i,j}}, \tag{3.2}$$

where $\phi_{i,j} = \mathrm{P}(x_{i,j} = 1|x_{i,k}, k \neq j; \theta_{j,k}, 1 \leq k \leq p) = \exp(\theta_{j,j} + \sum_{k\neq j}\theta_{j,k}x_{i,k})/\{1 + \exp(\theta_{j,j} + \sum_{k\neq j}\theta_{j,k}x_{i,k})\}$. It can be seen that this gives rise to a logistic regression problem where the j-th variable is taken as the response and is regressed on the remaining variables, and hence decomposes the problem into p separate logistic regressions, which are simple to solve.

In the joint neighborhood selection method proposed, Hoefling and Tibshirani (2009) and Wang et al. (2009) solve the following joint criterion problem:

$$\begin{aligned}
\max_{\boldsymbol{\Theta}} \quad & \sum_{j=1}^{p}\sum_{i=1}^{n}\Big[x_{i,j}\Big(\theta_{j,j} + \sum_{k\neq j}\theta_{j,k}x_{i,k}\Big) \\
& -\log\Big\{1 + \exp\Big(\theta_{j,j} + \sum_{k\neq j}\theta_{j,k}x_{i,k}\Big)\Big\}\Big] - \lambda\sum_{j<j'}|\theta_{j,j'}| \\
\text{subject to} \quad & \theta_{j,j'} = \theta_{j',j}, \ 1 \leq j < j' \leq p.
\end{aligned} \tag{3.3}$$

Notice that the penalty *jointly* imposes sparsity over all interaction effects, while the tuning parameter λ controls its degree. However, this method does not lead to solving p separate logistic problems due to the symmetry constraint $\theta_{j,j'} = \theta_{j',j}$. On the other hand, it reduces the number of parameters to be estimated by half, i.e., $p(p+1)/2$ for the joint method vs. p^2 for the neighborhood selection method.

Hoefling and Tibshirani (2009) and Wang et al. (2009) proposed an efficient iterative algorithm to solve this problem. The algorithm consists of two nested loops. In the outer loop, they follow the strategy in Friedman et al. (2010) to approximate the logistic log-likelihood in (3.3) by its Taylor series expansion. Specifically, we denote the estimate of $\theta_{j,j'}$ in the t-th iteration by $\theta_{j,j'}^{(t)}$, and write

$$\begin{aligned} & x_{i,j}\Big(\theta_{j,j} + \sum_{k\neq j}\theta_{j,k}x_{i,k}\Big) - \log\Big\{1 + \exp\Big(\theta_{j,j} + \sum_{k\neq j}\theta_{j,k}x_{i,k}\Big)\Big\} \\ \approx\ & -\frac{1}{2}w_{i,j}^{(t)}\Big(y_{i,j}^{(t)} - \theta_{j,j} - \sum_{k\neq j}\theta_{j,k}x_{i,k}\Big)^2 + C_{i,j}^{(t)}\ , \end{aligned} \tag{3.4}$$

where

$$\begin{aligned} p_{i,j}^{(t)} &= \frac{\exp(\theta_{j,j}^{(t)} + \sum_{k\neq j}\theta_{j,k}^{(t)}x_{i,k})}{1 + \exp(\theta_{j,j}^{(t)} + \sum_{k\neq j}\theta_{j,k}^{(t)}x_{i,k})}\ , \\ y_{i,j}^{(t)} &= \theta_{j,j}^{(t)} + \sum_{k\neq j}\theta_{j,k}^{(t)}x_{i,k} - \frac{p_{i,j}^{(t)} - x_{i,j}}{w_{i,j}^{(t)}}\ , \\ w_{i,j}^{(t)} &= p_{i,j}^{(t)}(1 - p_{i,j}^{(t)})\ , \end{aligned}$$

and $C_{i,j}^{(t)}$ is some constant unrelated to $\mathbf{\Theta}$. We define next the following quantities:

$$\begin{aligned} \boldsymbol{\theta} &= (\theta_{1,2}, \ldots, \theta_{j,j'}, \ldots, \theta_{p-1,p})^{\mathsf{T}}\ , \\ \boldsymbol{y}_j^{*} &= (\sqrt{w_{1,j}^{(t)}}y_{1,j}, \ldots, \sqrt{w_{n,j}^{(t)}}y_{n,j})^{\mathsf{T}}\ , \\ \boldsymbol{y}_j^{**} &= \boldsymbol{y}_j^{*} - \bar{y}_j, \text{where } \bar{y}_j = \frac{1}{n}\sum_{i=1}^{n}\sqrt{w_{i,j}^{(t)}}y_{i,j}^{(t)}\ , \\ \boldsymbol{x}_j^{*} &= (\sqrt{w_{1,j}^{(t)}}x_{1,j}, \ldots, \sqrt{w_{n,j}^{(t)}}x_{n,j})^{\mathsf{T}}\ , \\ \boldsymbol{x}_j^{**} &= \boldsymbol{x}_j^{*} - \bar{x}_j, \text{where } \bar{x}_j = \frac{1}{n}\sum_{i=1}^{n}\sqrt{w_{i,j}^{(t)}}x_{i,j}^{(t)}\ . \end{aligned} \tag{3.5}$$

We further define an $np \times 1$ column vector

$$\boldsymbol{\mathcal{X}}_{j,j'}^{**} = (\mathbf{0}_n{}^{\mathsf{T}}, \ldots, \mathbf{0}_n{}^{\mathsf{T}},\quad \underbrace{\boldsymbol{x}_{j'}^{**\mathsf{T}},}_{j\text{-th block}}\quad \mathbf{0}_n{}^{\mathsf{T}}, \ldots, \mathbf{0}_n{}^{\mathsf{T}},\quad \underbrace{\boldsymbol{x}_{j}^{**\mathsf{T}},}_{j'\text{-th block}}\quad \mathbf{0}_n{}^{\mathsf{T}}, \ldots, \mathbf{0}_n{}^{\mathsf{T}})^{\mathsf{T}}, \tag{3.6}$$

where $\mathbf{0}_n$ is an n-dimensional column vector of zeros. $\boldsymbol{\mathcal{X}}^{**}_{j,j'}$ consists of p blocks of size n, where the j-th block and the j'-th block are $\boldsymbol{x}^{**}_{j'}$ and $\boldsymbol{x}^{**}_{j}$, respectively, and all other blocks are zeros. Finally, let $\boldsymbol{\mathcal{Y}}^{**} = (\boldsymbol{y}_1^{**\intercal}, \ldots, \boldsymbol{y}_p^{**\intercal})^\intercal$ (an $np \times 1$ column vector) and $\boldsymbol{\mathcal{X}}^{**} = (\boldsymbol{\mathcal{X}}^{**}_{1,2}, \ldots, \boldsymbol{\mathcal{X}}^{**}_{j,j'}, \ldots, \boldsymbol{\mathcal{X}}^{**}_{p-1,p})$ (an $np \times p(p-1)/2$ matrix). Then, (3.3) can be rewritten as the following lasso problem:

$$\min_{\boldsymbol{\theta}} \frac{1}{2}\|\boldsymbol{\mathcal{Y}}^{**} - \boldsymbol{\mathcal{X}}^{**}\boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|_1. \tag{3.7}$$

In the inner loop of the algorithm, criterion (3.7) can be efficiently solved by shooting-type algorithms (Friedman et al., 2007). Letting $\widehat{\boldsymbol{\theta}}$ be the estimate obtained from (3.7), then for each $1 \le j \le p$, the main effects $\theta_{j,j}$'s in (3.4) are calculated as follows:

$$\widehat{\theta}_{j,j} = \frac{\bar{y}_j - \sum_{k \neq j} \widehat{\theta}_{j,k}\bar{x}_k}{\frac{1}{n}\sum_{i=1}^n \sqrt{w_{i,j}^{(t)}}}. \tag{3.8}$$

In summary, the algorithm consists of the following steps:

Step 1. Initialize $\boldsymbol{\Theta}^{(0)}$ by setting $\theta^{(0)}_{j,j'} = 0$ for all $1 \le j \neq j' \le p$ and $\theta^{(0)}_{j,j} = \log[p_j/(1-p_j)]$, where $p_j = \sum_{i=1}^n x_{i,j}/(n - \sum_{i=1}^n x_{i,j})$;

Step 2. Given the estimate in the t-th step, update $\boldsymbol{\Theta}^{(t+1)}$ by solving criteria (3.7) and (3.8);

Step 3. Repeat Step 2 until convergence.

3.3 Theoretical Properties

In this section, we present the asymptotic properties of the joint neighborhood selection method; the proofs can be found in the Appendices. Since in the Ising model the structure of the underlying network only depends on the interaction effects, we focus on the variant of the model with no main effects, which gives rise to the criterion

$$\max_{\boldsymbol{\theta}} \sum_{j=1}^{p}\sum_{i=1}^{n}\Big[x_{i,j}\Big(\sum_{j'\neq j}\theta_{j,j'}x_{i,j'}\Big) - \log\Big\{1+\exp\Big(\sum_{j'\neq j}\theta_{j,j'}x_{i,j'}\Big)\Big\}\Big] - \lambda\sum_{j<j'}|\theta_{j,j'}|, \quad (3.9)$$

where $\theta_{j,j'} = \theta_{j',j}$, $1 \le j < j' \le p$, and $\boldsymbol{\theta}$ is a vector with dimension $p(p-1)/2$ defined as $\boldsymbol{\theta} = (\theta_{1,2},\ldots,\theta_{j,j'},\ldots,\theta_{p-1,p})^{\top}$.

Let $\boldsymbol{\theta}^0$ be the true value of $\boldsymbol{\theta}$, and let $\boldsymbol{Q}^0$ be the population Fisher information matrix of the model in criterion (3.9) at $\boldsymbol{\theta}^0$ (refer to Appendix I for details). Further, let

$$\boldsymbol{\mathcal{X}}_{j,j'} = (\mathbf{0}_n{}^{\top},\ldots,\mathbf{0}_n{}^{\top},\quad \underbrace{\boldsymbol{x}_{j'}{}^{\top},}_{j\text{-th block}}\quad \mathbf{0}_n{}^{\top},\ldots,\mathbf{0}_n{}^{\top},\quad \underbrace{\boldsymbol{x}_{j}{}^{\top},}_{j'\text{-th block}}\quad \mathbf{0}_n{}^{\top},\ldots,\mathbf{0}_n{}^{\top})^{\top}, \quad (3.10)$$

and let $\boldsymbol{\mathcal{X}} = (\boldsymbol{\mathcal{X}}_{1,2},\ldots,\boldsymbol{\mathcal{X}}_{j,j'},\ldots,\boldsymbol{\mathcal{X}}_{p-1,p})$. Let $\boldsymbol{\mathcal{X}}^{(i,j)}$ be the $[(j-1)n+i]$-th row of $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{X}}^{(i)} = (\boldsymbol{\mathcal{X}}^{(i,1)},\ldots,\boldsymbol{\mathcal{X}}^{(i,p)})^{\top}$, and let $\boldsymbol{U}^0 = E(\boldsymbol{\mathcal{X}}^{(i)\top}\boldsymbol{\mathcal{X}}^{(i)})$. In addition, let $S = \{(j,j') : \theta^0_{j,j'} \neq 0, 1 \le j < j' \le p\}$ be the index set of all nonzero components of $\boldsymbol{\theta}^0$, whose cardinality is denoted by q, and let S^c be the complement of S. Finally, for any matrix $\boldsymbol{W}$ and subsets of row and column indices $\mathcal{U}$ and $\mathcal{V}$, let $\boldsymbol{W}_{\mathcal{U},\mathcal{V}}$ be the matrix consisting of rows $\mathcal{U}$ and columns $\mathcal{V}$ in $\boldsymbol{W}$, and let $\Lambda_{\min}(\cdot)$ and $\Lambda_{\max}(\cdot)$ denote the smallest and largest eigenvalue of a matrix.

Our results rely on the following regularity conditions:

(A) Dependency: There exist positive constants $\tau_{\min}$ and $\tau_{\max}$ such that

$$\Lambda_{\min}(\boldsymbol{Q}^0_{S,S}) \ge \tau_{\min} \quad \text{and} \quad \Lambda_{\max}(\boldsymbol{U}^0_{S,S}) \le \tau_{\max}\ ; \quad (3.11)$$

(B) Incoherence: There exists a constant $\tau \in (0,1)$ such that

$$\|\boldsymbol{Q}^0_{S^c,S}(\boldsymbol{Q}^0_{S,S})^{-1}\|_{\infty} \le 1-\tau\ . \quad (3.12)$$

Similar conditions have been assumed by Meinshausen and Buhlmann (2006), Ravikumar et al. (2009) and Peng et al. (2009). The most closely related conditions for binary data are those of Ravikumar et al. (2009), but because they fit regressions separately, their conditions are on the $p \times p$ matrices corresponding to the individual regressions, whereas ours are on the $p(p-1)/2 \times p(p-1)/2$ matrices corresponding to all the parameters combined. These conditions can be interpreted as a bound on the amount of dependence (A), and a bound on influence non-neighbors can have on a given node (B). Under these conditions, we establish the following results:

Theorem III.1. *(Parameter estimation). Suppose conditions (A) and (B) hold and* $\widehat{\boldsymbol{\theta}}$ *is the maximizer of the criterion* (3.9)*. If the tuning parameter* $\lambda = C_\lambda \sqrt{(\log p)/n}$ *for some constant* $C_\lambda > 16(2-\tau)/\tau$ *and if* $n > (4/C)q^3 \log(p)$ *for some constant* $C < \tau_{min}^2 \tau^2 / \max\{288(1-\tau)^2, 72\}$*, then with probability tending to 1,*

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\|_2 \leq M\sqrt{\frac{q \log p}{n}} \ , \tag{3.13}$$

for some constant $M > (2C_\lambda/\tau_{min})[1 + \tau/(8-4\tau)]$*.*

Theorem III.2. *(Structure estimation). Under conditions of Theorem III.1, if we further assume* $\theta_{min}^0 = \min_{(j,j')\in S} |\theta_{j,j'}^0| \geq 2M\sqrt{q\log(p)/n}$*, then with probability tending to 1,*

$$\widehat{\theta}_{j,j'} \neq 0 \textit{ for all } (j,j') \in S \textit{ and } \widehat{\theta}_{j,j'} = 0 \textit{ for all } (j,j') \in S^c \ .$$

The proofs of Theorems III.1 and III.2 are given in Appendix I.

3.4 Application to the Senate Voting Record

The dataset was obtained from the website of the US Congress (`http://www.senate.gov`). It contains the voting records of the 100 senators of the 109th Congress (January 3, 2005 — January 3, 2007) on 645 bills, resolutions, motions, debates and roll call votes that the Senate deliberated and voted on. The votes are recorded as one for “yes” and zero for “no”. Missing values (missed votes) for each senator were imputed with the majority vote of that senator’s party on that particular bill; the missing votes for the Independent Senator Jeffords were imputed with the Democratic majority vote. The number of imputed votes is fairly small, less than 5% of the total and less than 3% of the total votes for 90% of the senators, and we do not expect this imputation to have a significant effect on the analysis. Finally, we excluded bills from the analysis if the ”yes/no” proportion fell outside the interval $[0.3, 0.7]$, since the Senate votes on many procedural and other uncontroversial motions that do not reflect the real political dynamics in the Senate. This resulted in a total of 387 observations (votes) on 100 variables (senators). We applied the joint neighborhood selection method to estimate the network structure and results are shown in Figure 3.1.

A richer structure than that dictated by the presence of two political parties emerges, with four distinct communities, two Republican and two Democratic. As expected, the two political parties are well separated, with many positive dependence links within their members (green solid lines) and negative links across parties (red dashed lines). The two communities on the left side of the plot can be broadly described as representing the cores of the two parties, although there is additional structure. For example, a number of the more liberal Democrats (Obama, Boxer, Kennedy, Bingaman, Stabenow, Kerry, Lautenberg, Sarbanes, Mikulski, Wyden, Leahy, Dorgan) have the strongest negative associations with the more conservative Republicans (Roberts, Sessions, Hutchison, Coburn, Burr, Shelby, Allen, Cornyn), mostly from Southern states (see also related analysis of earlier congresses in Clinton et al. (2004) and de Leeuw (2006)). Further, a number of positive associations are detected between some of the more centrist Democrats (Lieberman, Nelson, Baucus, Landrieu, Schumer, Clinton); a detailed inspection of the votes suggests that these are mostly due to their positions on issues of national security and the economy. Similarly, there

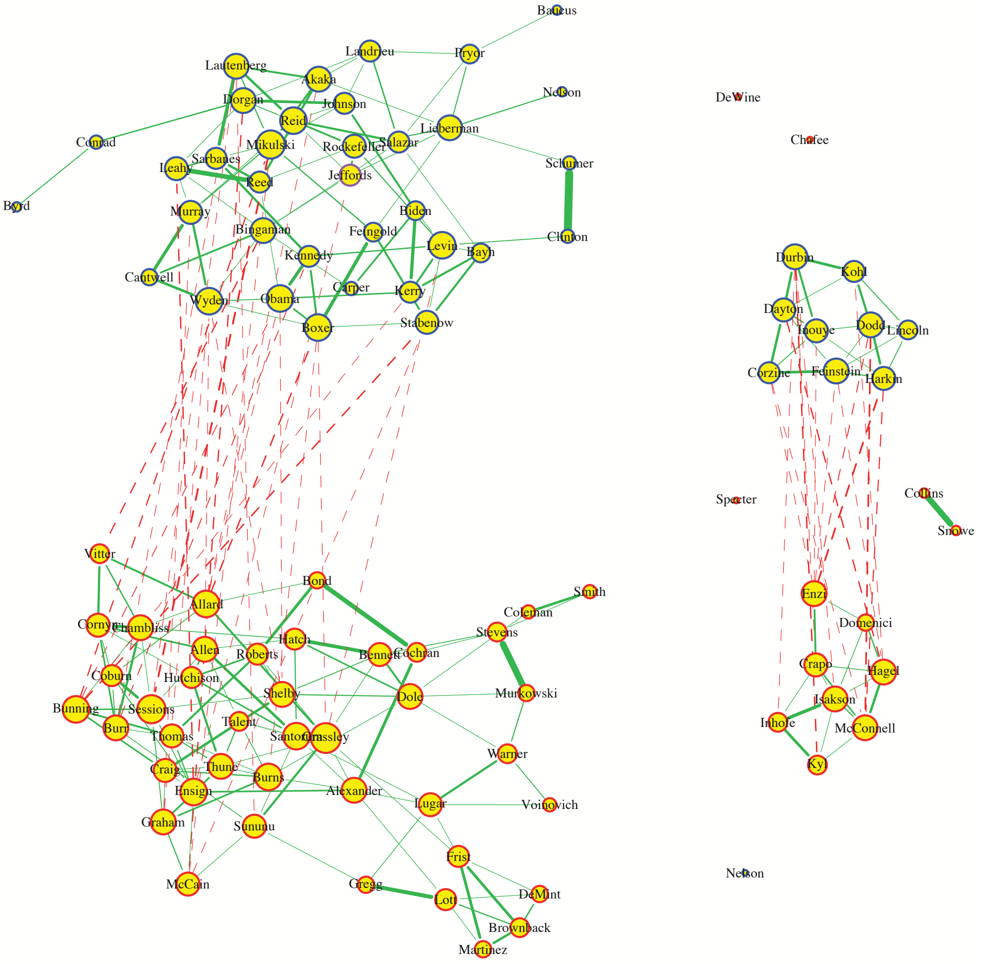


Figure 3.1: Voting dependencies between senators estimated by the joint neighborhood selection method. Each red (blue) circle represents a Republican (Democratic) senator, the circle size is proportional to the degree of the node. Senator Jeffords (the purple circle) is an independent senator. A solid green (dashed red) link represents a positive (negative) dependence between two senators. The width of each link is proportional to its associated $|\widehat{\theta}_{j,j'}|$. For clarity, all links with $|\widehat{\theta}_{j,j'}| \leq 0.1$ have the same width.

is a separate cluster of moderate Republicans (Grassley, Lugar, Alexander, Warner, Frist, Voinovich). A separate community of Republicans and Democrats emerges on the right side of the plot. An inspection of the votes suggests that they differ from the core members of their respective parties because of their voting record on several issues, including national security, confirmation votes on nominations, and certain regulatory and budget measures. Also of interest is the strong agreement between pairs of senators coming from the same state and party (Schumer-Clinton, Murray-Cantwell, Stevens-Murkowski, Hatch-Bennett, Collins-Snowe). Further, moderate Republicans DeWine, Chafee and Specter and the pro-life Democrat Nelson are represented as isolated nodes, thus confirming results of previous analysis by Clinton et al. (2004) and de Leeuw (2006) (albeit based on data from the 105th Congress). We also note that the Senate voting record from the 109th Congress was analyzed by Banerjee et al. (2008); however, the dataset they used turned out to have been contaminated with many votes from earlier Congresses starting from the 1990s, which led to a large number of missing votes for senators elected later. Since their imputation method was to impute “no” for all missing votes, the validity of their analysis is unclear and their results cannot be directly compared to ours. Overall, our analysis confirms known political patterns and provides new insights into the U.S. Senate's voting.

3.5 Extension to General Markov Networks

The joint neighborhood selection method can be extended to model general Markov networks consisting of categorical variables. Let $(x_{i,1}, \dots, x_{i,p})$ be the i-th observation, where $x_{i,j}$, $1 \leq j \leq p$, takes values in the discrete set $\{1, 2, \dots, D\}$ for some positive integer D. Denote by $z_{i,j}^{(1)}, \dots, z_{i,j}^{(D-1)}$ the dummy variables associated with $x_{i,j}$, i.e., $z_{i,j}^{(d)} = \mathrm{I}(x_{i,j} = d)$, $1 \leq d \leq D-1$, where $\mathrm{I}(\cdot)$ denotes the indicator function. Notice that we omit $z_{i,j}^{(D)}$ because it is redundant given the constraint $\sum_{d=1}^{D} z_{i,j} = 1$.

The criterion of joint neighborhood selection can be modified as follows:

$$\begin{aligned}\max_{\{\boldsymbol{\theta}_j^*:1\le j\le p\}\bigcup\{\boldsymbol{\theta}_{j,j'}^*:1\le j<j'\le p\}} \quad & \sum_{j=1}^{p}\sum_{i=1}^{n}\Big[\sum_{d=1}^{D-1} z_{i,j}^{(d)}\Big(\theta_j^{(d)}+\sum_{k\neq j}\sum_{d'=1}^{D-1}\theta_{j,k}^{(d,d')}z_{i,k}^{(d')}\Big) \\ & - \log\Big\{\sum_{d=1}^{D-1}\exp\Big(\theta_j^{(d)}+\sum_{k\neq j}\sum_{d'=1}^{D-1}\theta_{j,k}^{(d,d')}z_{i,k}^{(d')}\Big)\Big\}\Big] \\ & - \lambda\sum_{j<j'}\sqrt{\sum_{d=1}^{D-1}\sum_{d'=1}^{D-1}(\theta_{j,j'}^{(d,d')})^2} \\ \text{subject to} \quad & \theta_{j,j'}^{(d,d')}=\theta_{j',j}^{(d,d')},\ 1\le j<j'\le p, 1\le d,d'\le D-1\end{aligned} \tag{3.14}$$

In (3.14), $\theta_j^{(d)}$ corresponds to the main effect of variable j in class d and $\theta_{j,j'}^{(d,d')}$ to the interaction effect between variable j in class d and variable j' in class d'. Further, $\boldsymbol{\theta}_j^* = \{\theta_j^{(d)} : 1 \le d \le D-1\}$ collects all main effects associated with variable j and $\boldsymbol{\theta}_{j,j'}^* = \{\theta_{j,j'}^{(d,d')} : 1 \le d, d' \le D-1\}$ collects all interaction effects associated with variables j and j'. Here, we remove the edge between nodes j and j' only if *all* the elements in $\boldsymbol{\theta}_{j,j'}^*$ are zero. To achieve this, we use the group penalty proposed by Yuan and Lin (2007), where all elements in $\boldsymbol{\theta}_{j,j'}^*$ are regarded as a group and simultaneously estimated as zeros or nonzeros. Criterion (3.14) can be estimated by a modified LQA-shooting algorithm, in which the inner loop is replaced by a modified shooting algorithm for group lasso (Friedman et al., 2007).

CHAPTER IV

Estimating Heterogeneous Graphical Models for Discrete Data with an Application to Roll Call Voting

4.1 Introduction

In this chapter, we focus on the case of a Markov network for binary random variables, which generalizes easily to categorical data. An interesting application of such networks deals with the analysis of *roll call data* for the United States Congress. Such data have obviously received a lot of attention amongst political scientists (see for example the books by (Enelow and Hinich, 1984; Matthews and J.A., 1975; Morton, 1999; Poole and Rosenthal, 1997), but has also been an application area for statistical techniques, including principal component analysis (de Leeuw, 2006), multidimensional scaling (Diaconis et al., 2008), Bayesian models (Clinton et al., 2004) and Gaussian graphical models (Banerjee et al., 2008). However, all such techniques have focused on treating the votes as homogeneous, assuming all the votes represent the same underlying relationship among senators/congressmen. However, it is well known that there are certain subgroups of politicians whose voting behavior depends on the issue, and who form different alliances when voting, for example, on national security and health care. Therefore, treating votes as heterogeneous is more accurate, and can provide further insight into the voting behavior of different groups of senators on different issues. In our application, we focus on voting records on three types of bills: defense and national security, environment and energy, and healthcare issues. Voting on the latter category is typically more partisan than voting on defense and national security, and thus we expect to see different connections in different categories.

To accomplish the analysis allowing for heterogeneity, we develop a framework for fitting different Markov models for each category that are nevertheless *linked*, sharing nodes and having some common edges across all categories, while other edges are uniquely associated with a particular category. Asymptotic properties of the proposed estimator are also established. Note that for the Gaussian case, this problem was considered by Guo et al. (2011), who proposed a joint likelihood based estimation method that borrowed strength across categories.

The remainder of the chapter is organized as follows. Section 4.2 introduces the Markov network and addresses algorithmic issues, while Section 4.5 presents asymptotic results. Section 4.3 illustrates the performance of the joint estimation method using simulated data, and the US Senate's voting record is analyzed in Section 4.4. Some concluding remarks are drawn in Section 4.6.

4.2 Model and Estimation Algorithm

In this section, we present the Markov model for heterogeneous data, focusing on the special case of binary variables (also known as the Ising model). The extension to general categorical variables is briefly discussed in Section 4.6. We start by discussing estimation of *separate* models for each category and then develop a model for joint estimation.

4.2.1 Problem Setup and Separate Estimation Method

Suppose that data have been collected on p *binary* variables for K categories, with n_k observations for the k-th category. Let $\boldsymbol{x}_i^{(k)} = (x_{i,1}^{(k)}, \ldots, x_{i,p}^{(k)})$ denote a p-dimensional row vector containing the data for the i-th observation in the k-th category and assume that it is independent observation from an exponential family with density function:

$$\mathrm{f}_k(X_1, \ldots, X_p) = \frac{1}{\mathrm{Z}(\boldsymbol{\Theta}^{(k)})} \exp\Big(\sum_{j=1}^{p} \theta_{j,j}^{(k)} X_j + \sum_{1 \le j < j' \le p} \theta_{j,j'}^{(k)} X_j X_{j'} \Big). \tag{4.1}$$

The partition function $\mathrm{Z}(\boldsymbol{\Theta}^{(k)}) = \sum_{X_j \in \{0,1\}, 1 \le j \le p} \exp(\theta_{j,j}^{(k)} X_j + \sum_{1 \le j < j' \le p} \theta_{j,j'}^{(k)} X_j X_{j'})$ ensures that the density function in (4.1) is a proper one, integrating to one. The parameters $\theta_{j,j}^{(k)}$, $1 \le j \le p$ correspond to the main effect for variable X_j in the k-th category, while $\theta_{j,j'}^{(k)}$, $1 \le j < j' \le p$ to the interaction effect between variables X_j and $X_{j'}$. The underlying network associated with the k-th category is determined by the symmetric matrix $\boldsymbol{\Theta}^{(k)} = (\theta_{j,j'}^{(k)})_{p \times p}$. Specifically, if $\theta_{j,j'}^{(k)} = 0$, then X_j and $X_{j'}$ are conditionally independent in the k-th category given all the remaining variables and hence their corresponding nodes are *not* connected. For each category, criterion (4.1) is referred to as the Markov network in the machine learning literature, and as the log-linear model in the statistics literature, where $\theta_{j,j'}^{(k)}$ is also interpreted as the conditional log-odds-ratio between X_j and $X_{j'}$ given the other variables. Although general Markov networks allow higher order interactions (3-way, 4-way, etc), Ravikumar et al. (2010) pointed out that one can consider only the pairwise interaction effects without loss of generality, since higher order interactions can be converted to pairwise ones by introducing additional variables (Wainwright and Jordan, 2008).

The simplest way to deal with such heterogenous data is to estimate K separate Markov models. Specifically, if one further assumes sparsity for the k-th category, the structure of the underlying graph can be estimated by regularizing the log-likelihood using an ℓ_1 penalty:

$$\max_{\mathbf{\Theta}^{(k)}} \frac{1}{n_k}\sum_{i=1}^{n_k}\Big\{\sum_{j=1}^{p}\theta_{j,j}^{(k)}x_{i,j}^{(k)}+\sum_{1\le j<j'\le p}\theta_{j,j'}^{(k)}x_{i,j}^{(k)}x_{i,j'}^{(k)}\Big\}-\log \mathrm{Z}(\mathbf{\Theta}^{(k)})-\lambda\sum_{1\le j<j'\le p}|\theta_{j,j'}^{(k)}|. \quad (4.2)$$

The ℓ_1 penalty shrinks some interaction effects $\theta_{j,j'}^{(k)}$, $1 \le j < j' \le p$, to zero and λ controls the degree of sparsity. However, estimating (4.2) directly is computationally infeasible due to the nature of the partition function. To overcome this difficulty, we adopt a pseudo-likelihood estimation method proposed in Guo et al. (2010), based on:

$$\begin{aligned}\max_{\mathbf{\Theta}^{(k)}} \quad & \frac{1}{n_k}\sum_{i=1}^{n_k}\sum_{j=1}^{p}\Big[x_{i,j}^{(k)}\Big(\theta_{j,j}^{(k)}+\sum_{j'\neq j}\theta_{j,j'}^{(k)}x_{i,j'}^{(k)}\Big)\\ & -\log\Big\{1+\exp\Big(\theta_{j,j}^{(k)}+\sum_{j'\neq j}\theta_{j,j'}^{(k)}x_{i,j'}^{(k)}\Big)\Big\}\Big]\\ & -\lambda\sum_{1\le j<j'\le p}|\theta_{j,j'}^{(k)}|\ , \end{aligned} \quad (4.3)$$

where $\mathbf{\Theta}^{(k)}$ is restricted to be symmetric. Criterion 4.3 can be efficiently solved using a modified coordinate descent algorithm introduced in Guo et al. (2010), or the algorithm of Hoefling and Tibshirani (2009).

4.2.2 Joint Estimation of Heterogeneous Networks

We start by reparameterizing each $\theta_{j,j'}^{(k)}$ as

$$\theta_{j,j'}^{(k)} = \phi_{j,j'}\gamma_{j,j'}^{(k)},\ 1 \le j \neq j' \le p; 1 \le k \le K. \quad (4.4)$$

To avoid sign ambiguities between $\phi_{j,j'}$ and $\gamma_{j,j'}^{(k)}$, we restrict $\phi_{j,j'} \ge 0$, $1 \le j < j' \le p$. To preserve the symmetry of $\mathbf{\Theta}^{(k)}$, we also require $\phi_{j,j'} = \phi_{j',j}$ and $\gamma_{j,j'}^{(k)} = \gamma_{j',j}^{(k)}$, $1 \le j < j' \le p$ and $1 \le k \le K$. Moreover, for identifiability reasons, we restrict the diagonal elements $\phi_{j,j} = 1$ and $\gamma_{j,j}^{(k)} = \omega_{j,j}^{(k)}$. Note that $\phi_{j,j'}$ is a common factor across all K categories that controls the occurrence of common links shared across categories, while $\gamma_{j,j'}^{(k)}$ is an individual factor specific to the k-th category. The proposed joint estimation method considers maximizing the following penalized criterion:

$$\begin{aligned}\max_{\{\boldsymbol{\Phi}^{(k)},\boldsymbol{\Gamma}^{(k)}\}_{k=1}^{K}} \quad & \sum_{k=1}^{K}\frac{1}{n_k}\sum_{i=1}^{n_k}\sum_{j=1}^{p}\Big[x_{i,j}^{(k)}\Big(\theta_{j,j}^{(k)}+\sum_{j'\neq j}\theta_{j,j'}^{(k)}x_{i,j'}^{(k)}\Big) \\ & -\log\Big\{1+\exp\Big(\theta_{j,j}^{(k)}+\sum_{j'\neq j}\theta_{j,j'}^{(k)}x_{i,j'}^{(k)}\Big)\Big\}\Big] \\ & -\eta_1\sum_{1\leq j<j'\leq p}\phi_{j,j'}-\eta_2\sum_{1\leq j<j'\leq p}\sum_{k=1}^{K}|\gamma_{j,j'}^{(k)}| \ , \end{aligned} \tag{4.5}$$

where $\boldsymbol{\Phi}^{(k)} = (\phi_{j,j'})_{p\times p}$ and $\boldsymbol{\Gamma}^{(k)} = (\gamma_{j,j'}^{(k)})_{p\times p}$, with η_1 a tuning parameter controlling the sparsity of the common structure across the K networks. Specifically, if $\phi_{j,j'}$ is shrunk to zero, all $\theta_{j,j'}^{(1)},\ldots,\theta_{j,j'}^{(K)}$ are also zero, and hence there is no link between nodes j and j' in any of the K graphs. Similarly, η_2 is a tuning parameter controlling the sparsity of links for individual categories. Due to the nature of the ℓ_1 penalty, some of $\gamma_{j,j'}^{(k)}$'s will be shrunk to zero, resulting in a collection of graphs with individual differences. Note that this two-level penalty was originally proposed in Zhou and Zhu (2010) for group variable selection in linear regression.

To simplify estimation, we convert the criterion (4.5) to an equivalent criterion with only one tuning parameter:

$$\begin{aligned}\max_{\{\boldsymbol{\Theta}^{(k)}\}_{k=1}^{K}} \quad & \sum_{k=1}^{K}\frac{1}{n_k}\sum_{i=1}^{n_k}\sum_{j=1}^{p}\Big[x_{i,j}^{(k)}\Big(\theta_{j,j}^{(k)}+\sum_{j'\neq j}\theta_{j,j'}^{(k)}x_{i,j'}^{(k)}\Big) \\ & -\log\Big\{1+\exp\Big(\theta_{j,j}^{(k)}+\sum_{j'\neq j}\theta_{j,j'}^{(k)}x_{i,j'}^{(k)}\Big)\Big\}\Big] \\ & -\lambda\sum_{1\leq j<j'\leq p}\sqrt{\sum_{k=1}^{K}|\theta_{j,j'}^{(k)}|} \ , \end{aligned} \tag{4.6}$$

where $\lambda = 2\sqrt{\eta_1\eta_2}$. The equivalence between (4.5) and (4.6) can be formalized as follows ($\boldsymbol{A}\cdot\boldsymbol{B}$ denotes the Schur-Hadamard element-wise product of two matrices);

Proposition 1. *Let* $\{\widehat{\boldsymbol{\Theta}}^{(k)}\}_{k=1}^{K}$ *be a local minimizer of* (4.6)*. Then there exists a local minimizer of* (4.5)*,* $(\widehat{\boldsymbol{\Phi}}, \{\widehat{\boldsymbol{\Gamma}}^{(k)}\}_{k=1}^{K})$*, such that* $\widehat{\boldsymbol{\Theta}}^{(k)} = \widehat{\boldsymbol{\Phi}} \cdot \widehat{\boldsymbol{\Gamma}}^{(k)}$*, for all* $1 \leq k \leq K$*. On the other hand, if* $(\widehat{\boldsymbol{\Phi}}, \{\widehat{\boldsymbol{\Gamma}}^{(k)}\}_{k=1}^{K})$ *is a local minimizer of* (4.5)*, then there also exists a local minimizer of* (4.6)*,* $\{\widehat{\boldsymbol{\Theta}}^{(k)}\}_{k=1}^{K}$*, such that* $\widehat{\boldsymbol{\Theta}}^{(k)} = \widehat{\boldsymbol{\Phi}} \cdot \widehat{\boldsymbol{\Gamma}}^{(k)}$*, for all* $1 \leq k \leq K$*.*

The proof of this proposition is similar to the proofs of Lemma 1 and Theorem 1 in Zhou and Zhu (2010) and is omitted here.

4.2.3 Algorithm and Model Selection

Criterion (4.6) leads to an efficient estimation algorithm based on the local linear approximation. Specifically, letting $(\theta_{j,j'}^{(k)})^{[t]}$ denote the estimates from the t-th iteration, we approximate $\sqrt{\sum_{k=1}^{K} |\theta_{j,j'}^{(k)}|} \approx \sum_{k=1}^{K} |\theta_{j,j'}^{(k)}| / \sqrt{\sum_{k=1}^{K} |(\theta_{j,j'}^{(k)})^{[t]}|}$, when $\theta_{j,j'}^{(k)} \approx (\theta_{j,j'}^{(k)})^{[t]}$. Thus, at the $(t+1)$-th iteration, problem (4.6) is decomposed into K individual optimization problems:

$$\begin{aligned}
\max_{\boldsymbol{\Theta}^{(k)}} \quad & \frac{1}{n_k} \sum_{i=1}^{n_k} \sum_{j=1}^{p} \Big[x_{i,j}^{(k)} \Big(\theta_{j,j}^{(k)} + \sum_{j' \neq j} \theta_{j,j'}^{(k)} x_{i,j'}^{(k)} \Big) \\
& - \log \Big\{ 1 + \exp \Big(\theta_{j,j}^{(k)} + \sum_{j' \neq j} \theta_{j,j'}^{(k)} x_{i,j'}^{(k)} \Big) \Big\} \Big] \\
& - \lambda \sum_{1 \leq j < j' \leq p} \Big(\sum_{k=1}^{K} |(\theta_{j,j'}^{(k)})^{[t]}| \Big)^{-1/2} |\theta_{j,j'}^{(k)}| \, . \qquad (4.7)
\end{aligned}$$

Note that criterion (4.7) is a variant of criterion (4.3) with a weighted ℓ_1 penalty and hence can be solved by the JOSE algorithm in Guo et al. (2010). For numerical stability, we threshold $\sqrt{\sum_{k=1}^{K} |(\theta_{j,j'}^{(k)})^{[t]}|}$ at 10^{-10}. The algorithm is summarized as follows:

Step 1. Initialize $\widehat{\theta}_{j,j'}^{(k)}$'s $(1 \leq j, j' \leq p; 1 \leq k \leq K)$ using the estimates from the separate estimation method;

Step 2. For each $1 \leq k \leq K$, update $\widehat{\theta}_{j,j'}^{(k)}$'s by solving (4.7) using the JOSE algorithm in Guo et al. (2010);

Step 3. Repeat Step 2 until convergence.

The tuning parameter λ in (4.6) controls the sparsity of the resulting estimator and it can be selected using cross-validation. Specifically, for each $1 \leq k \leq K$, we randomly split the data in the k-th category into D subsets with similar sizes and denote the index set of the observations in the d-th subset as $\mathcal{T}_d^{(k)}$, $1 \leq d \leq D$. Then λ is selected by maximizing

$$\begin{aligned}\frac{1}{D}\sum_{d=1}^{D}\sum_{k=1}^{K}\frac{1}{|\mathcal{T}_d^{(k)}|}\sum_{i\in\mathcal{T}_d^{(k)}}\sum_{j=1}^{p} \quad & x_{i,j}^{(k)}\Big\{(\widehat{\theta}_{j,j}^{(k)})^{[-d]}(\lambda)+\sum_{j'\neq j}(\widehat{\theta}_{j,j'}^{(k)})^{[-d]}(\lambda)x_{i,j'}^{(k)}\Big\} \\ - \quad & \log\Big[1+\exp\Big\{(\widehat{\theta}_{j,j}^{(k)})^{[-d]}(\lambda)+\sum_{j'\neq j}(\widehat{\theta}_{j,j'}^{(k)})^{[-d]}(\lambda)x_{i,j'}^{(k)}\Big\}\Big]\end{aligned} \tag{4.8}$$

where $|\mathcal{T}_d^{(k)}|$ is the cardinality of $\mathcal{T}_d^{(k)}$ and $(\widehat{\theta}_{j,j'}^{(k)})^{[-d]}(\lambda)$ is the joint estimate of $\theta_{j,j'}^{(k)}$ based on all observations except those in $\mathcal{T}_d^{(1)}\bigcup\ldots\bigcup\mathcal{T}_d^{(K)}$, as well as the tuning parameter λ.

4.3 Simulation Study

In this section, we evaluate the performance of the joint estimation method on three synthetic examples, each with $p = 50$ variables and $K = 3$ categories. The network structure in each example is composed of two parts: the common structure across all categories and the individual structure specific to a category. The common structures in these examples are a chain graph, a nearest neighbor graph and a scale-free graph. These graphs are generated as follows:

Example 1: Chain Graph. A chain graph is generated by connecting nodes 1 to p in increasing order, as shown in Figure 4.1 (A1).

Example 2: Nearest Neighbor Graph. The data generating mechanism of the nearest neighbor graph is adapted from Li and Gui (2006). Specifically, we generate p points randomly on a unit square, calculate all $p(p-1)/2$ pairwise distances, and find three nearest neighbors of each point in terms of these distances. The nearest neighbor network is obtained by linking any two points that are nearest neighbors of each other. Figure 4.1 (B1) illustrates a nearest-neighbor graph.

Example 3: Scale-free Graph. A scale-free graph has a power-law degree distribution and can be simulated by the Barabasi-Albert algorithm (Barabasi and Albert, 1999). A realization of a scale-free network is depicted in Figure 4.1 (C1).

In each example, the network for the k-th category ($k = 1, \ldots, K$) is created by randomly adding links to the common structure. The individual links in different categories are disjoint and have the same degree of sparsity, measured by ρ, the ratio of the number of individual links to the number of common links. In particular, $\rho = 0$ corresponds to identical networks for all three categories. In the simulation study, we consider ρ=0, 1/4 and 1, gradually increasing the proportion of individual links (Figure 4.1). Given the graphs, the symmetric parameter matrix $\mathbf{\Theta}^{(k)}$ is generated as follows. Each $\theta^{(k)}_{j,j'} = \theta^{(k)}_{j',j}$ corresponding to a link between nodes j and j' is uniformly drawn from $[-1, -0.5] \cup [0.5, 1]$, whereas all other elements are set to zero. Then we generate the data using Gibbs sampling. Specifically, suppose the i-th iteration sample has been drawn and is denoted as $(x^{(k)}_1)^{[t]}, \ldots, (x^{(k)}_p)^{[t]}$; then, in the $(t+1)$-th iteration, we draw $(x^{(k)}_j)^{[t+1]}$, $1 \leq j \leq p$, from the Bernoulli distribution:

$$(x^{(k)}_j)^{[t+1]} \sim \text{Bernoulli}\Big(\frac{\exp(\theta^{(k)}_{j,j} + \sum_{j' \neq j} \theta^{(k)}_{j,j'} (x^{(k)}_{j'})^{[t]})}{1 + \exp(\theta^{(k)}_{j,j} + \sum_{j' \neq j} \theta^{(k)}_{j,j'} (x^{(k)}_{j'})^{[t]})}\Big). \tag{4.9}$$

To ensure that the simulated observations are close to i.i.d. samples from the target distribution, the first 1,000,000 rounds are discarded (burn-in) and the data are col-

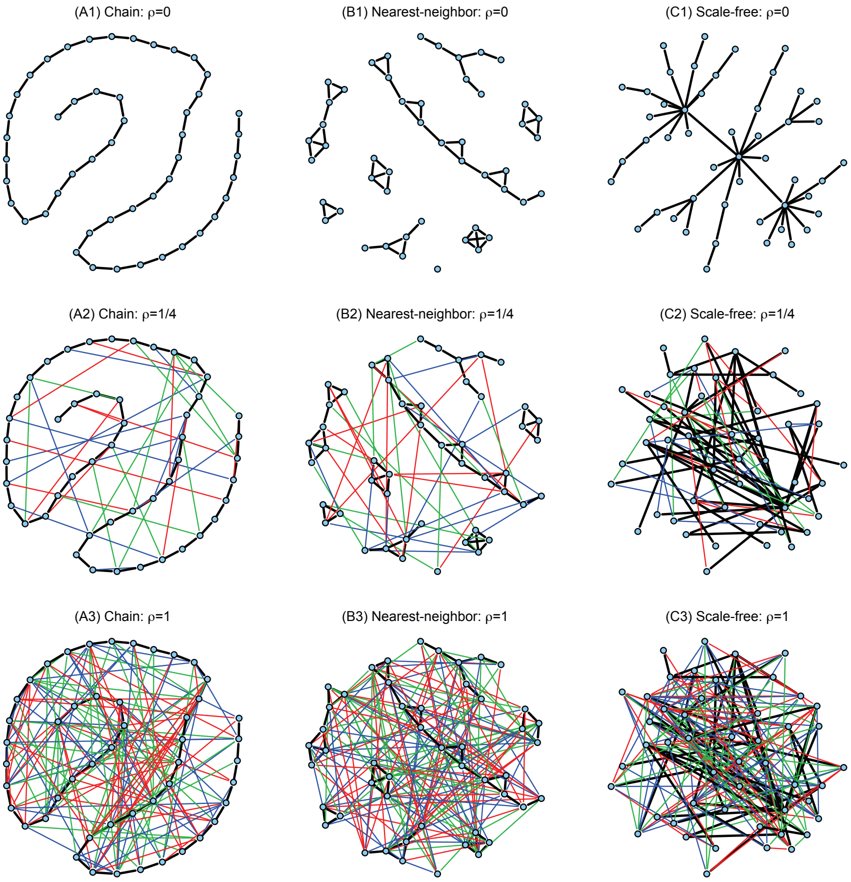


Figure 4.1: The networks used in three simulated examples. The black lines represent the common structure, whereas the red, blue and green lines represent the individual links in the three categories. ρ is the ratio of the number of individual links to the number of common links.

lected every 100 iterations from the sampler. In the simulation study, we consider a balanced scenario and an unbalanced scenario. The former consists of $n_k = 200$ observations in each category, whereas the latter has three unbalanced categories with sample sizes $n_1 = 150$, $n_2 = 300$ and $n_3 = 450$.

We compared the structure estimation results of the joint estimation method and the separate estimation method using ROC curves, which dynamically characterize the sensitivity (proportion of correctly identified links) and the specificity (proportion of correctly excluded links) by varying the tuning parameter λ. Figure 4.2 shows the ROC curves averaged over 50 replications from the three examples in the balanced scenario. It can be seen that the curves estimated by the joint estimation method dominate those of the separate estimation method when the proportion of individual links is low. As ρ increases, the structures become more and more different, and the joint and separate methods move closer together. This is expected, since the joint estimation method is designed to take advantage of common structure. The results in the unbalanced scenario exhibit a similar pattern (Figure 4.3).

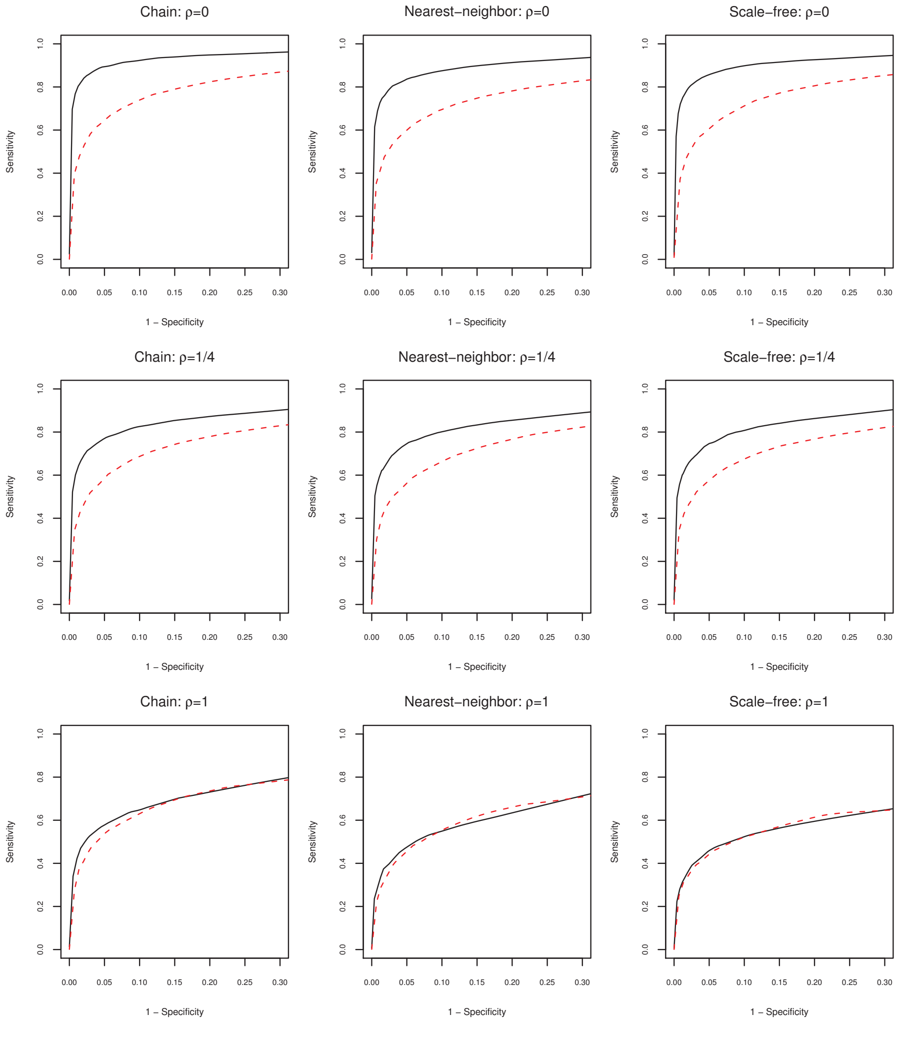


Figure 4.2: Results for the balanced scenario ($n_1 = n_2 = n_3 = 200$). The ROC curves are averaged over 50 replications. ρ is the ratio between the number of individual links and the number of common links.

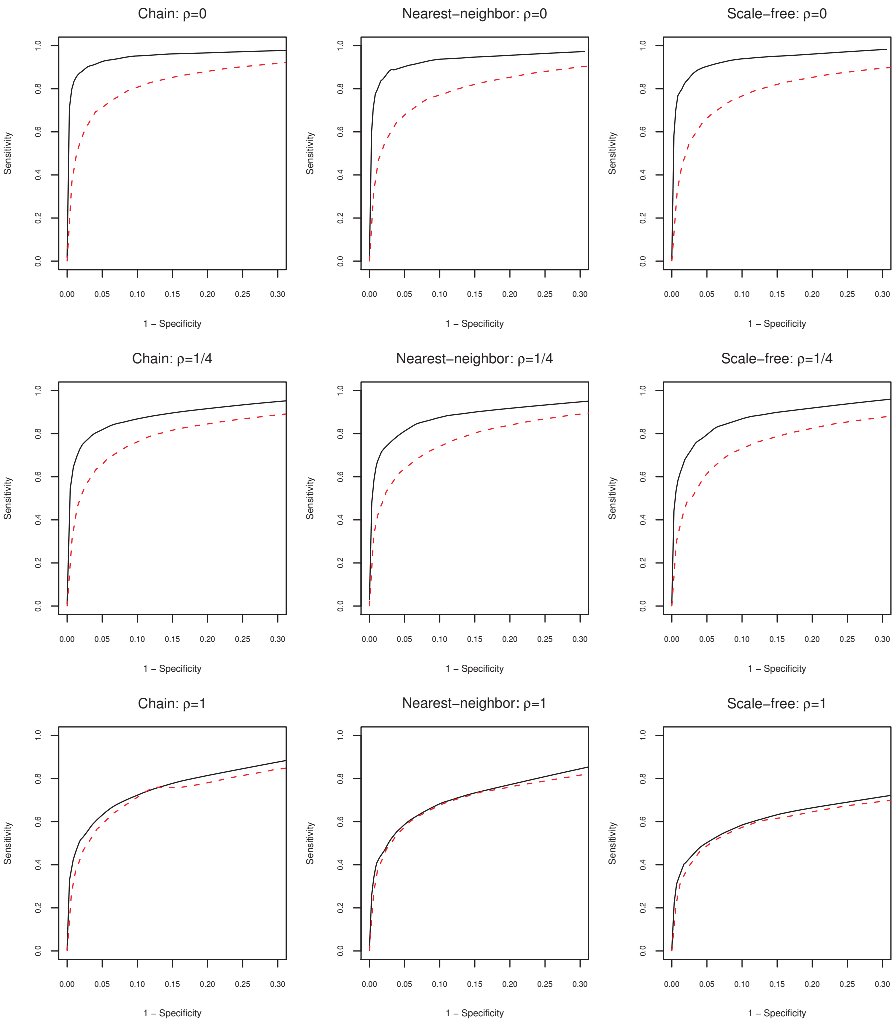


Figure 4.3: Results for the unbalanced scenario ($n_1 = 150$, $n_2 = 300$, $n_3 = 450$). The ROC curves are averaged over 50 replications. ρ is the ratio between the number of individual links and the number of common links.

4.4 Analysis of the U.S. Senate voting records

We applied the proposed joint estimation method to the voting records of the U.S. Senate from the 109th Congress covering the period 2005-2006. The data were obtained directly from the Senate's website (www.senate.gov). The variables correspond to the 100 senators, and the observations to the 645 votes that the Senate deliberated and voted on during that period, which include bills, resolutions, motions, debates and roll call votes. The votes are recorded as "yes" (encoded as "1") and “no” (encoded as “0”). Missing observations were replaced with the majority vote of the senator’s party on that particular vote. The bills with a “yes/no” proportion greater than 90% or less than 10% were excluded from the analysis. Three categories of votes were extracted from bills, resolutions and motions: 1) defense and security issues (133); 2) environment and energy issues (34); 3) health and medical care issues (46). The tuning parameter for the proposed method was selected through cross-validation. Following Li and Gui (2006), we used a bootstrap procedure with the proposed estimator to evaluate the confidence of the estimated edges. We only keep the robust edges in the estimated networks and remove those with occurrence frequencies less than some cut-off value in the bootstrap procedure.

The network representation, depicting both the common and the individual structures with a cut-off for inclusion of 0.5, is given in Figure 4.4. The common network estimated by the joint estimation method is shown in the top left panel of the Figure. As expected, members of the two political parties are clearly separated. There are many more associations between Democratic senators than Republican ones and this pattern holds for both the common and individual structures. One possible explanation may be that the Democrats were in the opposition, thus voting more like a block. Further, the Independent senator Jeffords is very “close” to the Democratic caucus, while the moderate Republicans Collins, Snow, Chafee and Specter (who switched to the Democratic party in early 2009) are closely positioned together, thus confirming results of previous analyses by Clinton et al. (2004) and de Leeuw (2006) (albeit based on data from the 105th Congress). Other interesting patterns emerging from the analysis are that the more moderate members of two parties are located closer to the center of their respective “clouds”’ (e.g. Warner, Voinovich, Smith on the Republican side and Levin, Reid, Mikulski, Rockefeller on the Democratic side), the close ties of the liberal Democrats Kennedy, Boxer and Nelson (Florida), the close voting records of senators from the same state Murkowski and Stevens (Alaska) and

Cantwell and Murray (Washington).

Examining the individual networks for the three categories shown in the remaining panels of Figure 4.4 we note that a lot of additional positive associations amongst Democrats emerge, primarily for defense and healthcare issue, thus indicating a stronger ideological cohesion for these two categories. Some stable negative associations emerge for the environment and healthcare categories.

Commenting on some selected patterns for individual senators, a strong dependence can be observed between Biden and Kerry on environmental and health care issues, but less so on defense, whereas Schumer and Clinton (Democratic senators from New York) are in strong agreement on defense, but less so on the other two categories. Interestingly, in general there is a lot of positive dependence among the Democratic senators on defense and health care issues (demonstrated by the thickness of the links), and very little on the environment, while for Republican senators the strengths of the associations are about the same for all three categories; an exception is the strong association of Murkwoski and Stevens on environmental and energy issues, given that they come from the oil rich state of Alaska. The overall weaker associations on defense votes can be partially explained by the fact that a number of them reflect some financial aspect (budget approval, appropriation, etc). In general, the model captures the basic common structure, as well as meaningful differences across the various categories.

4.5 Asymptotic Properties

In this section, we study the asymptotic properties of the proposed joint estimation method. Since the structure of the underlying network only depends on the interaction effects, we focus on a variant of the model without main effects. Specifically, for each

$k = 1, \ldots, K$,

$$\max_{\{\mathbf{\Theta}^{(k)}\}_{k=1}^{K}} \quad \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i=1}^{n_k} \sum_{j=1}^{p} \Big[x_{i,j}^{(k)} \Big(\sum_{j' \neq j} \theta_{j,j'}^{(k)} x_{i,j'}^{(k)} \Big) - \log \Big\{ 1 + \exp \Big(\sum_{j' \neq j} \theta_{j,j'}^{(k)} x_{i,j'}^{(k)} \Big) \Big\} \Big]$$
$$-\lambda \sum_{1 \leq j < j' \leq p} \sqrt{\sum_{k=1}^{K} |\theta_{j,j'}^{(k)}|} \ . \qquad (4.10)$$

We will show that the estimator in criterion (4.10) is consistent in terms of both parameter estimation and model selection, when p and n go to infinity and the tuning parameter λ goes to zero at some appropriate rate.

Before stating the main results, we introduce necessary notation and regularity conditions. For each $k = 1, \ldots, K$, denote $\boldsymbol{\theta}^{(k)} = (\theta_{1,2}^{(k)}, \ldots, \theta_{j,j'}^{(k)}, \ldots, \theta_{p-1,p}^{(k)})$ as a $p(p-1)/2$-dimensional vector, recording all upper triangular elements in $\mathbf{\Theta}^{(k)}$. Let $\overline{\boldsymbol{\theta}}^{(k)}$ be the true value of $\boldsymbol{\theta}^{(k)}$. Let $\overline{\boldsymbol{Q}}^{(k)}$ be the population Fisher information matrix of the model in criterion (4.10) (see the Appendix for a precise definition) and let $\mathcal{X}_{(i)}^{(k)}$ be a matrix with p rows and $p(p-1)/2$ columns, whose (j, j')-th column is composed of zeros except the j-th (j'-th) component being $x_{i,j'}$ ($x_{i,j}$). In addition, we define $\overline{\boldsymbol{U}}^{(k)} = E[\boldsymbol{\mathcal{X}}_{(i)}^{(k)\intercal} \boldsymbol{\mathcal{X}}_{(i)}^{(k)}]$. To index the zero and nonzero elements, let $S_k = \{(j, j') : \theta_{j,j'}^{(k)} \neq 0, 1 \leq j < j' \leq p\}$ and $S_k^c = \{(j, j') : \theta_{j,j'}^{(k)} = 0, 1 \leq j < j' \leq p\}$, and let $S_\cap = \bigcap_{k=1}^{K} S_k$, $S_\cup = \bigcup_{k=1}^{K} S_k$. The cardinalities of S_k and $S_\cup$ are denoted by q_k and q, respectively. For any matrix $\boldsymbol{W}$ and subsets of row and column indices $\mathcal{U}$ and $\mathcal{V}$, let $\boldsymbol{W}_{\mathcal{U},\mathcal{V}}$ be the matrix consisting of rows $\mathcal{U}$ and columns $\mathcal{V}$ in $\boldsymbol{W}$. Finally, let $\Lambda_{\min}(\cdot)$ and $\Lambda_{\max}(\cdot)$ denote the smallest and largest eigenvalue of a matrix, respectively.

The asymptotic properties of the joint estimation method rely on the following regularity conditions:

(A) Nonzero elements bounds: There exist positive constants $\gamma_{\min}$ and $\gamma_{\max}$ such that

(i) $\min_{1 \leq k \leq K} \min_{(j,j') \in S_k} |\overline{\theta}_{j,j'}^{(k)}| \geq \gamma_{\min}$;

(ii) $\max_{1\le k\le K}\max_{(j,j')\in S_k\setminus S_\cap}|\overline{\theta}^{(k)}_{j,j'}|\le\gamma_{\max}$.

(B) Dependency: There exist positive constants $\tau_{\min}$ and $\tau_{\max}$ such that for any $k=1,\dots,K$,

$$\Lambda_{\min}(\overline{\boldsymbol{Q}}^{(k)}_{S_k,S_k})\ge\tau_{\min}\quad\text{and}\quad\Lambda_{\max}(\overline{\boldsymbol{U}}^{(k)}_{S_k,S_k})\le\tau_{\max}\ . \tag{4.11}$$

(C) Incoherence: There exists a constant $\tau\in(1-\sqrt{\gamma_{\min}/4\gamma_{\max}},1)$ such that for any $k=1,\dots,K$,

$$\|\overline{\boldsymbol{Q}}^{(k)}_{S_k^c,S_k}(\overline{\boldsymbol{Q}}^{(k)}_{S_k,S_k})^{-1}\|_\infty\le 1-\tau\ . \tag{4.12}$$

Condition (A) enforces a lower bound on the magnitudes of all nonzero elements, as well as an upper bound on the magnitudes of those nonzero elements associated with individual links. Conditions (B) and (C) bound the amount of dependence and the influence that the non-neighbors can have on a given node, respectively. Conditions similar to (B) and (C) were also assumed by Meinshausen and Buhlmann (2006), Ravikumar et al. (2010), Peng et al. (2009) and Guo et al. (2010). Our conditions are most closely related to those of Guo et al. (2010), but here they are extended to the heterogenous data setting.

Theorem IV.1. *(Parameter estimation). Suppose all regularity conditions hold. If the tuning parameter* $\lambda=C_\lambda\sqrt{(\log p)/n}$ *for some constant* $C_\lambda>(8-4\tau)\sqrt{\gamma_{\min}}/(1-\tau)$ *and if* $\min\{n/q^3,n_1/q_1^3,\dots,n_K/q_K^3\}>(4/C)\log p$ *for some constant* $C=\min\{\tau_{\min}^2\tau^2/288(1-\tau)^2,\tau_{\min}^2\tau^2/72,\tau_{\min}\tau/48\}$*, then there exists a local maximizer of the proposed criterion* (4.10)*,* $\{\widehat{\boldsymbol{\theta}}^{(k)}\}_{k=1}^K$*, such that, with probability tending to 1,*

$$\sum_{k=1}^K\|\widehat{\boldsymbol{\theta}}^{(k)}-\overline{\boldsymbol{\theta}}^{(k)}\|_2\le M\sqrt{\frac{q\log p}{n}}\ , \tag{4.13}$$

for some constant $M>(2KC_\lambda/\tau_{\min}\sqrt{\gamma_{\min}})(3-2\tau)/(2-\tau)$.

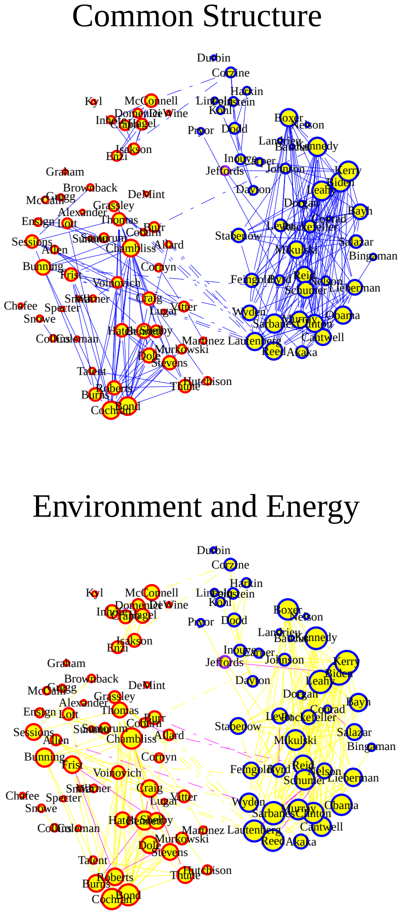
Theorem IV.2. *(Structure selection). Under conditions of Theorem IV.1, with probability tending to 1, we have,*

$$\begin{aligned}\widehat{\theta}_{j,j'}^{(k)} &\neq 0, \quad \textit{for all} \quad (j,j') \in S_k, k = 1, \ldots, K; \\ \widehat{\theta}_{j,j'}^{(k)} &= 0, \quad \textit{for all} \quad (j,j') \in S_k^c, k = 1, \ldots, K \ .\end{aligned}$$

Theorems IV.1 and IV.2 establish the consistency in terms of parameter estimation and structure selection, respectively. The proofs are given in the Appendix.

4.6 Concluding Remarks

We have proposed a joint estimation method for the analysis of heterogenous Markov networks motivated by an application on Senate voting patterns. The method allows the estimation of the networks' common structure by borrowing strength across categories, and allows for individual differences. Asymptotic properties of the method have been established. In particular, we show that the convergence rate is similar to the rate for Gaussian graphical models in a similar context (Guo et al., 2010). The proposed method can be extended to deal with general categorical data with more than two levels using the strategy described in Ravikumar et al. (2010) and Guo et al. (2010). The most interesting feature emerging from the analysis is the existence of more stable associations for the Democrats, both in terms of the common structure and the healthcare and defense categories.


Common Structure

Defense and Security

Environment and Energy

Health and Medical Issues

Figure 4.4: The common and individual structures for the Senate voting data. The nodes represent the 100 senators, with red, blue and purple node colors corresponding to Republican, Democrat, or Independent (Senator Jeffords), respectively. A solid line corresponds to a positive interaction effect and a dashed line to a negative interaction effect. The width of a link is proportional to the magnitude of the corresponding overall interaction effect. For each individual network, the links that only appear in this category are highlighted in purple.

CHAPTER V

Graphical Models with Ordinal Variables

5.1 Introduction

The dependence between *ordinal* variables is not covered by existing graphical models. However, data with such structure have become prevalent recently. For example, each movie available on the Netflix website can be rated by the people watching it. The rating is based on a five point scale and can serve as a guide for future movie watchers. Similar online rating systems are available for books, electronics, travel, restaurants, etc (Koren et al., 2009).

Ordinal variables are very common in survey questionnaires, where respondents are asked to rate an item or to express their level agreement with a particular issue under consideration. Such responses are known to be rated on a Likert scale (Babbie, 2010) and a popular model to analyze such data is the polychotomous Rasch model (von Davier and Carstensen, 2010) that obtains interval level estimates on a continuum, an idea that we explore in this work as well. Another area modeling ordinal variables is regression analysis, where an ordinal response is fitted by a set of numerical covariates. A number of estimation methods for this model exist, including the proportional odds model (Walker and Duncan, 1967; McCullagh, 1980), the partial proportional odds model (Peterson, 1990), the probit model (Bliss, 1935), etc. A comprehensive review of ordinal regression is given in McCullagh and Nelder (1989) and O'Connell (2005).

The objective of this study is to introduce a graphical model for ordinal variables and discuss its efficient estimation under the assumption of sparsity in the dependence structure. The proposed model assumes that the ordinal variables are generated by discretizing the marginal distributions of a latent multivariate Gaussian distribution and the relationships of these ordinal variables are described by the underlying Gaussian graphical model. An EM-like algorithm is developed to efficiently estimate the latent network.

The remainder of the chapter is organized as follows. Section 5.2 presents the probit graphical model and discusses algorithmic and model selection issues. Section 5.3 evaluates the performance of the proposed method by several synthetic examples and Section 5.4 applies the model to explore the network structure between movies from their user ratings.

5.2 Methodology

5.2.1 Probit Graphical Model

Suppose we have p ordinal random variables $X_1, \ldots, X_p$, where $X_j \in \{1, 2, \ldots, K_j\}$ for some integer K_j, which is the number of the ordinal levels in variable j. In the proposed probit graphical model, we assume that there exist p latent random variables $Z_1, \ldots, Z_p$ from a joint Gaussian distribution with mean zero and covariance matrix $\mathbf{\Sigma} = (\sigma_{j,j'})_{p \times p}$, respectively. Without loss of generality, we further assume that Z_j's have unit variances ($\sigma_{j,j} = 1$ for $j = 1, \ldots, p$), i.e., the Z_j's marginally follow standard Gaussian distributions. Each observed variable X_j is discretized from its latent counterpart Z_j. Specifically, for the j-th variable ($j = 1, \ldots, p$), we assume that $(-\infty, +\infty)$ is split into K_j disjointed intervals by a set of thresholds $-\infty = \theta_0^{(j)} < \theta_1^{(j)} < \ldots < \theta_{K_j - 1}^{(j)} < \theta_{K_j}^{(j)} = +\infty$, such that $X_j = k$ if and only if Z_j

locates in interval $[\theta_{k-1}^{(j)}, \theta_k^{(j)}]$. The distribution of Z_j indicates that

$$\Pr(X_j = k) = \Pr(\theta_{k-1}^{(j)} \leq Z_j < \theta_k^{(j)}) = \Phi(\theta_k^{(j)}) - \Phi(\theta_{k-1}^{(j)}), \tag{5.1}$$

where $\Phi(\cdot)$ denotes the cumulative density function of the standard normal distribution.

Letting $\mathbf{\Omega} = \mathbf{\Sigma}^{-1} = (\omega_{j,j'})_{p\times p}$, $\boldsymbol{X} = (X_1, \ldots, X_p)$ and $\boldsymbol{Z} = (Z_1, \ldots, Z_p)$, so that the joint density function of $(\boldsymbol{X}, \boldsymbol{Z})$ can be written as:

$$\begin{aligned} \mathrm{f}(\boldsymbol{X}, \boldsymbol{Z} \mid \mathbf{\Omega}, \mathbf{\Theta}) &= \mathrm{f}(Z_1, \ldots, Z_p \mid \mathbf{\Omega}) \prod_{j=1}^{p} \mathrm{f}_{\mathbf{\Theta}}(X_j | Z_j; \mathbf{\Theta}) \\ &= \frac{\det(\mathbf{\Omega})}{(2\pi)^{p/2}} \exp(-\frac{1}{2} \boldsymbol{Z} \mathbf{\Omega} \boldsymbol{Z}^{\intercal}) \prod_{j=1}^{p} \mathrm{I}(\theta_{X_j - 1}^{(j)} \leq Z_j < \theta_{X_j}^{(j)}) \end{aligned} \tag{5.2}$$

where $\mathbf{\Theta} = \{\theta_k^{(j)} : j = 1, \ldots, p; k = 1, \ldots, K_j\}$, $\mathrm{I}(\cdot)$ is the indicator function and $\mathbf{\Omega}$ the covariance matrix of $Z_1, \ldots, Z_p$. Thus, the marginal probability density function of the observed data is given by

$$\begin{aligned} &\mathrm{f}(\boldsymbol{X} \mid \mathbf{\Omega}, \mathbf{\Theta}) = \\ &\int \cdots \int \mathrm{f}(\boldsymbol{X}, Z_1 = z_1, \ldots, Z_p = z_p \mid \mathbf{\Omega}, \mathbf{\Theta}) dz_p \cdots dz_1 \end{aligned} \tag{5.3}$$

Let $x_{i,j}$ and $z_{i,j}$ be the i-th realization of the observed variable X_j and the latent variable Z_j, respectively. Next, we consider maximizing an ℓ_1-regularized marginal log-likelihood function of the observed data as follows:

$$\sum_{i=1}^{n} \log \mathrm{f}(\boldsymbol{x}_i \mid \mathbf{\Omega}, \mathbf{\Theta}) - \lambda \sum_{j \neq j'} |\omega_{j,j'}|. \tag{5.4}$$

where $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,p})$. The model maximizing criterion (5.4) is referred to as *probit graphical model*, which is motivated by the probit regression model (Bliss, 1935) and the polychotomous Rasch model (von Davier and Carstensen, 2010). The tuning parameter λ in criterion (5.4) controls the degree of the sparsity in the underlying network. When λ is large enough, some $\omega_{j,j'}$'s can be shrunk to zeros, resulting in the removal of the corresponding links in the underlying network. Numerically, it is difficult to solve criterion (5.4) directly due to the existence of the integral in (5.3). We introduce next an EM-like algorithm to estimate (5.4) in an iterative manner.

5.2.2 Algorithm for Probit Graphical Model

Criterion (5.4) depends on the parameters $\boldsymbol{\Theta}$ and $\boldsymbol{\Omega}$ and the latent variable $\boldsymbol{Z}$. The former has a closed-form estimator. Specifically, for each $j = 1, \ldots, p$, we set

$$\widehat{\theta}_k^{(j)} = \begin{cases} -\infty, & \text{if } k = 0; \\ \Phi^{-1}(\sum_{i=1}^n \mathrm{I}(x_{i,j} \leq k)/n), & \text{if } k = 1, \ldots, K_j - 1; \\ +\infty, & \text{if } k = K_j. \end{cases} \tag{5.5}$$

where Φ^{-1} is the inverse function of the cumulative density function of standard normal distribution. We can show that $\widehat{\boldsymbol{\Theta}}$ consistently estimates $\boldsymbol{\Theta}$. The estimation of $\boldsymbol{\Omega}$, on the other hand, is nontrivial due to the multiple integrals in criterion (5.3). To address this problem, we applied the EM algorithm to solving (5.4), where the latent variables $z_{i,j}$'s ($i = 1, \ldots, n; j = 1, \ldots, p$) are treated as "missing data" and are imputed in the E-step, and the parameter $\boldsymbol{\Omega}$ is estimated in the M-step.

Suppose $\widehat{\boldsymbol{\Omega}}$ is the updated estimate of $\boldsymbol{\Omega}$ updated in the M-step, then the E-step computes the conditional expectation of the joint log-likelihood given the estimates $\widehat{\boldsymbol{\Theta}}$ and $\widehat{\boldsymbol{\Omega}}$:

$$\begin{aligned} \mathrm{Q}(\boldsymbol{\Omega}) &= \sum_{i=1}^n \mathrm{E}[\log f(\boldsymbol{x}_i, \boldsymbol{z}_i \mid \boldsymbol{\Theta}, \boldsymbol{\Omega}) \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}] \\ &= \frac{n}{2}[\log\det(\boldsymbol{\Theta}) - \mathrm{trace}(\boldsymbol{S}\boldsymbol{\Theta}) - p\log(2\pi)] \end{aligned} \tag{5.6}$$

where $\boldsymbol{z}_i = (z_{i,1}, \ldots, z_{i,p})$ and trace$(\cdot)$ is the matrix trace. Criterion (5.6) is usually referred to as Q-function in the literature. $\boldsymbol{S}$ is a $p \times p$ matrix whose (j, j')-th element is $s_{j,j'} = 1/n \sum_{i=1}^{n} \mathrm{E}(z_{i,j} z_{i,j'} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})$ $(1 \leq j, j' \leq p)$. Given $\boldsymbol{x}_i$, the conditional distribution of $z_{i,1}, \ldots, z_{i,p} \mid \boldsymbol{x}_i$ is equivalent to that of $z_{i,1}, \ldots, z_{i,p} \mid \theta^{(1)}_{x_{i,1}-1} \leq z_{i,1} \leq \theta^{(1)}_{x_{i,1}}, \ldots, \theta^{(p)}_{x_{i,p}-1} \leq z_{i,p} \leq \theta^{(p)}_{x_{i,p}}$, which the follows a truncated multivariate Gaussian distribution defined on a hyper-cube $[\theta^{(1)}_{x_{i,1}-1}, \theta^{(1)}_{x_{i,1}}] \times \ldots \times [\theta^{(p)}_{x_{i,p}-1}, \theta^{(p)}_{x_{i,p}}]$. Therefore, $\mathrm{E}(z_{i,j} z_{i,j'} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})$ is the second moment of a truncated multivariate Gaussian distribution and it can be directly estimated using the algorithms proposed by Tallis (1961), Lee (1979) and Leppard and Tallis (1989). Nevertheless, the computational cost of these direct estimation algorithms is extremely high and thus not suitable for even moderate size problems. An alternative approach is based on the Markov-chain-Monte-Carlo (MCMC) method. Specifically, we randomly generate a sequence of samples from the conditional $\boldsymbol{z}_i \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}$ using a Gibbs sampler from a multivariate truncated normal distribution (Kotecha and Djuric, 1999) and then $\mathrm{E}(z_{i,j} z_{i,j'} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})$ is estimated by the empirical second moment of the conditional from these samples. Although the MCMC approach is faster than the direct estimation method, it is still lack of efficiency for large scale networks. To address the computational issue, we develop an efficient approximate estimation algorithm whose details are discussed in Section 5.2.3.

The M-step updates $\boldsymbol{\Omega}$ by maximizing the ℓ_1-regularized Q-function (up to a constant and a factor):

$$\max_{\boldsymbol{\Omega}} \; \log\det(\boldsymbol{\Omega}) - \mathrm{trace}(\boldsymbol{S}\boldsymbol{\Omega}) - \lambda \sum_{j \neq j'} |\omega_{j,j'}|, \tag{5.7}$$

Criterion (5.7) can be solved efficiently by a few existing algorithms such as graphical lasso (Friedman et al., 2008) and SPICE (Rothman et al., 2008). The maximizer of (5.7) is denoted by $\widetilde{\boldsymbol{\Omega}}$. Nevertheless, the estimated covariance matrix, $\widetilde{\boldsymbol{\Sigma}} = \widetilde{\boldsymbol{\Omega}}^{-1}$, does not necessarily possess unit diagonal elements as assumed by the probit graphical model. Therefore, we post-process $\widetilde{\boldsymbol{\Sigma}}$ by scaling it to a unit-diagonal matrix $\widehat{\boldsymbol{\Sigma}}$ and update $\widehat{\boldsymbol{\Omega}} = \widehat{\boldsymbol{\Sigma}}^{-1}$, which will be used in the E-step of the next iteration.

5.2.3 Approximation of the Conditional Expectation

Noting that when $j = j'$, the corresponding conditional expectation is the second moment of the conditional $z_{i,j} \mid \boldsymbol{x}_i$, i.e., $\mathrm{E}(z_{i,j}^2 \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})$; when $j \neq j'$, we use the mean field theory (Peterson and Anderson, 1987) to approximate $E(z_{i,j} z_{i,j'} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}) \approx E(z_{i,j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}) E(z_{i,j'} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})$. With this approximation, it is sufficient to estimate the first moment $E(z_{i,j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})$ and the second moment $E(z_{i,j}^2 \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})$, respectively. We need to point out that, in general, the latent variable $z_{i,j}$ not only depends on $x_{i,j}$, but also on all other observed variables $\boldsymbol{x}_{i,-j} = (x_{i,1}, \ldots, x_{i,j-1}, x_{i,j+1}, \ldots, x_{i,p})$. By applying the iterate expectation equation, we can express the first and second moments of the conditional $z_{i,j} \mid \boldsymbol{x}_i$ as follows:

$$E(z_{i,j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}) = E[E(z_{i,j} \mid \boldsymbol{z}_{i,-j}, x_{i,j}; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}) \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}] \tag{5.8}$$

$$E(z_{i,j}^2 \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}) = E[E(z_{i,j}^2 \mid \boldsymbol{z}_{i,-j}, x_{i,j}; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}) \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}] \tag{5.9}$$

where $\boldsymbol{z}_{i,-j} = (z_{i,1}, \ldots, z_{i,j-1}, z_{i,j+1}, \ldots, z_{i,p})$. The interior expectation in (5.8) and (5.9) are relatively straightforward to compute. Indeed, given parameter $\widehat{\boldsymbol{\Omega}}$, $z_{i,1}, \ldots, z_{i,p}$ jointly follow a multivariate Gaussian distribution with mean zero and covariance matrix $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Omega}}^{-1}$. The property of Gaussian distribution shows that, given $\boldsymbol{z}_{i,-j}$, the conditional $z_{i,j} \mid \boldsymbol{z}_{i,-j}$ follows a Gaussian distribution with mean $\widetilde{\mu}_{i,j} = \widehat{\boldsymbol{\Sigma}}_{j,-j} \widehat{\boldsymbol{\Sigma}}_{-j,-j}^{-1} \boldsymbol{z}_{i,-j}^{\intercal}$ and variance $\widetilde{\sigma}_{i,j}^2 = 1 - \widehat{\boldsymbol{\Sigma}}_{j,-j} \widehat{\boldsymbol{\Sigma}}_{-j,-j}^{-1} \widehat{\boldsymbol{\Sigma}}_{-j,j}$, respectively. Moreover, given the observed data $x_{i,j}$, the conditional $z_{i,j} \mid \boldsymbol{z}_{i,-j}, x_{i,j}$ in the RHS of equation (5.8) is equivalent to $z_{i,j} \mid \boldsymbol{z}_{i,-j}, \theta_{x_{i,j}-1}^{(j)} \leq z_{i,j} \leq \theta_{x_{i,j}}^{(j)}$, which follows a truncated Gaussian distribution defined on interval $[\theta_{x_{i,j}-1}^{(j)}, \theta_{x_{i,j}}^{(j)}]$. The following lemma gives the closed-form expression of the first and second moments of the truncated Gaussian distribution.

Lemma V.1. *Suppose that a random variable Y follows a Gaussian distribution with mean μ_0 and variance σ_0. Then, for any constant t_1 and t_2, $Y \mid t_1 \leq Y \leq t_2$ follows a truncated Gaussian distribution defined on $[t_1, t_2]$. Let $\xi_1 = (t_1 - \mu_0)/\sigma_0$ and $\xi_2 = (t_2 - \mu_0)/\sigma_0$, then the first and second moments of $Y \mid t_1 \leq Y \leq t_2$ are:*

$$\begin{aligned}
\mathrm{E}(Y \mid t_1 \leq Y \leq t_2) &= \mu_0 + \frac{\phi(\xi_1) - \phi(\xi_2)}{\Phi(\xi_2) - \Phi(\xi_1)}\sigma_0 && (5.10)\\
\mathrm{E}(Y^2 \mid t_1 \leq Y \leq t_2) &= \mu_0^2 + \sigma_0^2 + 2\frac{\phi(\xi_1) - \phi(\xi_2)}{\Phi(\xi_2) - \Phi(\xi_1)}\mu_0\sigma_0 \\
&\quad + \frac{\xi_1\phi(\xi_1) - \xi_2\phi(\xi_2)}{\Phi(\xi_2) - \Phi(\xi_1)}\sigma_0^2 && (5.11)
\end{aligned}$$

For more properties of the truncated Gaussian distribution, we refer the readers to Johnson et al. (1994).

Let $\delta_{i,j,k} = (\theta_k^{(j)} - \widetilde{\mu}_{i,j})/\widetilde{\sigma}_{i,j}$, then by applying Lemma V.1 to the conditional $z_{i,j} | \boldsymbol{z}_{i,-j}, x_{i,j}$, we obtain:

$$\begin{aligned}
\mathrm{E}(z_{i,j} | \boldsymbol{z}_{i,-j}, x_{i,j}; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}) &= \widetilde{\mu}_{i,j} + a_{i,j}\widetilde{\sigma}_{i,j}\ , && (5.12)\\
\mathrm{E}(z_{i,j}^2 | \boldsymbol{z}_{i,-j}, x_{i,j}; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}) &= \widetilde{\mu}_{i,j}^2 + \widetilde{\sigma}_{i,j}^2 + 2a_{i,j}\widetilde{\mu}_{i,j}\widetilde{\sigma}_{i,j} + b_{i,j}\widetilde{\sigma}_{i,j}^2 && (5.13)
\end{aligned}$$

where $a_{i,j} = [\phi(\delta_{i,j,x_{i,j}-1}) - \phi(\delta_{i,j,x_{i,j}}]/[\Phi(\delta_{i,j,x_{i,j}}) - \Phi(\delta_{i,j,x_{i,j}-1})]$ and $b_{i,j} = [\delta_{i,j,x_{i,j}-1}\phi(\delta_{i,j,x_{i,j}-1}) - \delta_{i,j,x_{i,j}}\phi(\delta_{i,j,x_{i,j}})]/[\Phi(\delta_{i,j,x_{i,j}}) - \Phi(\delta_{i,j,x_{i,j}-1})]$, respectively.

Now we plug equations (5.12) and (5.13) into (5.8) and (5.9), respectively. Since $\widetilde{\mu}_{i,j}$, $a_{i,j}$ and $b_{i,j}$ depend on the latent variables $z_{i,j}$'s, the outer expectations in (5.8) and (5.9) depend on the following items: $\mathrm{E}(\widetilde{\mu}_{i,j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})$, $\mathrm{E}(a_{i,j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})$, $\mathrm{E}(b_{i,j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})$ and $\mathrm{E}(a_{i,j}\widetilde{\mu}_{i,j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})$. Note that $\widetilde{\mu}_{i,j}$ is a linear function of $\boldsymbol{z}_{i,-j}$ and $\widetilde{\sigma}_{i,j}$ is a constant irrelevant to the latent data. For each $i = 1, \ldots, n$ and $j = 1, \ldots, p$, the conditional expectation of $\widetilde{\mu}_{i,j}$ is

$$\mathrm{E}(\widetilde{\mu}_{i,j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Omega}}) = \widehat{\boldsymbol{\Sigma}}_{j,-j} \widehat{\boldsymbol{\Sigma}}_{-j,-j}^{-1} \mathrm{E}(\boldsymbol{z}_{i,-j} | \boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Omega}})^{\mathsf{T}} \tag{5.14}$$

Nevertheless, $a_{i,j}$ and $b_{i,j}$ are nonlinear functions of $\widetilde{\mu}_{i,j}$, and thus of $\boldsymbol{z}_{i,-j}$. Therefore, we consider the following approximations:

$$\mathrm{E}(a_{i,j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Omega}}) \approx \frac{\phi(\widetilde{\delta}_{i,j,x_{i,j}-1}) - \phi(\widetilde{\delta}_{i,j,x_{i,j}})}{\Phi(\widetilde{\delta}_{i,j,x_{i,j}}) - \Phi(\widetilde{\delta}_{i,j,x_{i,j}-1})} \tag{5.15}$$

$$\mathrm{E}(b_{i,j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Omega}}) \approx \frac{\widetilde{\delta}_{i,j,x_{i,j}-1}\phi(\widetilde{\delta}_{i,j,x_{i,j}-1}) - \widetilde{\delta}_{i,j,x_{i,j}}\phi(\widetilde{\delta}_{i,j,x_{i,j}})}{\Phi(\widetilde{\delta}_{i,j,x_{i,j}}) - \Phi(\widetilde{\delta}_{i,j,x_{i,j}-1})} \tag{5.16}$$

where $\widetilde{\delta}_{i,j,x_{i,j}} = [\theta_k^{(j)} - \mathrm{E}(\widetilde{\mu}_{i,j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Omega}})]/\widetilde{\sigma}_{i,j}$. Finally, we approximate $\mathrm{E}(a_{i,j}\widetilde{\mu}_{i,j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}) \approx \mathrm{E}(a_{i,j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})\mathrm{E}(\widetilde{\mu}_{i,j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})$. Therefore, (5.8) and (5.9) can be approximated by

$$\begin{aligned}
&\mathrm{E}(z_{i,j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}) \approx \\
&\widehat{\boldsymbol{\Sigma}}_{j,-j} \widehat{\boldsymbol{\Sigma}}_{-j,-j}^{-1} \mathrm{E}({\boldsymbol{z}_{i,-j}}^{\mathsf{T}} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}) \\
&+ \frac{\phi(\widetilde{\delta}_{i,j,x_{i,j}-1}) - \phi(\widetilde{\delta}_{i,j,x_{i,j}})}{\Phi(\widetilde{\delta}_{i,j,x_{i,j}}) - \Phi(\widetilde{\delta}_{i,j,x_{i,j}-1})} \widetilde{\sigma}_{i,j}
\end{aligned} \tag{5.17}$$

$$\begin{aligned}
&\mathrm{E}(z_{i,j}^2 \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}) \approx \\
&\widehat{\boldsymbol{\Sigma}}_{j,-j} \widehat{\boldsymbol{\Sigma}}_{-j,-j}^{-1} \mathrm{E}({\boldsymbol{z}_{i,-j}}^{\mathsf{T}} \boldsymbol{z}_{i,-j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}) \widehat{\boldsymbol{\Sigma}}_{-j,-j}^{-1} \widehat{\boldsymbol{\Sigma}}_{j,-j}^{\mathsf{T}} + \widetilde{\sigma}_{i,j}^2 + \\
&2 \frac{\phi(\widetilde{\delta}_{i,j,x_{i,j}-1}) - \phi(\widetilde{\delta}_{i,j,x_{i,j}})}{\Phi(\widetilde{\delta}_{i,j,x_{i,j}}) - \Phi(\widetilde{\delta}_{i,j,x_{i,j}-1})} [\widehat{\boldsymbol{\Sigma}}_{j,-j} \widehat{\boldsymbol{\Sigma}}_{-j,-j}^{-1} \mathrm{E}({\boldsymbol{z}_{i,-j}}^{\mathsf{T}} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})] \widetilde{\sigma}_{i,j} \\
&+ \frac{\delta_{i,j,x_{i,j}-1}^{(j)} \phi(\widetilde{\delta}_{i,j,x_{i,j}-1}) - \widetilde{\delta}_{i,j,x_{i,j}} \phi(\widetilde{\delta}_{i,j,x_{i,j}})}{\Phi(\widetilde{\delta}_{i,j,x_{i,j}}) - \Phi(\widetilde{\delta}_{i,j,x_{i,j}-1})} \widetilde{\sigma}_{i,j}^2
\end{aligned} \tag{5.18}$$

Equations (5.17) and (5.18) establish the recursive relationships among the elements in $\mathrm{E}(\boldsymbol{z}_i \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})$ and $\mathrm{E}({\boldsymbol{z}_i}^{\mathsf{T}} \boldsymbol{z}_i \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})$, respectively. Therefore, it is natural that they can be estimated by an iterative procedure. Algorithm 5.2.3 summarizes the main steps of the proposed combined estimation procedure outlined in Sections 5.2.2 and 5.2.3.

```
1: Initialize $\mathrm{E}(z_{i,j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}) \approx \mathrm{E}(z_{i,j} \mid x_{i,j}; \widehat{\boldsymbol{\Theta}})$ and $\mathrm{E}(z_{i,j}^2 \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}) \approx \mathrm{E}(z_{i,j} \mid x_{i,j}; \widehat{\boldsymbol{\Theta}})$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$;
2: Initialize $s_{j,j'} = 1/n \sum_{i=1}^{n} \mathrm{E}(z_{i,j} \mid x_{i,j}; \widehat{\boldsymbol{\Theta}}) \mathrm{E}(z_{i,j'} \mid x_{i,j'}; \widehat{\boldsymbol{\Theta}})$ for $1 \leq j \neq j' \leq p$ and $s_{j,j} = 1/n \sum_{i=1}^{n} \mathrm{E}(z_{i,j} \mid x_{i,j}; \widehat{\boldsymbol{\Theta}})$ for $j = 1, \ldots, p$, then estimate $\widehat{\boldsymbol{\Omega}}$ by maximizing criterion (5.7);
   {Start outer loop}
3: repeat
4:    E-step: estimate $\boldsymbol{S}$ in (5.6);
      {Start inner loop}
5:    repeat
6:       for $i = 1$ to n do
7:          if $j = j'$ then
8:             Update $\mathrm{E}(z_{i,j}^2 \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})$ using RHS of equation (5.18) for $j = 1, \ldots, p$;
9:          else
10:            Update $\mathrm{E}(z_{i,j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})$ using RHS of equation (5.17) for $j = 1, \ldots, p$ and then set $\mathrm{E}(z_{i,j} z_{i,j'} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}) = \mathrm{E}(z_{i,j} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}}) \mathrm{E}(z_{i,j'} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})$ for $1 \leq j \neq j' \leq p$;
11:         end if
12:      end for
13:      Update $s_{j,j'} = 1/n \sum_{i=1}^{n} \mathrm{E}(z_{i,j} z_{i,j'} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Omega}})$ for $1 \leq j, j' \leq p$;
14:   until The inner loop converges;
15:   M-step: update $\widehat{\boldsymbol{\Omega}}$ by maximizing criterion (5.7);
16: until The outer loop converges.
```

In Algorithm 5.2.3, Lines 1–2 initialize the expectation of the conditional $z_{i,j} \mid \boldsymbol{x}_i$ and the parameter $\widehat{\boldsymbol{\Omega}}$. Lines 3–16 establish the outer loop which iteratively computes the E-step and the M-step. In the E-step, Lines 5–14 consist of the inner loop which recursively estimate the first and second moments of the conditional $z_{i,j} \mid \boldsymbol{x}_i$. It can be seen that the complexity of the inner loop is $O(np^2)$, which is the same as that of the Graphical Lasso algorithm in the M-step. Therefore, the overall complexity of Algorithm 5.2.3 is $O(np^2)$.

5.2.4 Model Selection

In the probit graphical model, the tuning parameter λ controls the sparsity of the resulting estimator and it can be selected using the cross validation method. Specifically, we randomly split the observed data $\boldsymbol{X}$ into D subsets with similar sizes and denote the index set of the observations in the d-th subset as $\mathcal{T}_d$ $(d = 1, \ldots, D)$. For any pre-specified λ, we denote $\widehat{\boldsymbol{\Omega}}_\lambda^{[-d]}$ as a maximizer of criterion (5.4) estimated by Algorithm 5.2.3 using all observations except those in $\mathcal{T}_d$. We also denote $\widehat{\boldsymbol{\Theta}}^{[-d]}$ and $\boldsymbol{S}^{[d]} = (s_{j,j'}^{[d]})_{p\times p}$ as the analogs of $\widehat{\boldsymbol{\Theta}}$ and $\boldsymbol{S}$ in Section 5.2.2 but restricted on data in $\mathcal{T}_d^c$ and $\mathcal{T}_d$, respectively. In particular, each element of $\boldsymbol{S}^{[d]}$ is defined as $s_{j,j'}^{[d]} = (1/|\mathcal{T}_d|)\sum_{i\in\mathcal{T}_d} \mathrm{E}(z_{i,j}z_{i,j'} \mid \boldsymbol{x}_i; \widehat{\boldsymbol{\Theta}}^{[-d]}, \widehat{\boldsymbol{\Omega}}_\lambda^{[-d]})$ $(1 \leq j, j' \leq p)$, where $|\mathcal{T}_d|$ is the cardinality of $\mathcal{T}_d$. Given $\widehat{\boldsymbol{\Theta}}^{[-d]}$ and $\widehat{\boldsymbol{\Omega}}_\lambda^{[-d]}$, $\boldsymbol{S}^{[d]}$ can be estimated by the algorithm introduced in Section 5.2.3, i.e., the inner loop of Algorithm 5.2.3. Thus, the optimal tuning parameter can be selected by maximizing the following criterion:

$$\max_{\lambda} \ \log\det(\widehat{\boldsymbol{\Omega}}_\lambda^{[-d]}) - \mathrm{trace}(\boldsymbol{S}^{[d]}\widehat{\boldsymbol{\Omega}}_\lambda^{[-d]}) - p\log(2\pi) \tag{5.19}$$

5.3 Simulated Examples

In this section, we use two sets of simulated experiments to illustrate the performance of the probit graphical model. The first set aims at comparing the computational cost of the three methods estimating the Q-function in E-step, namely the direct estimation, the MCMC estimation and the approximation algorithm. The second set compares the performance of the probit graphical model using the approximation algorithm to that of the Gaussian graphical model.

5.3.1 Computational Cost and Performance

In this experiment, we simulate a low-dimensional data set with $p = 5$ variables and $n = 10$ observations. Specifically, we define the underlying inverse covariance matrix $\boldsymbol{\Omega}$ as a tri-diagonal matrix with 1s on the main diagonal and 0.5 on the first off diagonal. Then, for $i = 1, \ldots, n$, we generate the latent data $\boldsymbol{z}_i = (z_{i,1}, \ldots, z_{i,p})$ from $N(\mathbf{0}, \boldsymbol{\Omega}^{-1})$ and discretize them as follows: for each $j = 1, \ldots, p$, set

$$\theta_k^{(j)} = \begin{cases} -\infty, & \text{if } k = 0; \\ \Phi^{-1}(0.2) & \text{if } k = 1; \\ \Phi^{-1}(0.4) & \text{if } k = 2; \\ +\infty, & \text{if } k = 3. \end{cases} \tag{5.20}$$

and $x_{i,j} = \sum_{k=1}^{3} \mathrm{I}(z_{i,j} \geq \theta_k^{(j)})$ $(i = 1, \ldots, n; j = 1, \ldots, p)$, i.e., the value of $x_{i,j}$ is k if it locates in interval $[\theta_{k-1}^{(j)}, \theta_k^{(j)}]$.

Table 5.1: Comparison of the CPU time. "Probit-Direct", "Probit-Gibbs" and "Probit-Approximate" represent the probit graphical model with direct, Gibbs sampling and approximate estimation to the second moment in E-step. The quantities plotted correspond to the median CPU time over different tuning parameters and the numbers in the parentheses to their median absolute deviation.

Method	CPU time in seconds
Probit-Direct	3310.21 (199.95)
Probit-Gibbs	46.17 (1.51)
Probit-Approximate	0.04 (0.03)

The probit graphical model estimated using the direct, Gibbs sampling and approximate methods is applied to this data set and the computational cost shown in Table 5.1. We can see that the median CPU time of the approximate estimation is only about 1/1,000 of that of Gibbs sampling and about 1/80,000 of that of the direct estimation method. Therefore, the approximate estimation is orders of magnitude more efficient that its competitors and hence suitable for large scale problems.

To evaluate the estimation performance of the competing methods we use the Frobenius and entropy loss functions defined next. Given the estimate $\widehat{\mathbf{\Omega}}^{(k)}$, the Frobenius (FL) and the entropy loss (EL) are given by:

$$\begin{aligned} FL &= \|h(\mathbf{\Omega}) - h(\widehat{\mathbf{\Omega}})\|_F^2 / \|h(\mathbf{\Omega})\|_F^2 \ . \\ EL &= \text{trace}(\mathbf{\Omega}^{-1}\widehat{\mathbf{\Omega}}) - \log[\det(\mathbf{\Omega}^{-1}\widehat{\mathbf{\Omega}})] - p \ , \end{aligned} \tag{5.21}$$

where the function $h(\cdot)$ in (5.21) scales the matrix to the one with unit diagonal elements.

The performance of the three estimation methods are shown in Figure 5.1. We can see that the Frobenius loss of the approximate estimation is slightly higher than those of the other two methods when the tuning parameter λ is relative small. When λ gets larger, the losses of the direct and the Gibbs estimations increase dramatically compared to the approximate one. It can be concluded that the approximate estimation method leads to very large computational improvements with marginal sacrifices in terms of estimation efficiency.

5.3.2 High-dimensional Experiments

In this section, we evaluate the performance of the proposed method by simulation study. These examples simulate four types of network structures: a scale-free graph, a hub graph, a nearest-neighbor graph and a block graph. Each network consists of $p = 50$ nodes. The details of these networks are described as follows:

Example 1: Scale-free Graph. A scale-free graph has a power-law degree distribution and can be simulated by the Barabasi-Albert algorithm (Barabasi and Albert, 1999). A realization of a scale-free network is depicted in Figure 5.2 (A).

Example 2: Hub Graph. A hub graph consists of a few high-degree nodes (hubs)

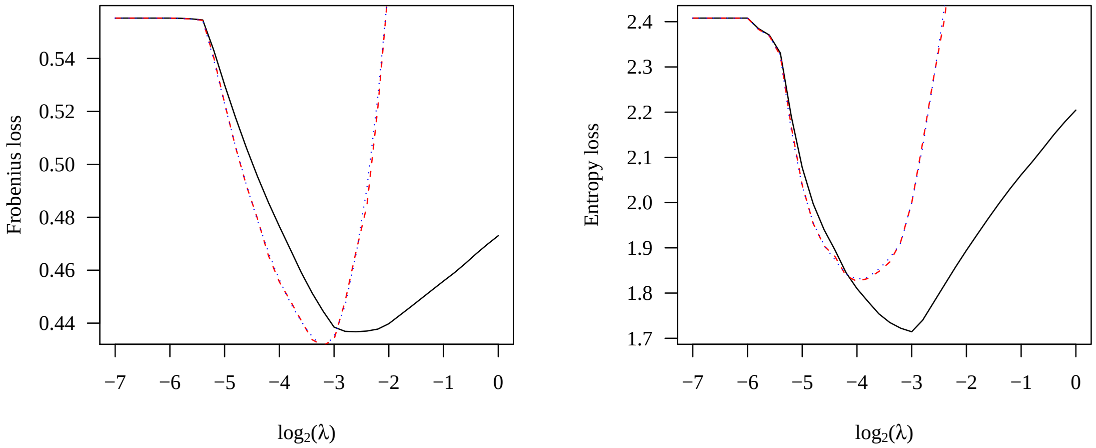


Figure 5.1: The comparison of Frobenius loss and Entropy loss over different tuning parameters. The direct estimation, the Gibbs sampling estimation and the approximation estimation are represented by blue dotted, red dashed and black solid lines.

and a large amount of low-degree nodes. In this example, we follow the simulation setting in Peng et al. (2009) and generate a hub graph by inserting a few hub nodes into a very sparse graph. Specifically, the graph consists of three hubs with degrees around eight, and the other 47 nodes with degrees at most three. An example hub graph is shown in Figure 5.2 (B).

Example 3: Nearest-neighbor Graph. To generate the nearest neighbor graphs, we slightly modify the data generating mechanism described in Li and Gui (2006). Specifically, we generate p points randomly on a unit square, calculate all $p(p-1)/2$ pairwise distances, and find the m nearest neighbors of each point in terms of these distances. The nearest neighbor network is obtained by linking any two points that are m-nearest neighbors of each other. The integer m controls the degree of sparsity of the network and the value $m = 5$ was chosen in the simulation study. Figure 5.2 (C) exhibits one realization of the

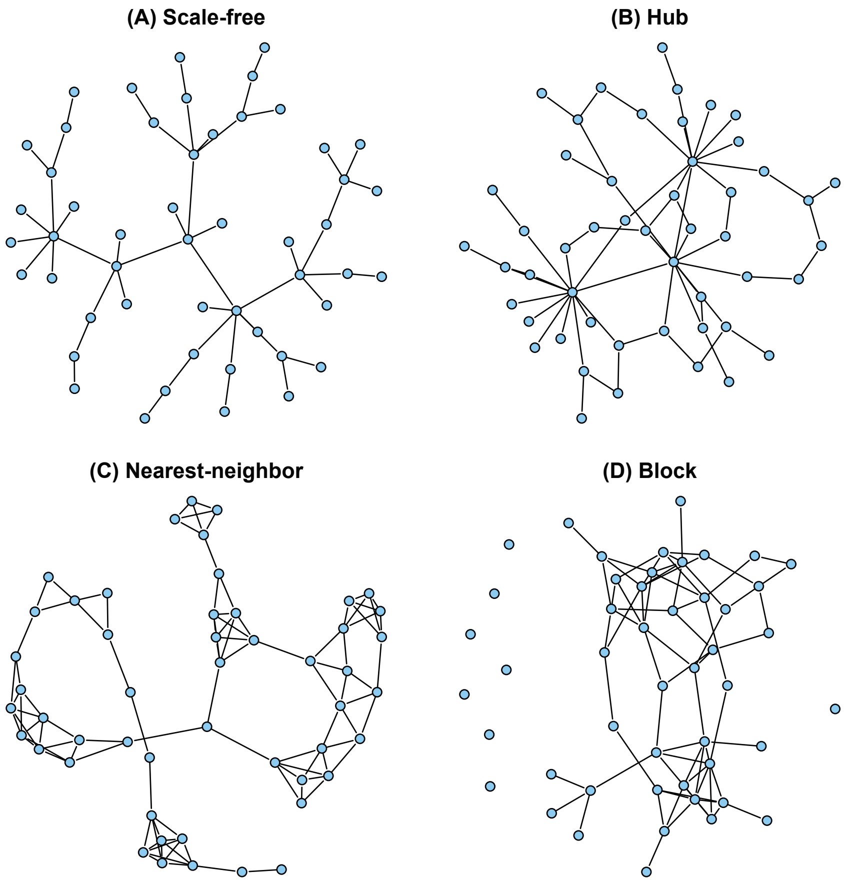


Figure 5.2: Illustration of the networks used in four simulated examples: scale-free graph, hub graph, nearest-neighbor graph and block graph.

nearest-neighbor network.

Example 4: Block Graph. In this setting, we generate a block graph using a symmetric random adjacency matrix with two dense blocks. Specifically, the blocks associated with nodes 1–20 and nodes 21–30 have densities 0.2 and 0.5, respectively, whereas all other parts in the matrix have a density 0.02 (background density). Figure 5.2 (D) illustrates such a random graph with two blocks.

The ordinal data sets are generated as follows. The first step is to generate the inverse covariance matrix $\boldsymbol{\Omega}$ of the latent multivariate Gaussian distribution. Specifically, each off-diagonal element $\omega_{j,j'}$ is drawn uniformly from $[-1,-0.5]\cup[0.5,1]$

if nodes j and j' are linked by an edge, otherwise $\omega_{j,j'} = 0$. Further, the diagonal elements were all set to be 2 to ensure the positive definiteness. The second step is to generate the latent data $\boldsymbol{z}_i = (z_{i,1}, \ldots, z_{i,p})$ i.i.d. from $N(\mathbf{0}, \boldsymbol{\Omega}^{-1})$. Finally, the continuous latent data $\boldsymbol{z}_i$'s are discretized into ordinal scale with three levels by thresholding. Specifically, for each $j = 1, \ldots, p$, we set

$$\theta_k^{(j)} = \begin{cases} -\infty, & \text{if } k = 0; \\ \Phi^{-1}(0.1) & \text{if } k = 1; \\ \Phi^{-1}(0.2) & \text{if } k = 2; \\ +\infty, & \text{if } k = 3. \end{cases} \tag{5.22}$$

and set $x_{i,j} = \sum_{k=1}^{3} \mathrm{I}(z_{i,j} \geq \theta_k^{(j)})$ $(i = 1, \ldots, n; j = 1, \ldots, p)$. For each example, we tried different sample sizes: n=50, 100, 200 and 500, respectively. In each setting, we generate 50 replicated data sets randomly.

We compare the proposed probit graphical model with two other methods. The first one applies the graphical lasso algorithm to the ordinal data $\boldsymbol{X}$ directly and the second one applies graphical lasso to the latent numerical data $\boldsymbol{Z}$. We refer to the second method as an oracle method because it simulates an ideal situation where $\boldsymbol{Z}$ is exactly recovered. This never happens in real data analysis, but we still put it here as a benchmark. In this work, the receiver operating characteristic (ROC) curves was used to evaluate the accuracy of network structure estimation. The ROC curve plots the sensitivity (the proportion of correctly detected links) against the false positive rate (the proportion of mis-identified zeros) over a range of values of the tuning parameter λ. In addition, the Frobenius loss and the entropy loss defined in (5.21) and (5.21) were used to evaluate the performance of the parameter estimation.

Figure 5.3 shows the ROC curves for all simulated examples. The curves are averaged over 50 replications. The oracle model provides a benchmark curve for each setting (blue dotted line in each panel). When the sample size is relatively small

(n=50, 100 or 200), it turns out that the probit model (dark solid line) dominates the Gaussian model (red dashed line). When the sample size gets larger, the two methods exhibit similar performance.

Table 5.2 summarizes the parameter estimation measured by the Frobenius loss and the entropy loss. The results were averaged over 50 replications and the tuning parameter λ was selected using the cross validation method introduced in Section 5.2.4. There is no doubt that the oracle model performs the best and its result provides a benchmark for other competitors. It is more informative to compare the other two models based on the observed data $\boldsymbol{X}$. We can see that the Frobenius losses from the probit model are consistently lower than those from the Gaussian model. The advantage is more significant when the sample sizes are moderate (n=100 or 200). In terms of the entropy loss, we can see that the the probit model outperforms the Gaussian model for relative large sample sizes, such as n=200 or 500.

Table 5.2: The Frobenius losses and entropy losses estimated by probit graphical model, the oracle model and the Gaussian model. The oracle model and the Gaussian model applies the Glasso algorithm to the latent data $\boldsymbol{Z}$ and the observed data $\boldsymbol{X}$, respectively. The results are averaged over 50 replications and the corresponding standard deviations and recorded in the parenthesis.

Example	n	Frobenius Loss			Entropy Loss		
		Gaussian	Oracle	Probit	Gaussian	Oracle	Probit
Scale-free	50	2.3 (0.12)	0.7 (0.05)	2.2 (0.13)	12.0 (0.73)	3.1 (0.29)	23.1 (1.83)
	100	2.2 (0.13)	0.4 (0.08)	1.7 (0.09)	9.4 (0.68)	1.9 (0.29)	10.1 (0.45)
	200	1.7 (0.12)	0.3 (0.02)	1.2 (0.04)	6.4 (0.33)	1.1 (0.10)	5.4 (0.26)
	500	0.9 (0.05)	0.1 (0.01)	0.7 (0.04)	3.3 (0.19)	0.5 (0.05)	2.7 (0.19)
Hub	50	1.2 (0.06)	0.3 (0.02)	1.1 (0.04)	21.2 (1.32)	5.8 (0.70)	29.4 (1.76)
	100	1.1 (0.10)	0.1 (0.01)	0.8 (0.03)	15.9 (1.03)	3.2 (0.27)	15.1 (0.64)
	200	0.8 (0.05)	0.1 (0.01)	0.6 (0.01)	11.9 (0.39)	1.8 (0.23)	10.4 (0.33)
	500	0.6 (0.02)	0.0 (0.00)	0.5 (0.01)	9.1 (0.16)	0.7 (0.06)	7.5 (0.16)
Nearest-neighbor	50	1.4 (0.04)	0.6 (0.02)	1.3 (0.06)	16.5 (0.80)	5.6 (0.30)	25.6 (2.04)
	100	1.3 (0.08)	0.4 (0.02)	1.0 (0.02)	12.1 (0.52)	3.5 (0.36)	12.4 (0.76)
	200	1.0 (0.04)	0.2 (0.01)	0.7 (0.03)	8.6 (0.32)	2.0 (0.11)	7.5 (0.17)
	500	0.6 (0.03)	0.1 (0.01)	0.5 (0.02)	5.5 (0.12)	0.8 (0.02)	4.5 (0.19)
Random-block	50	1.8 (0.05)	0.7 (0.05)	1.7 (0.04)	14.8 (1.04)	4.7 (0.46)	23.5 (1.76)
	100	1.6 (0.16)	0.4 (0.02)	1.3 (0.03)	10.7 (1.10)	2.9 (0.27)	11.3 (0.46)
	200	1.3 (0.05)	0.2 (0.03)	0.9 (0.05)	7.2 (0.19)	1.6 (0.11)	6.3 (0.32)
	500	0.7 (0.03)	0.1 (0.01)	0.6 (0.03)	4.1 (0.15)	0.7 (0.06)	3.5 (0.13)

5.4 Application to Movie Rating Records

In the section, we applied the probit graphical model to Movielens, a data set recording the rating scores for 1682 movies rated by 943 users. The rating scores have five levels, where one corresponds to strong dissatisfaction and five to strong satisfaction. In the original data matrix, more than 90% entries are missed. To address this, we selected a sub-matrix with 193 users and 32 movies such that the averaged proportion of missed ratings in these movies is less than 15%. Each missing value in the selected sub-matrix is imputed by the median of those observed values in the same column (movie).

The probit graphical model is applied to the sub-matrix and the estimated network is illustrated in Figure 5.4. It turns out that the estimated network consists of a large connected community as well as a few isolated nodes. The large community mainly consists of the mass marketed commercial movies, especially those science fiction movies. These movies usually require high budget productions and bet for success in box-office through famous directors and stars as well as exciting visual effects. For example, the Star War series, including Star War (1977) and its two sequels Empire Strike Back (1980) and Return of the Jedi (1983), were directed or produced by George Lucas; the Terminator series (1984, 1991) were directed by James Cameron; E.T. (1982), Jurassic Park (1993) and the Indiana Jones series, including Raiders of Lost Ark (1981) and the Last Crusade (1989), were directed by Steven Spielberg. We can see that usually movies in the same series have strong connections (represented by relatively wide lines in the Figure), indicating the existence of significant dependence relationships between the ratings of these movies. Examples include the Star War series, the Alien series, the Terminator series and the Indiana Jones series. In addition, Raiders of the Lost Ark (1981) and Back to the Future (1985) are two hub nodes each having 16 connections to other movies and both of them were directed or produced by Steven Spielberg.

On the other hand, the isolated nodes represent a family of art-oriented comedies, which attract the audience by plot and intension rather than visual effects. Examples include the crime comedies (Pulp Fiction (1994), Silence of the Lambs (1991) and Fargo (1996)) and the romanic comedies (When Harry Met Sally (1989) and Princess Bride (1987)). These art comedies do not show significant dependence neither between each other nor with those commercial movies in the large community.

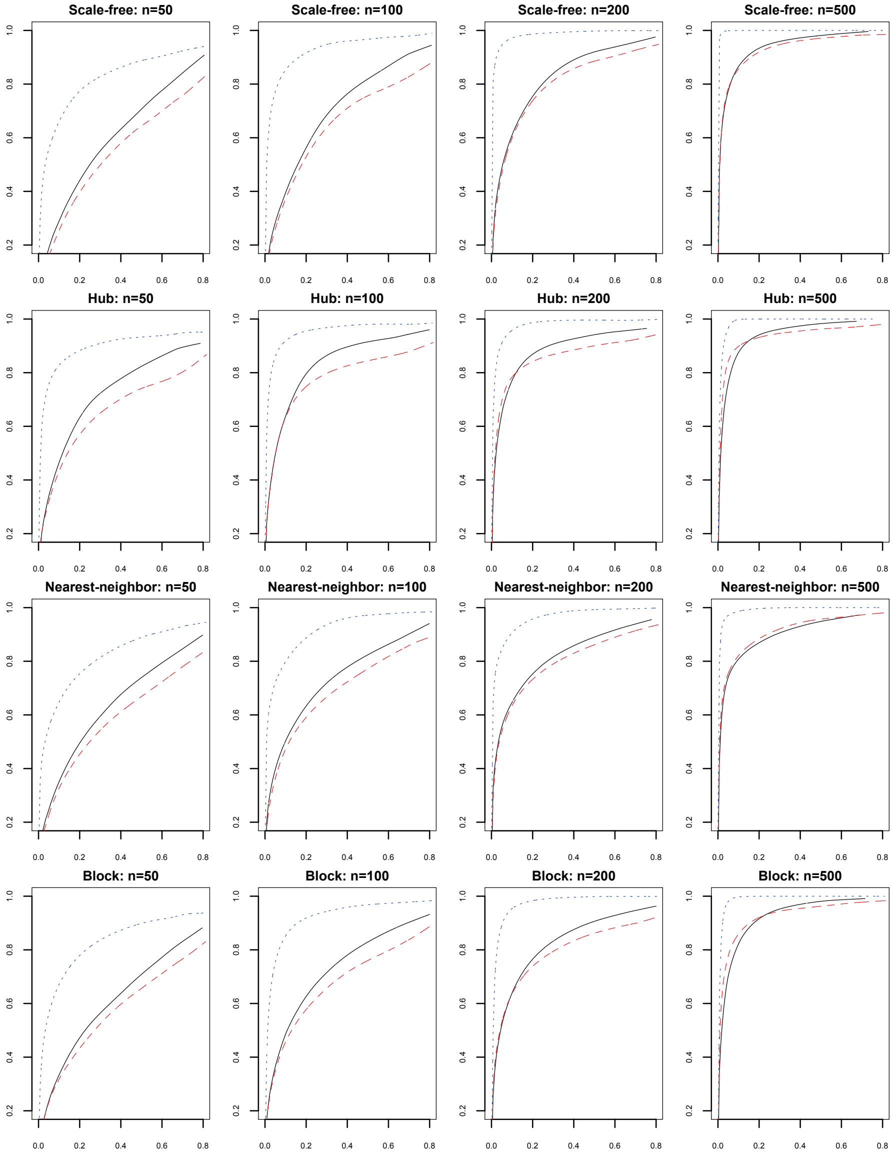


Figure 5.3: The ROC curves estimated by probit graphical model (solid dark line), the oracle model (dotted blue line) and the Gaussian model (dashed red line). The oracle model and the Gaussian model applies the graphical lasso algorithm to the latent data $\boldsymbol{Z}$ and the observed data $\boldsymbol{X}$, respectively.

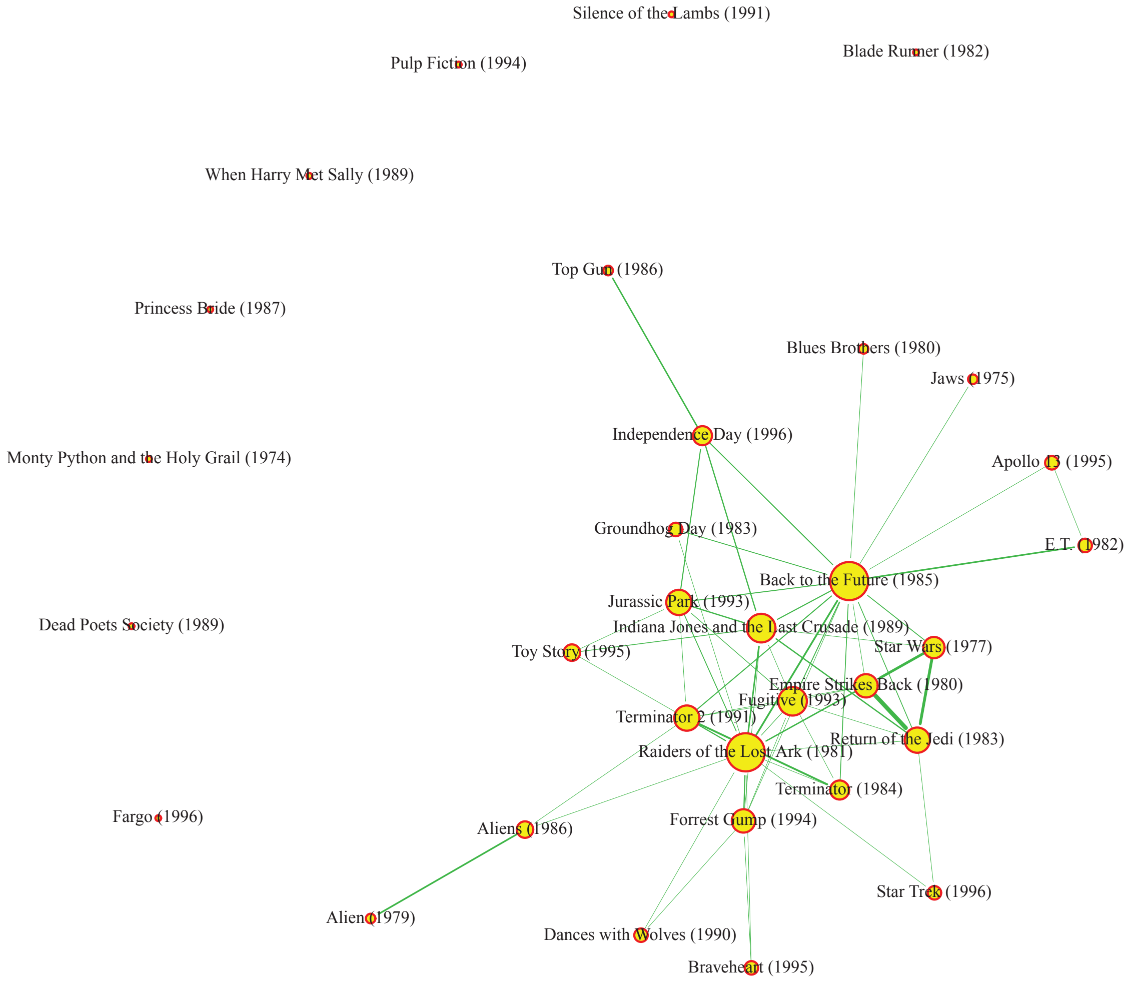


Figure 5.4: The network estimated by the probit graphical model. The nodes represent the movies labeled by their titles. The area of a node is proportional to its degree and the width of a link is proportional to the magnitude of the corresponding partial correlations.

CHAPTER VI

Pairwise Variable Selection for High-dimensional Model-based Clustering

6.1 Introduction

All the existing variable selection methods for model-based clustering choose informative variables in a “one-in-all-out” manner; that is, a variable is selected if it is informative for at least one pair of clusters and removed only if it is non-informative for all clusters. However, in many practical situations, one may be interested in identifying which variables are discriminative for which specific pairs of clusters. A toy example illustration of such a scenario is shown in Figure 6.1. There are three clusters present in this two-dimensional data set; the first variable discriminates between clusters 2 and 3, while the second variable discriminates between clusters 1 and 2. We believe that such situations arise often in high-dimensional data, for example, in data obtained from high-throughput expression technologies.

To address this problem, this paper proposes a *pairwise* variable selection method for high-dimensional model-based clustering. Specifically, a *pairwise fusion* penalty is introduced to penalize the difference between (all) pairs of cluster centers for each variable and shrink the centroids of non-separable clusters to some identical value. If all cluster centroids associated with a variable are “fused,” this variable is regarded

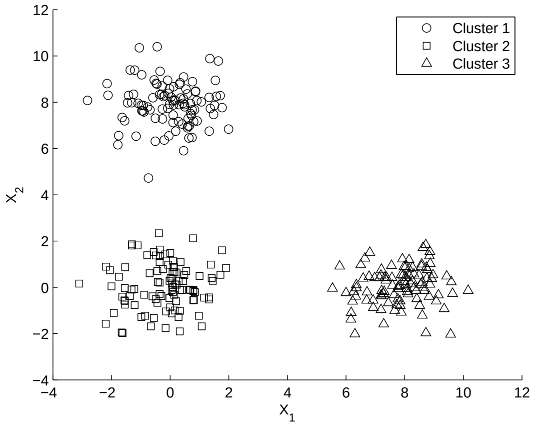


Figure 6.1: A toy example. Variable 1 is informative for separating clusters 2 and 3, and variable 2 is informative for separating clusters 1 and 2.

as non-informative and removed from the model. Otherwise, the pairwise fusion penalty has the flexibility of only fusing the centroids of non-separable clusters for this variable.

The remainder of the chapter is organized as follows: Section 6.2 introduces the pairwise fusion penalty, and Section 6.3 discusses algorithmic issues. The performance of the proposed clustering technique on synthetic and real data is demonstrated in Sections 6.4 and 6.5, respectively. Finally, some concluding remarks are drawn in Section 6.6.

6.2 Problem Formulation and Pairwise Fusion

Suppose n samples have been collected on p variables and organized in a data matrix $\boldsymbol{X} = (x_{i,j})_{n\times p}$. Without loss of generality we can assume that the data are centered for each variable, i.e., $\sum_{i=1}^{n} x_{i,j} = 0$, for all $1 \leq j \leq p$. In model-based clustering, a K-cluster problem is described by a K-component Gaussian mixture model. Specifically, the observations $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,p})$ are assumed to be independent and generated from the density

$$f(\boldsymbol{x}_i) = \sum_{k=1}^{K} w_k \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{6.1}$$

where $\phi(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ denotes the Gaussian density function with mean vector $\boldsymbol{\mu}_k = (\mu_{k,1}, \ldots, \mu_{k,p})$ and covariance matrix $\boldsymbol{\Sigma}_k$,

$$\phi(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{p/2} \text{det}(\boldsymbol{\Sigma}_k)^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^{\intercal}\right\} . \tag{6.2}$$

The "weights" w_k's ($w_k \geq 0$ for all $1 \leq k \leq K$ and $\sum_{k=1}^{K} w_k = 1$) are the mixing coefficients, capturing the contribution of the k-th cluster. We also introduce the following notation: the mean parameters $\mu_{k,j}$'s can be collected in a $K \times p$ matrix, with rows corresponding to clusters and columns to variables,

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_{1,1} & \mu_{1,2} & \cdots & \mu_{1,j} & \cdots & \mu_{1,p} \\ \mu_{2,1} & \mu_{2,2} & \cdots & \mu_{2,j} & \cdots & \mu_{2,p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mu_{K,1} & \mu_{K,2} & \cdots & \mu_{K,j} & \cdots & \mu_{K,p} \end{bmatrix} .$$

We use $\boldsymbol{\mu}_k = (\mu_{k,1}, \ldots, \mu_{k,p})$ to represent the mean parameters for the k-th cluster (k-th row vector of $\boldsymbol{\mu}$), and $\boldsymbol{\mu}_{(j)} = (\mu_{1,j}, \ldots, \mu_{K,j})^{\intercal}$ to represent the mean parameters for the j-th variable (j-th column vector of $\boldsymbol{\mu}$).

The log-likelihood of the data matrix $\boldsymbol{X}$ is then given by,

$$\log p(\boldsymbol{X}|\boldsymbol{\Theta}) = \sum_{i=1}^{n} \log \Big\{ \sum_{k=1}^{K} w_k \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \Big\}, \tag{6.3}$$

where $\boldsymbol{\Theta} = \{w_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$ is the parameter set of interest. The log-likelihood (6.3) can be maximized using an expectation-maximization (EM) algorithm, which in the E-step imputes the cluster membership of the samples and in the M-step estimates the mixing coefficients, the mean parameters and the covariance matrices. The number of clusters K can be selected using, for example, a Bayesian information criterion (BIC) or another similar criterion. Given the estimate $\widehat{\boldsymbol{\Theta}}$, an observation $\boldsymbol{x}^* = (x_1^*, \ldots, x_p^*)$ is assigned to the cluster which achieves

$$\arg \max_{1 \leq k \leq K} \widehat{w}_k \phi(\boldsymbol{x}^*; \widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}}_k). \tag{6.4}$$

6.2.1 The Pairwise Fusion Penalty

Since our focus here is on variables defined as informative in terms of differences in the cluster *means*, we make a further simplifying assumption that the covariance matrices are the same for all clusters and are diagonal, i.e., $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2)$ for all $1 \leq k \leq K$. An alternative would be to impose a shrinkage penalty on the covariance matrices as well as the means, as in Xie et al. (2008), and consider a variable non-informative for a pair of clusters only if it has both the same mean and the same covariance structure in both clusters. This does not seem to be important for the applications we have in mind, such as gene selection in expression data clustering, since the main effects are normally contained in the means. Moreover, this is a common assumption in high-dimensional settings, since it significantly reduces the number of parameters to be estimated. There is also theoretical justification for estimating the covariance matrix by a diagonal matrix for discriminant analysis in high dimensions (Bickel and Levina, 2004). In addition, imposing an additional penalty on the variances results in a dramatic increase in computational cost, and, in our experience, very small empirical gains.

Given our focus on pairwise variable selection, we propose maximizing the following criterion for estimating the parameters of the Gaussian mixture model:

$$\sum_{i=1}^{n} \log \Big\{ \sum_{k=1}^{K} w_k \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}) \Big\} - \lambda \sum_{j=1}^{p} \sum_{1 \leq k < k' \leq K} |\mu_{k,j} - \mu_{k',j}|, \tag{6.5}$$

where λ is a tuning parameter. We refer to $\sum_{j=1}^{p}\sum_{1\le k<k'\le K}|\mu_{k,j}-\mu_{k',j}|$ as the *pairwise fusion penalty* (PFP). The aim of the penalty is to shrink the difference between every pair of cluster centers for each variable j. Due to the singularity of the absolute value function, some differences are shrunken to exactly zero, resulting in some cluster means $\hat{\mu}_{k,j}$'s having identical values. Notice that we are not shrinking the means to zero, only towards each other; zero has no special meaning here and the data do not need to be centered. If $\hat{\mu}_{k,j}=\hat{\mu}_{k',j}$, then variable j is considered to be "non-informative" for separating cluster k and cluster k', though it may be informative for separating other clusters. Moreover, if all cluster means for a variable are shrunken to the same value, that variable is considered non-informative for clustering purposes and can be removed from the model.

6.2.2 The Adaptive Pairwise Fusion Penalty

To further improve on (6.5), we apply the popular adaptive penalization (Zou, 2006) by considering

$$\sum_{i=1}^{n}\log\Big\{\sum_{k=1}^{K}w_k\phi(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\Sigma})\Big\}-\lambda\sum_{j=1}^{p}\sum_{1\le k<k'\le K}\tau_{k,k'}^{(j)}|\mu_{k,j}-\mu_{k',j}|, \qquad (6.6)$$

where $\tau_{k,k'}^{(j)}$ are pre-specified weights. We call this version adaptive pairwise fusion penalty (APFP). The intuition is that if variable j is informative for separating clusters k and k', we would like the corresponding $\tau_{k,k'}^{(j)}$ to be small; thus, the difference between $\mu_{k,j}$ and $\mu_{k',j}$ is lightly penalized. On the other hand, for a non-informative variable j for clusters k and k', we would like the corresponding $\tau_{k,k'}^{(j)}$ to be large and hence the difference between $\mu_{k,j}$ and $\mu_{k',j}$ is heavily penalized. In our implementation, we compute the weights from the unpenalized estimates as

$$\tau_{k,k'}^{(j)}=|\widetilde{\mu}_{k,j}-\widetilde{\mu}_{k',j}|^{-1}\ ,$$

where $\widetilde{\mu}_{k,j}$ is the estimate of $\mu_{k,j}$ without any penalization ($\lambda = 0$).

It is interesting to compare our approach to the ℓ_1-regularized method proposed by Pan and Shen (2006) and the ℓ_∞-regularized method proposed by Wang and Zhu (2007). Note that Pan and Shen (2006) proposed an ℓ_1 penalty without adaptive weights, but for a fair comparison here we use adaptive versions of all the methods. Pan and Shen (2006) proposed to use the criterion,

$$\sum_{i=1}^{n} \log\Big\{\sum_{k=1}^{K} w_k \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma})\Big\} - \lambda \sum_{j=1}^{p}\sum_{k=1}^{K} \tau_{k,j}^{\ell_1} |\mu_{k,j}|, \tag{6.7}$$

where $\tau_{k,j}^{\ell_1}$'s are adaptive weights defined as $\tau_{k,j}^{\ell_1} = 1/|\widetilde{\mu}_{k,j}|$ for all $1 \leq k \leq K$ and $1 \leq j \leq p$. Here $\widetilde{\mu}_{k,j}$ is the estimate from model-based clustering method without penalty. Notice that the data are required to be centered, and the ℓ_1 penalty shrinks the individual $\mu_{k,j}$'s towards zero (the global mean) and removes variable j from the model if all $\widehat{\mu}_{k,j}$ for $1 \leq k \leq K$ are set to zero. However, it cannot identify variables that are non-informative for separating particular subsets of clusters, especially when the common mean of these clusters is different from zero. On the other hand, the ℓ_∞-regularized criterion proposed by Wang and Zhu (2007) is

$$\sum_{i=1}^{n} \log\Big\{\sum_{k=1}^{K} w_k \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma})\Big\} - \lambda \sum_{j=1}^{p} \tau_j^{\ell_\infty} \max(|\mu_{1,j}|, \ldots, |\mu_{k,j}|, \ldots, |\mu_{K,j}|), \tag{6.8}$$

where the adaptive weight $\tau_j^{\ell_\infty} = 1/\max(|\widetilde{\mu}_{1,j}|, \ldots, |\widetilde{\mu}_{k,j}|, \ldots, |\widetilde{\mu}_{K,j}|)$. Unlike the ℓ_1 penalty which shrinks each $\mu_{k,j}$ individually, the ℓ_∞ norm penalizes the maximum magnitude of the cluster means for each variable. If the largest cluster mean for variable j is shrunk to zero, then all other means for the j-th variable are automatically zero, and the variable can be eliminated from the model. However, this penalty is also unable to identify specific clusters that can be separated by a particular variable.

6.2.3 Model Selection

There are two parameters to be selected, the number of clusters K and the tuning parameter λ. We select them using a BIC-type criterion, defined by

$$\mathrm{BIC}(K, \lambda) = -2 \sum_{i=1}^{n} \log \Big\{ \sum_{k=1}^{K} \widehat{w}_k \phi(\boldsymbol{x}_i; \widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}}) \Big\} + d \log n, \tag{6.9}$$

where $\{\widehat{w}_k, \widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}}\}_{k=1}^{K}$ are estimated with K clusters and the tuning parameter λ. The degrees of freedom d are defined as the number of distinct nonzero estimates. Specifically, $d = K - 1 + p + e(\widehat{\boldsymbol{\mu}})$, where $e(\widehat{\boldsymbol{\mu}})$ is the number of distinct nonzero elements in $\{\widehat{\mu}_{k,j}\}$. This definition is similar to the degrees of freedom for fused Lasso (Tibshirani et al., 2005).

6.3 The Optimization Algorithm

The optimization of the objective function (6.6) is non-trivial. As in classical model-based clustering, we employ an EM algorithm to maximize the log-likelihood function subject to the penalty constraint. Let $\Delta_{i,k}$ be the indicator of whether $\boldsymbol{x}_i$ is from cluster k, that is, $\Delta_{i,k} = 1$ if $\boldsymbol{x}_i$ belongs to cluster k, and $\Delta_{i,k} = 0$ otherwise. If the missing data $\Delta_{i,k}$ were observed, the penalized log-likelihood function for the complete data is given by

$$\sum_{i=1}^{n} \sum_{k=1}^{K} \Delta_{i,k} \{\log w_k + \log \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma})\} - \lambda \sum_{j=1}^{p} \sum_{1 \le k < k' \le K} \tau_{k,k'}^{(j)} |\mu_{k,j} - \mu_{k',j}|. \tag{6.10}$$

Our algorithm follows closely the EM algorithm for the standard (unpenalized) Gaussian mixture model (McLachlan and Peel, 2002); the main difference is in estimating $\mu_{k,j}$ in the M-step. The EM algorithm iterates between two alternating steps and produces a sequence of estimates $\widehat{\boldsymbol{\Theta}}^{(t)}$, $t = 0, 1, 2, \ldots$ We start with the E-step given the current parameter estimates $\widehat{\boldsymbol{\Theta}}^{(t)}$.

E-step

In this step, we impute values for the unobserved $\Delta_{i,k}$ by

$$\widehat{\Delta}_{i,k}^{(t+1)} = \mathrm{E}(\Delta_{i,k}|\boldsymbol{X}, \widehat{\boldsymbol{\Theta}}^{(t)}) = \Pr(\Delta_{i,k} = 1|\boldsymbol{X}, \widehat{\boldsymbol{\Theta}}^{(t)}) = \frac{\widehat{w}_k^{(t)} \phi(\boldsymbol{x}_i; \widehat{\boldsymbol{\mu}}_k^{(t)}, \widehat{\boldsymbol{\Sigma}}^{(t)})}{\sum_{k'=1}^{K} \widehat{w}_{k'}^{(t)} \phi(\boldsymbol{x}_i; \widehat{\boldsymbol{\mu}}_{k'}^{(t)}, \widehat{\boldsymbol{\Sigma}}^{(t)})}. \tag{6.11}$$

Plugging them into (6.10), we obtain the so-called penalized Q-function:

$$Q_P(\boldsymbol{\Theta}, \widehat{\boldsymbol{\Theta}}^{(t)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \widehat{\Delta}_{i,k}^{(t+1)} \{\log w_k + \log \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma})\} - \lambda \sum_{j=1}^{p} \sum_{1 \le k < k' \le K} \tau_{k,k'}^{(j)} |\mu_{k,j} - \mu_{k',j}|.$$

M-step

The goal is to update the parameter estimates via

$$\widehat{\boldsymbol{\Theta}}^{(t+1)} = \arg\max_{\boldsymbol{\Theta}} Q_p(\boldsymbol{\Theta}, \widehat{\boldsymbol{\Theta}}^{(t)}). \tag{6.12}$$

Specifically,

$$\frac{\partial Q_p}{\partial w_k} = 0 \;\Rightarrow\; \widehat{w}_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\Delta}_{i,k}^{(t+1)} \tag{6.13}$$

$$\frac{\partial Q_p}{\partial \sigma_j^2} = 0 \;\Rightarrow\; (\widehat{\sigma}_j^{(t+1)})^2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \widehat{\Delta}_{i,k}^{(t+1)} (x_{i,j} - \widehat{\mu}_{k,j}^{(t)})^2, \; 1 \le j \le p, \tag{6.14}$$

and

$$\widehat{\boldsymbol{\mu}}^{(t+1)} = \arg\min_{\boldsymbol{\mu}} \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} \left\{ \widehat{\Delta}_{i,k}^{(t+1)} \sum_{j=1}^{p} \frac{(x_{i,j} - \mu_{k,j})^2}{(\widehat{\sigma}_j^{(t)})^2} \right\} + \lambda \sum_{j=1}^{p} \sum_{1 \le k < k' \le K} \tau_{k,k'}^{(j)} |\mu_{k,j} - \mu_{k',j}| \tag{6.15}$$

The optimization of (6.15) is nontrivial and is discussed in detail next.

Estimation of the cluster means

In general, objective function (6.15) can be transformed into a quadratic programming problem, and solved by a commercially available package. This approach, however, can be inefficient in practice, especially for a large number of variables p. Thus, we propose a more efficient iterative algorithm based on the standard local quadratic approximation (Fan and Li, 2001). Local quadratic approximation has been used in a number of variable selection procedures and its convergence properties have been studied by Fan and Li (2001) and Hunter and Li (2005). Specifically, we approximate

$$|\mu_{k,j}^{(s+1)} - \mu_{k',j}^{(s+1)}| \approx \frac{(\mu_{k,j}^{(s+1)} - \mu_{k',j}^{(s+1)})^2}{2|\widehat{\mu}_{k,j}^{(s)} - \widehat{\mu}_{k',j}^{(s)}|} + \frac{1}{2}|\widehat{\mu}_{k,j}^{(s)} - \widehat{\mu}_{k',j}^{(s)}| \; , \tag{6.16}$$

where s is the iteration index (different from t, which is used to denote different iterations of the EM algorithm, whereas s is used to denote iterations of the local quadratic approximation within the M-step), and $\widehat{\mu}^{(s)}$ are the estimates from the previous iteration. This approximation converts the minimization in (6.15) into a generalized ridge (quadratic) problem, which can be solved in closed form. For example, for each j (notice that (6.15) can be decomposed into p separate minimization problems), we solve (iteratively over s)

$$\min_{\boldsymbol{\mu}_{(j)}^{(s+1)}} \frac{1}{2(\widehat{\sigma}_j^{(t)})^2} \sum_{i=1}^{n} \sum_{k=1}^{K} \widehat{\Delta}_{i,k}^{(t+1)} (x_{i,j} - \mu_{k,j}^{(s+1)})^2 + \lambda \sum_{1 \le k < k' \le K} \tau_{k,k'}^{(j)} \frac{(\mu_{k,j}^{(s+1)} - \mu_{k',j}^{(s+1)})^2}{2|\widehat{\mu}_{k,j}^{(s)} - \widehat{\mu}_{k',j}^{(s)}|} \; . \tag{6.17}$$

For numerical stability, we threshold the absolute value of $\widehat{\mu}_{k,j}^{(s)} - \widehat{\mu}_{k',j}^{(s)}$ at a lower bound of 10^{-10}, and at the end of the iterations, set all estimates equal to 10^{-10} to zero.

We note that the M-step of maximizing the penalized Q-function does not have closed form solutions, and its maximizer is obtained iteratively. Therefore, strictly speaking, our algorithm is an expectation-conditional maximization (ECM) algorithm (Meng and Rubin, 1993), which replaces the M-step of EM by a sequence of conditional maximization steps, each maximizing the penalized Q-function over $\mathbf{\Theta}$, but with some of its elements fixed at their previous values. By Theorem 3 in Meng and Rubin (1993), our algorithm is guaranteed to converge to a stationary point.

6.4 Numerical Results

In this section, we illustrate the performance of the proposed pairwise variable selection method on three synthetic examples with four clusters for Simulations 1 and 3 and five clusters for Simulation 2. We compare four methods: Gaussian mixture model-based clustering without a penalty, the adaptive ℓ_1 penalty (6.7), the adaptive ℓ_∞ (6.8) and our proposed adaptive pairwise fusion penalty (6.6). We refer to them as "GMM", "AL1", "ALP" and "APFP" respectively. The non-adaptive PFP method was also applied and is generally dominated by APFP; its results are omitted for space considerations. In Simulations 1 and 2, the same number of observations, i.e., 20, are generated from each cluster, while in Simulation 3, we generate different number of observations for different clusters. The number of clusters K and the tuning parameter λ are selected using the BIC criterion, as described in Section 6.2.3. For benchmarking purposes, we also calculate the solution by specifying the true number of clusters, namely $K = 4$ for Simulations 1 and 3 and $K = 5$ for Simulation 2, and only select λ using BIC. We repeat this 50 times for each simulation and record the average clustering error rates as compared to the true cluster labels, and average selection rate for both informative and non-informative variables. To compute the clustering error rates, the predicted class labels are calculated by a majority vote, i.e., if most data points in a particular predicted cluster belong to a true cluster k ($1 \leq k \leq K$), then all data points in this predicted cluster are labeled as k.

The performance of the EM algorithm in model-based clustering depends on the choice of the initial values for the parameters since the likelihood function is not convex, and the algorithm can only converge to a local maximum. To get a good starting value, we first fit 100 GMMs (without penalty) with different random initial values, and use the estimate with the highest likelihood as a starting value for the EM algorithm. In our simulations, the EM algorithm usually converged after about 100 iterations.

Table 6.1: Means of informative variables in Simulations 1–3.

Simulation	Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
1 & 3	1–10	2.5	0	0	-2.5	—
	11–20	1.5	1.5	-1.5	-1.5	—
2	1–10	2.5	2.5	0	0	-2.5
	11–20	-2.5	0	0	0	2.5
	21–30	2.5	0	0	-2.5	-2.5

Simulation 1

In this scenario, there are four clusters and $p = 220$, with the first 20 being informative and the remaining ones non-informative. The variables were generated according to the following mechanism: the first 20 are independently distributed $N(\mu_{k,j}, \sigma^2)$ for cluster k, whereas the remaining 200 variables are all i.i.d. $N(0, 1)$ for all four clusters. Table 6.1 gives the means for the first 20 variables. For example, in cluster 1, variables 1–10 all have the same mean value 2.5, and variables 11–20 all have the same mean value 1.5. Figure 6.2 (left panel) illustrates the distribution of the informative variables. Notice that variables 1–10 are non-informative for separating clusters 2 and 3, while variables 11–20 are non-informative for separating clusters 1 and 2 (as well as clusters 3 and 4). We consider two values of the common variance, $\sigma^2 = 1$ and $\sigma^2 = 4$. The former creates a high "signal-to-noise ratio (SNR)" scenario, while the latter simulates a situation where the "signal-to-noise ratio" is low.

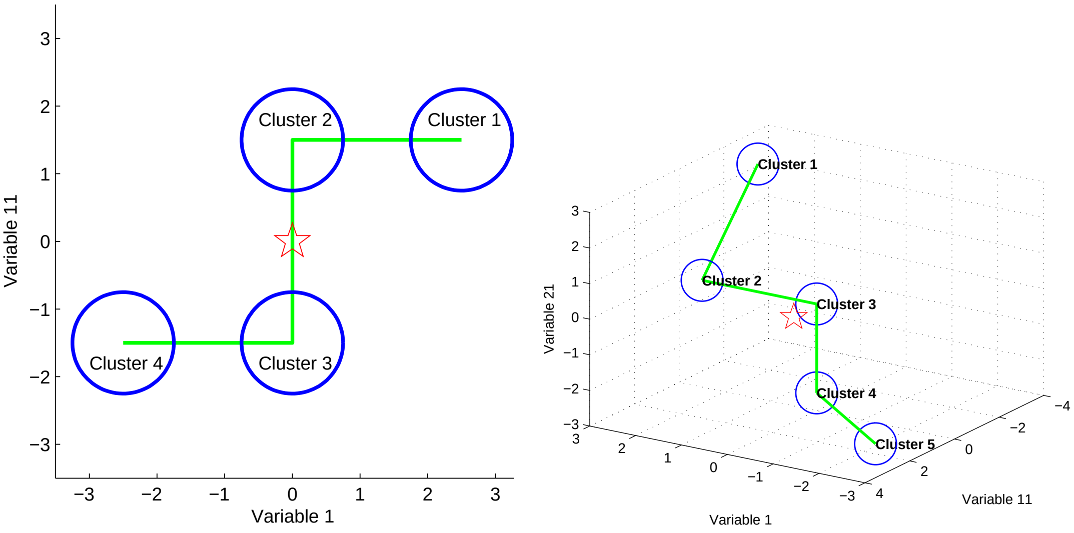


Figure 6.2: The distribution of informative variables in Simulation 1 (left) and Simulation 2 (right). The red star indicates the position of the overall sample mean.

Simulation 2

A five cluster scenario is considered. There are a total of $p = 230$ variables with the first 30 informative and the other 200 non-informative. Similarly to Simulation 1, the informative variables are independently distributed as $N(\mu_{k,j}, \sigma^2)$ for cluster k, whereas the remaining 200 variables are all i.i.d. $N(0, 1)$ for all five clusters. Table 6.1 gives the mean values for the informative variables, and Figure 6.2 (right panel) illustrates the distribution of the informative variables. Notice that variables 1–10 are non-informative for separating clusters 1 and 2, as well as clusters 3 and 4; variables 11–20 are non-informative for separating clusters 2, 3 and 4; and variables 21–30 are non-informative for separating clusters 2 and 3, as well as clusters 4 and 5. We, again, consider $\sigma^2 = 1$ (high signal-to-noise ratio) and $\sigma^2 = 4$ (low signal-to-noise ratio).

Simulation 3

This simulation is designed to test the proposed method on unbalanced data, i.e., data where clusters have different sample sizes. All the settings in this simulation are the same as in Simulation 1 (high SNR), except that the sample size for clusters 3 and 4 has been increased to 200. Therefore, there are two small clusters (1 and 2) with 20 observations each and two large clusters (3 and 4) with 200 observations each.

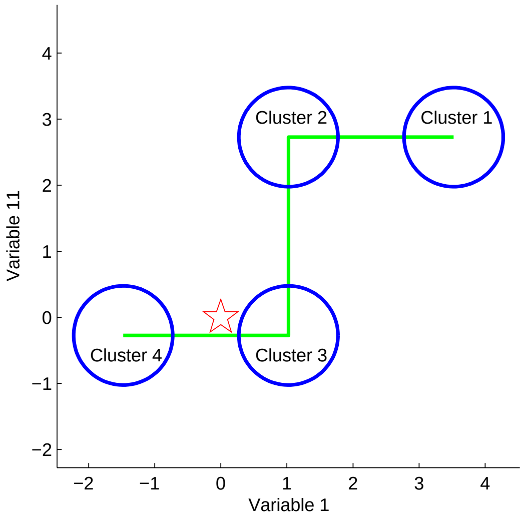


Figure 6.3: Simulation 3. The sample sizes of clusters 1, 2, 3 and 4 are 20, 20, 200, and 200, respectively. The red star indicates the position of the overall sample mean, and the plot is shifted to show centered data.

The results over 50 replications for all simulation scenarios are summarized in Table 6.2. When the signal-to-noise ratio in Simulations 1 and 2 is high, all four methods select the correct number of clusters and the error rates are very close to zero. On the other hand, in the low signal-to-noise ratio setting, GMM and ALP completely fail to select the correct number of clusters, and have a high error rate. The performance of the AL1 and APFP methods also degrade, but both are still able to select the correct number of clusters most of the time. Further, the error rate of

Table 6.2: Prediction and variable selection results for Simulations 1–3. Each table cell gives average(SD) over 50 repetitions. "K" is the average number of selected clusters, "ER" is the average clustering error rate, "ER (correct K)" is the average error rate when K is set to the true value rather than selected by BIC, "Info" is the average proportion of selected informative variables, and "Noninfo" is the average proportion of selected non-informative variables. "High SNR" corresponds to $\sigma^2 = 1$, and "Low SNR" corresponds to $\sigma^2 = 4$.

Sim. (SNR)	Method	K	ER (%)	ER (correct K)	Info (%)	Noninfo (%)
1 (High)	GMM	3 (0)	25 (0)	0 (0)	100 (100)	100 (100)
	AL1	4 (0)	0 (0)	0 (0)	100 (100)	7.1 (7.1)
	ALP	4 (0)	0 (0)	0 (0)	100 (100)	2.4 (2.4)
	APFP	4 (0)	0 (0)	0 (0)	100 (100)	0.5 (0.5)
1 (Low)	GMM	3 (0)	33 (4.9)	20.6 (8.5)	100 (100)	100 (100)
	AL1	3.8 (0.6)	19.2 (14.9)	14.2 (10.7)	100 (100)	6 (6)
	ALP	3 (0)	34.1 (14.5)	14.4 (14)	95.9 (95.9)	4 (4)
	APFP	3.7 (0.6)	19.2 (16.7)	15.1 (12.6)	100 (100)	2.3 (2.3)
2 (High)	GMM	3 (0)	40 (0)	0 (0.2)	100 (100)	100 (100)
	AL1	5 (0)	0 (0)	0 (0)	100 (100)	6.9 (6.9)
	ALP	5 (0)	0 (0.1)	0 (0.1)	100 (100)	1.8 (1.8)
	APFP	5 (0)	0 (0)	0 (0)	100 (100)	1.1 (1.1)
2 (Low)	GMM	3 (0)	40.3 (0.7)	15.3 (5.3)	100 (100)	100 (100)
	AL1	4.7 (0.6)	11.7 (9.8)	8.3 (5.3)	100 (100)	10 (10)
	ALP	3 (0)	40.1 (0.4)	5.8 (3)	100 (100)	5.2 (5.2)
	APFP	4.7 (0.5)	11.7 (7.7)	9.2 (5.5)	100 (100)	2.4 (2.4)
3	GMM	3 (0)	4.5 (0)	0 (0)	100 (100)	100 (100)
	AL1	4 (0)	0 (0)	0 (0)	100 (100)	8.1 (8.1)
	ALP	3.9 (0.2)	0.3 (1.1)	0 (0)	100 (100)	5.9 (5.9)
	APFP	4 (0.1)	0 (0)	0 (0)	100 (100)	0.2 (0.2)

the APFP method is comparable with that of the AL1 method. In terms of variable selection, AL1, ALP and APFP are able to identify the informative variables, but APFP is more effective than ALP and AL1 at removing non-informative variables. The results for Simulation 3 are very similar to those of Simulation 1 with high SNR, which shows that unbalanced data do not affect performance of any of the methods.

If a variable is non-informative for separating a pair of clusters, and the corresponding estimated means are also the same, we consider this correct "fusion". Table 6.3 summarizes these results. Specifically, each row in the table gives the proportion of correctly fused variables (average over 50 replications) out of the ten that are non-informative for separating the corresponding pair of clusters (indicated in the third column). For example, the first row shows that for the APFP method, on average 91.6% of the variables among the first ten are correctly fused for clusters 2 and 3. It is also clear that APFP dominates both AL1 and ALP in terms of correctly fusing the cluster means. Although AL1 and ALP can correctly fuse some cluster means (e.g., in the first and second row), these results are artifacts. For example, in Simulation 1, the means of clusters 2 and 3 for variables 1–10 are all equal to zero, which happens to be the value that the ℓ_1 penalty shrinks to. The same reasoning applies to clusters 2, 3 and 4 for variables 11-20 in Simulation 2. On the other hand, in Simulation 1, although clusters 1 and 2 (as well as clusters 3 and 4) have the same mean value for variables 11–20, the AL1 method fails to fuse them, since their mean value is different from zero. The ALP method only shrinks the cluster mean with the largest magnitude, such as the means of clusters 1 and 2 and cluster 3 and 4 for variables 11–20 in Simulation 1. We can also see that both AL1 and ALP are unable to perform pairwise variable selection for unbalanced clusters in Simulation 3. In contrast to Simulation 1, the overall sample mean in Simulation 3 (red star in Figure 6.3) does not lie at the centroid of the four cluster means. This explains why AL1 fails to identify non-separable clusters 2 and 3 for variables 11–20 and ALP fails to identify non-separable clusters 3 and 4, which they were able to identify in Simulation 1. The APFP method identifies the correct structure in all these scenarios.

6.5 Applications to Gene Expression Data

In this section, we apply the pairwise fusion method to two gene microarray data sets. To illustrate the method, we pre-select a subset of genes from each data by ranking the genes according to their variance and only using the top 100 and bottom 100 genes. We anticipate that high variance genes are more informative than low variance genes for clustering purposes, although, as the results below show, this is not always true. Notice that selection does not use any class label information. The obtained 200 variables (genes) are centered before clustering.

6.5.1 The SRBCT Data

This data set contains the expression profiles of 2308 genes, obtained from 83 tissue samples of small round blue cell tumors (SRBCT) of childhood cancer (Khan et al., 2001). The 83 samples are classified into four tumor subtypes: Ewing's sarcoma (EWS), rhabdomyosarcoma (RMS), neuroblastoma (NB), and Burkitt's lymphoma (BL).

The results in Table 6.4 (SRBCT) show that all these methods select six clusters via BIC and produce the same error rate of 1.4%. Table 6.5 shows the confusion matrix for the APFP method. Each row corresponds to a tumor subtype, and each column to an identified cluster. It can be seen that subtype EWS is split into clusters 2 and 6, and subtype RMS into clusters 1 and 3. This result suggests possible existence of heterogeneous structures within these two subtypes.

From Table 6.4, we can also see that both GMM and ALP select all 200 genes, while APFP selects 92 from the top 100 genes and 66 from the bottom 100 genes, and AL1 selects all top 100 genes and 88 from the bottom 100 genes. This is a somewhat unexpected result. To further investigate this issue, two F-statistics and their p-values were computed for each gene; the first one compares the four tumor subtypes, while the second one the six identified clusters. The results are shown in Figure 6.4. Notice that although genes with a large variance tend to be informative (since they tend to have small p-values as shown in the left panels of Figure 6.4), genes with a small variance are not necessarily non-informative for clustering. The right panels in Figure 6.4 show that among the bottom 100 genes by variance there is a number of genes with relatively small p-values, both for discriminating the true subtypes and the found clusters. These turn out to be the genes that are selected by the APFP method from the bottom 100 genes. Further, the left panels in Figure 6.4 show that some of top 100 genes have large p-values. Indeed, the four genes that have the largest p-values are not selected by APFP. Overall, Figure 6.4 provides insight into why 66 genes are selected by the APFP method from the bottom 100 group, and why some of the genes in the top 100 group are not selected. The selection of all the genes by the L1 method is obviously not satisfactory.

Figure 6.5 shows the results for pairwise fusion. The rows correspond to the 92 (out of top 100) genes selected by the APFP method and the column to pairs of clusters. There are a total of 15 pairs formed from the six identified clusters. A black (white) spot indicates that the estimated means of the corresponding gene for the two clusters are different (the same). For example, the gene with ID “435953” is non-informative for separating clusters 1 and 3, as well as clusters 2 and 5, and clusters 4 and 6. It can be seen that most genes are informative for only a subset of clusters. Compared to the “one-in-all-out” approach, this result is more informative for describing the functions of a gene with respect to discriminating different tumor subtypes.

6.5.2 PALL Data Set

This data set contains gene expression profiles for 12,625 genes from 248 patients (samples) with pediatric acute lymphoblastic leukemia (PALL), see Yeoh et al. (2002) for more details. The samples are classified into six tumor subtypes: T-ALL (43 cases), E2A-PBX1 (27 cases), TEL-AML (79 cases), hyperdiploid>50 (64 cases), BCR-ABL (15 cases) and MLL (20 cases). The original data had a large number of missing intensities and the following pre-processing was applied. All intensity values less than one were set to one; then all intensities were transformed to log-scale. Further, all genes with log-intensities equal to zero for more than 80% of the samples were discarded, thus leaving 12,083 genes for further consideration. From the preprocessed data, the top and bottom 100 genes were selected according to the overall variance criterion described above. All variables were centered.

From Table 6.4 (PALL), we can see that GMM, AL1 and APFP methods select 12, 7 and 9 clusters, respectively, and produce comparable error rates (25%∼27%), all of which are significantly lower than that of ALP (41.1%). Table 6.6 shows the confusion matrix for the APFP method. Unlike the results on the SRBCT data, the clusters discovered by APFP are generally not consistent with the six subtypes. However, subtypes E2A-PBX1 and T-ALL are largely captured by clusters 3 and 7, most samples in subtype hyperdiploid>50 are assigned to clusters 4 and 6, while TEL-AML is split amongst clusters 1, 2 and 9. This result suggests the possible presence of a more complex structure in some of the subtypes.

Figure 6.6 shows the scatter plot of variance vs p-values obtained from the two F-statistics as described above. Once again, genes with a large variance do not necessarily correspond to small p-values, and vice versa. Figure 6.7 provides a detailed illustration of the gene functions with respect to discriminating different tumor subtypes.

6.6 Conclusions

We have developed a method for simultaneously clustering high-dimensional data and selecting informative variables, by employing a penalized model-based clustering framework. In particular, the proposed method penalizes the difference between the cluster means for each pair of clusters and for each variable, which allows one to identify and remove non-informative variables for selected subsets of clusters. This allows to gain more insight into the function of particular variables and potentially discover heterogeneous structures that other available methods are unable to capture. Our numerical work suggests that this penalty proves more effective in removing non-informative variables than an ℓ_1 penalty method, and provides better interpretation. Possible extensions include allowing for different variances and fusing variances as well as the means, as discussed at the start of Section 6.2.1, as well as extensions to non-Gaussian data. Applications to problems other than clustering are another possibility; a similar penalty for simultaneously selecting factors and collapsing levels in ANOVA was proposed by Bondell and Reich (2009) while this paper was under review.

Table 6.3: Pairwise variable selection results for Simulations 1–3. "Pair" corresponds to non-separable cluster pairs for the variables in the corresponding row. For example, the first row indicates that variables 1–10 are non-informative for separating clusters 2 and 3. The numbers in the following columns show what proportion of variables of the set are identified as non-informative for separating a given pair of clusters by each method. The optimal value is 10 in each case. All results are averages (SDs) over 50 repetitions.

Sim. (SNR)	Variables	Pair	AL1(%)	ALP(%)	APFP(%)
1 (High)	1–10	2/3	96.6 (5.2)	0.2 (1.4)	91.6 (9.1)
	11–20	1/2	0.2 (1.4)	40.8 (18.9)	91.8 (8.5)
		3/4	0 (0)	42.2 (21.4)	92.2 (7.9)
1 (Low)	1–10	2/3	95.6 (9.3)	6 (21.4)	79.8 (17.6)
	11–20	1/2	1 (3.0)	85 (16.2)	78.2 (21.2)
		3/4	0.4 (2.0)	79.6 (14.1)	84 (13.4)
2 (High)	1–10	1/2	0.2 (1.41)	0.2 (1.41)	84.2 (12.3)
		3/4	34.6 (28.1)	0.4 (2.0)	87.4 (9.7)
	11–20	2/3	98 (5.0)	0.2 (1.4)	94 (8.1)
		2/4	97.6 (4.8)	0.4 (2.0)	93.4 (8.2)
		3/4	97.2 (4.5)	0.2 (1.4)	93.2 (8.9)
	21–30	2/3	30.2 (30.1)	0.4 (2.0)	83.8 (12.1)
		4/5	0 (0)	0 (0)	88.2 (10.6)
2 (Low)	1–10	1/2	0.2 (1.41)	17 (10.9)	72.4 (17)
		3/4	73 (14.7)	0 (0)	74.4 (18.5)
	11–20	2/3	94.8 (6.46)	0 (0)	89.2 (11.2)
		2/4	95.4 (5.4)	0 (0)	89.4 (9.8)
		3/4	95.4 (6.1)	0 (0)	89 (10.2)
	21–30	2/3	76.8 (14.9)	0 (0)	67.8 (21.8)
		4/5	0 (0)	21.2 (13.8)	74.4 (16.8)
3	1–10	2/3	0.2 (1.4)	0.4 (2.0)	94.6 (6.8)
	11–20	1/2	0.2 (1.4)	60.8 (14.7)	92.6 (6.6)
		3/4	0 (0)	0 (0)	96.8 (6.2)

Table 6.4: Clustering results for the SRBCT and PALL data sets. “Top 100” and “Bottom 100” correspond to the number of genes that are selected from the top 100 and bottom 100 genes respectively, as ranked by overall variance.

Data	Method	K	Error rate (%)	Top 100 (%)	Bottom 100 (%)
SRBCT	GMM	6	1.4	100	100
	AL1	6	1.4	100	88
	ALP	6	1.4	100	100
	APFP	6	1.4	92	66
PALL	GMM	12	25.7	100	100
	AL1	7	24.7	94	100
	ALP	5	41.1	100	100
	APFP	9	27.0	89	99

Table 6.5: Confusion matrix of the APFP method for the SRBCT data. Rows correspond to tumor subtypes, and columns to identified clusters.

Subtype	C1	C2	C3	C4	C5	C6
EWS	0	18	0	0	0	11
RMS	16	0	9	0	0	0
NB	1	0	0	0	17	0
BL	0	0	0	11	0	0

Table 6.6: Confusion matrix of the APFP method for the PALL data. Rows correspond to tumor subtypes, and columns to identified clusters.

Subtype	C1	C2	C3	C4	C5	C6	C7	C8	C9
BCR-ABL	0	0	0	2	6	7	0	0	0
E2A-PBX1	0	0	25	0	0	1	0	1	0
hyperdiploid>50	1	1	0	35	0	24	0	2	1
MLL	1	0	2	0	13	0	0	4	0
TEL-AML	30	18	0	0	0	0	0	0	31
T-ALL	0	0	0	0	5	0	33	5	0

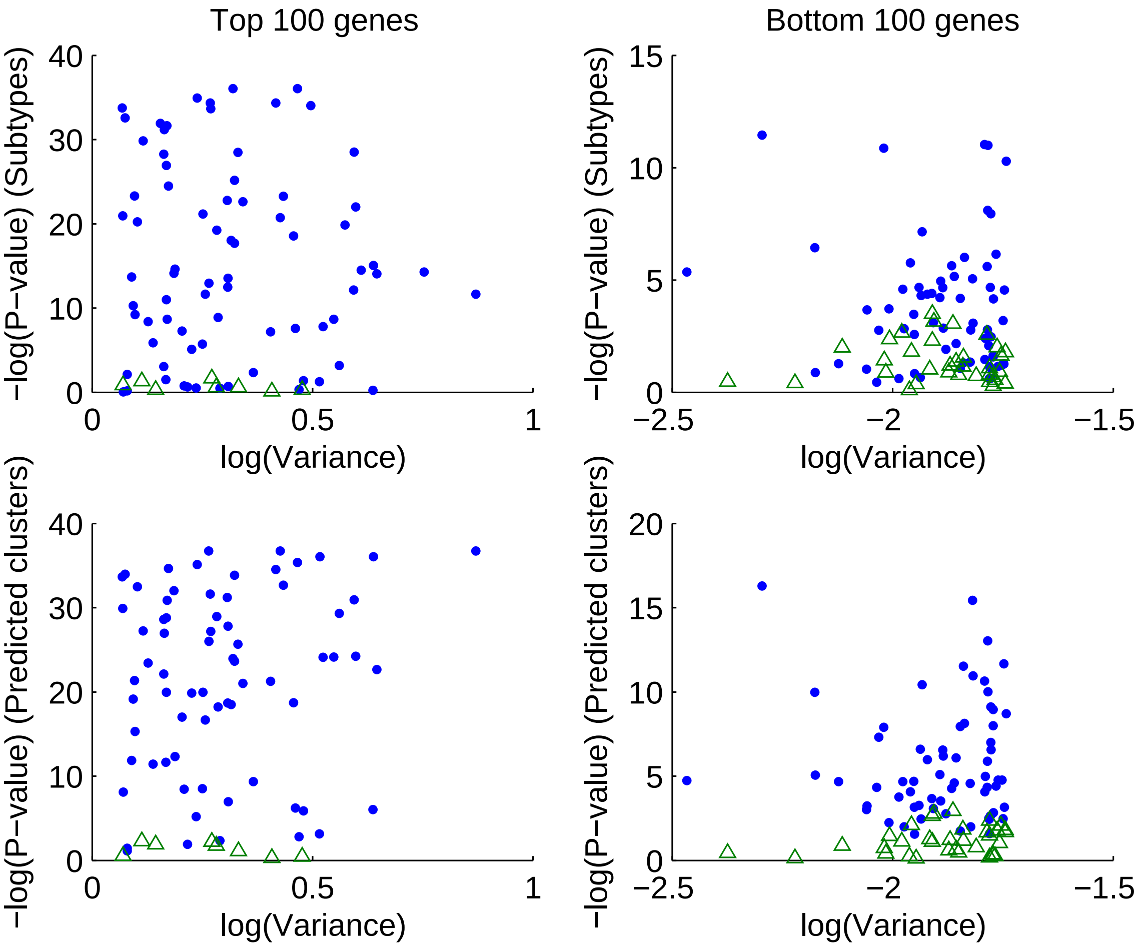


Figure 6.4: Plots of the negative logarithm p-values vs variance for SRBCT data. The left column is the top 100 genes (largest overall variances), and the right column is the bottom 100 genes. The upper row is negative logarithm p-values corresponding to an F-statistics comparing four tumor subtypes, and the lower row is the negative logarithm p-values for the six identified clusters. Triangles denote the genes that are not selected by the APFP method.

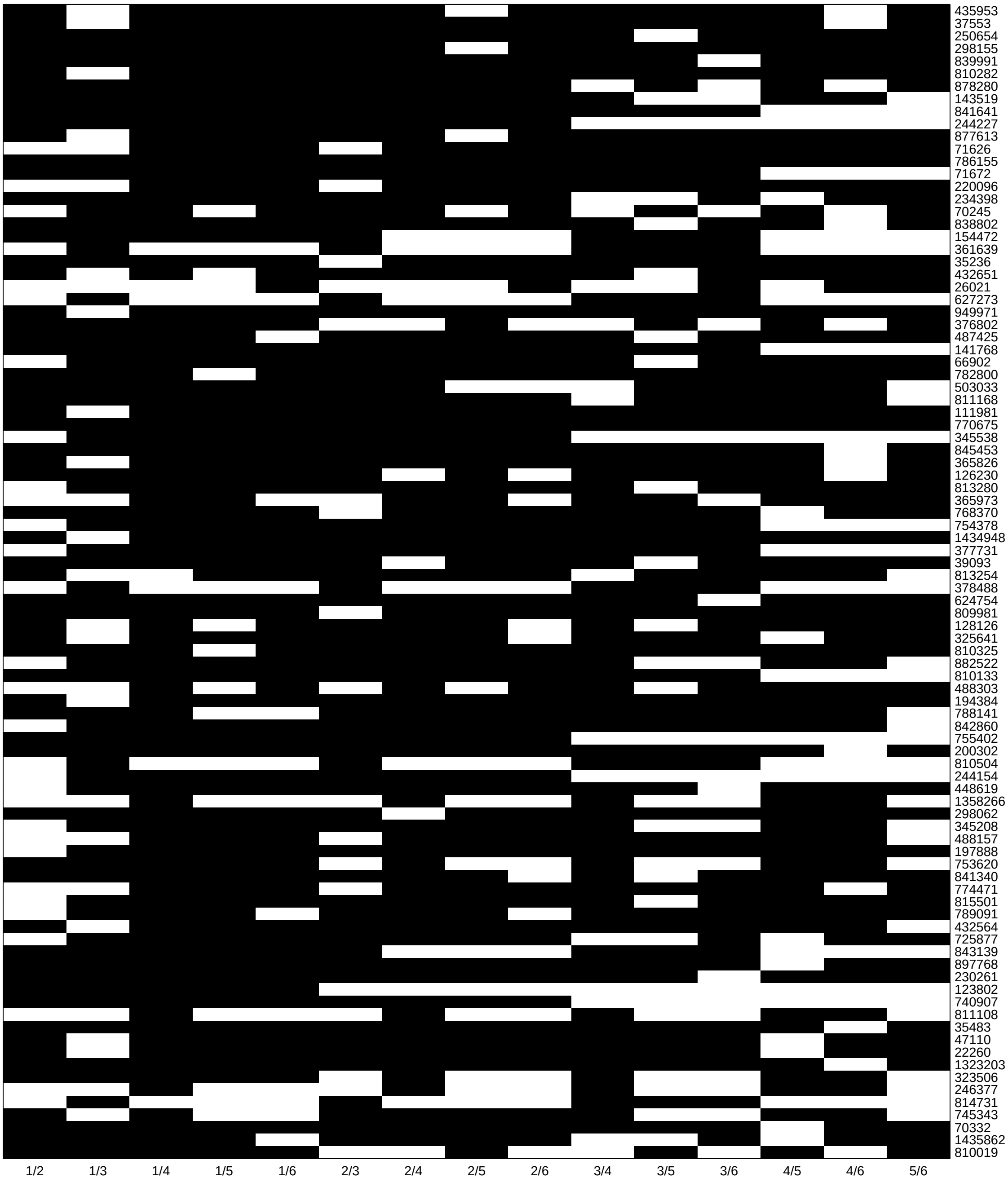


Figure 6.5: Pairwise variable selection results for the APFP method on the SRBCT data with top 100 genes. Each row corresponds to a gene. Each column corresponds to a cluster pair; for example, "1/2" indicates clusters 1 and 2. A black (white) spot indicates that the estimated means of the corresponding gene for the two clusters are different (the same). For example, gene "435953" is non-informative for separating clusters 1 and 3, 2 and 5, and 4 and 6.

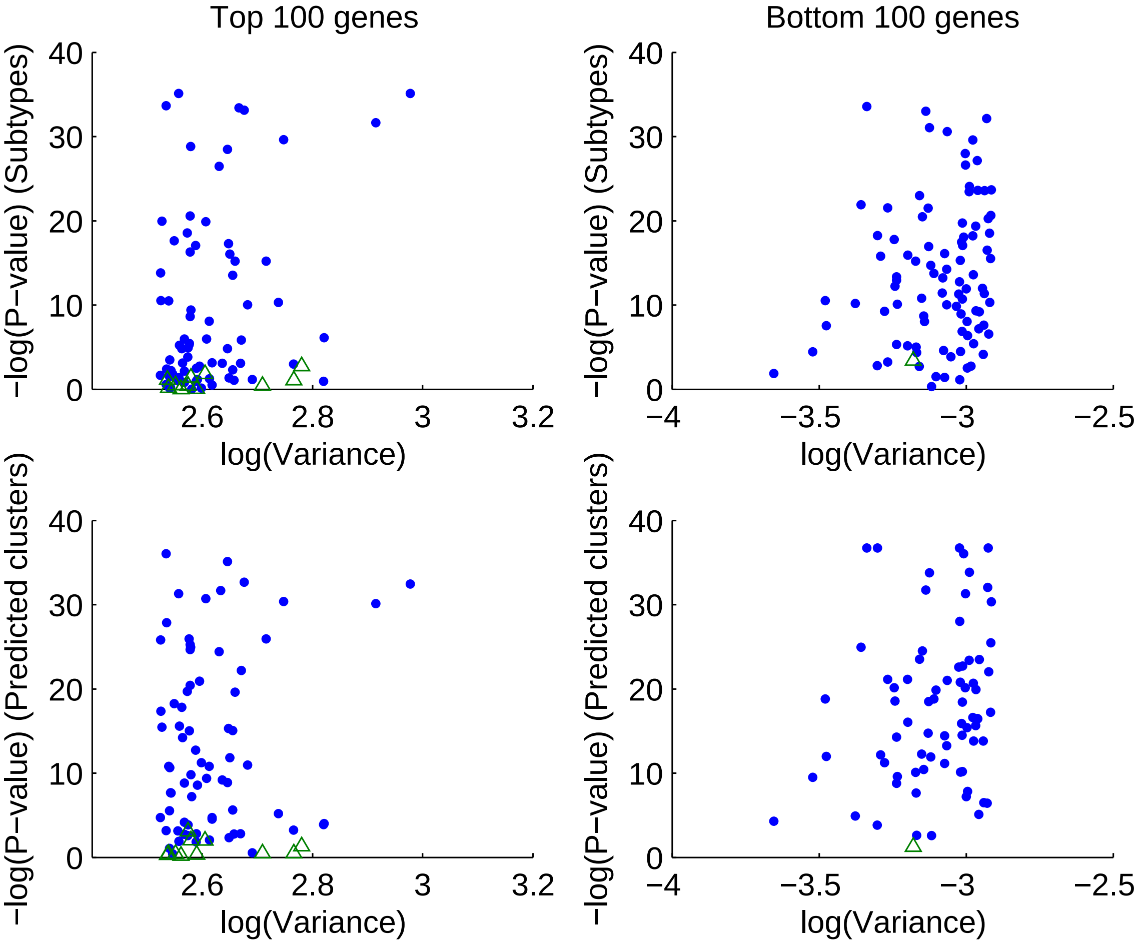


Figure 6.6: Plots of the negative logarithm p-values vs variance for PALL data. The left column is the top 100 genes (largest overall variances), and the right column is the bottom 100 genes. The upper row is negative logarithm p-values corresponding to an F-statistics comparing four tumor subtypes, and the lower row is the negative logarithm p-values for the six identified clusters. Triangles denote the genes that are not selected by the APFP method.

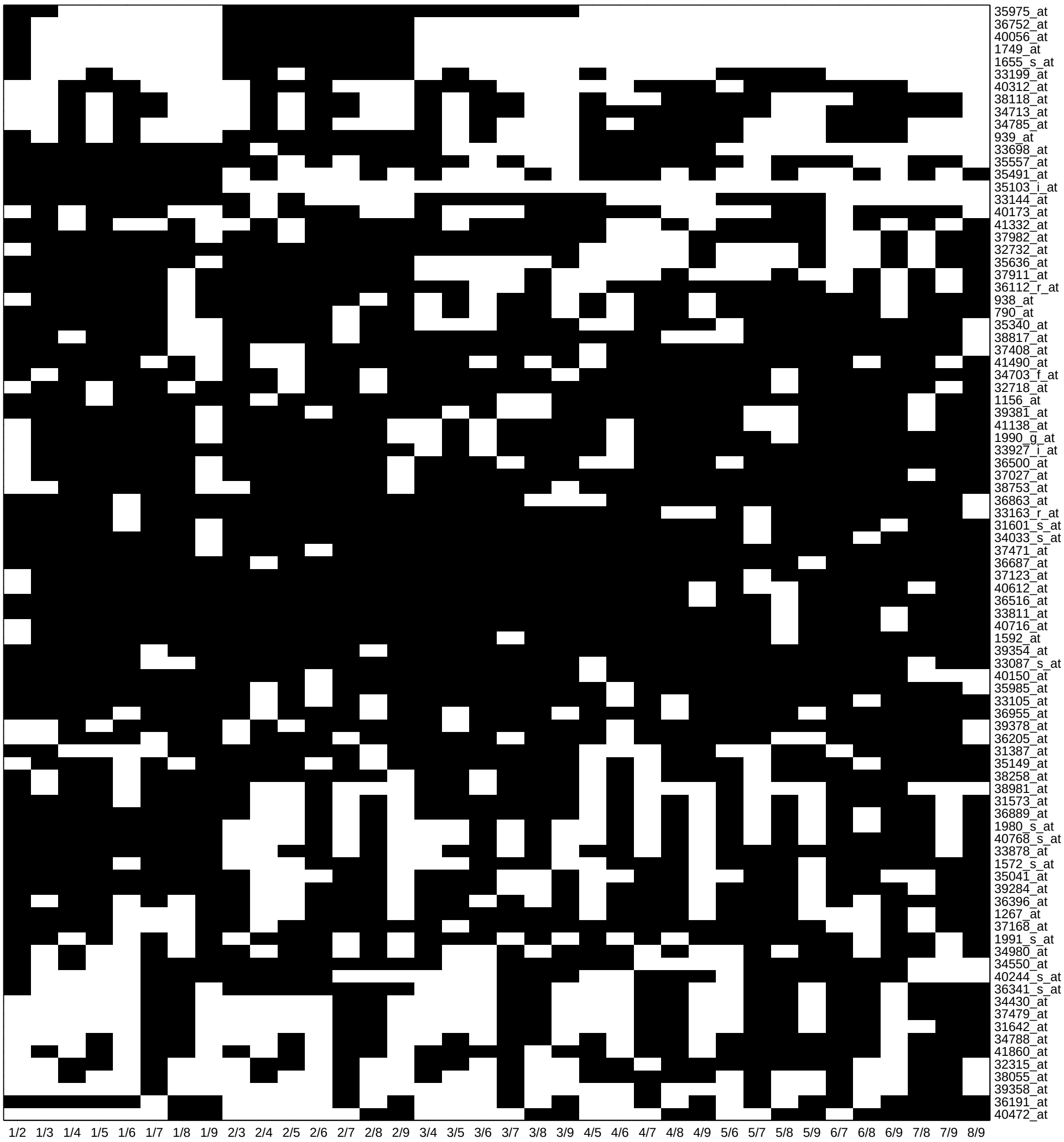


Figure 6.7: Pairwise variable selection results for the APFP method on the PALL data with top 100 genes. Each row corresponds to a gene. Each column corresponds to a cluster pair; for example, "1/2" indicates clusters 1 and 2. A black (white) spot indicates that the estimated means of the corresponding gene for the two clusters are different (the same).

CHAPTER VII

Sparse Fused Principal Component Analysis

7.1 Introduction

In this chapter, we propose another version of PCA with sparse components motivated by the following empirical considerations. In many application areas, some variables are highly correlated and form natural “blocks”. For example, in the meat spectra example discussed in Section 4, the spectra exhibit high correlations within the high and low frequency regions, thus giving rise to such a block structure. Something analogous occurs in the image data, where the background forms one natural block, and the foreground one or more such blocks. In such cases, the loadings of the block tend to be of similar magnitude. The proposed technique is geared towards exploring such block structures and producing sparse principal components whose loadings are of the same sign and magnitude, thus significantly aiding interpretation of the results. We call this property *fusion* and introduce a penalty that forces “fusing” of loadings of highly correlated variables in addition to forcing small loadings to zero. We refer to this method as sparse fused PCA (SFPCA).

The remainder of the paper is organized as follows: the technical development and computing algorithm for our method are presented in Section 7.2. An illustration of the method based on simulated data is given in Section 7.3. In Section 7.4, we apply the new method to several real datasets. Finally, some concluding remarks are drawn in Section 7.5.

7.2 The Model and its Estimation

7.2.1 Preliminaries and Sparse Variants of PCA

Let $\boldsymbol{X} = (x_{i,j})_{n\times p}$ be a data matrix comprised of n observations and p variables, whose columns are assumed to be centered. As noted above, PCA reduces the dimensionality of the data by constructing linear combinations of the original variables that have maximum variance; i. e., for $k = 1, \cdots, p$, define

$$\boldsymbol{\alpha}_k = \arg\max_{\boldsymbol{\alpha}} Var(\boldsymbol{X}\boldsymbol{\alpha}), \quad \text{subject to } \boldsymbol{\alpha}_k'\boldsymbol{\alpha}_k = 1, \boldsymbol{\alpha}_k'\boldsymbol{\alpha}_j = 0 \ \text{ for all } j \neq k, \tag{7.1}$$

where $\boldsymbol{\alpha}_k$ is a p-dimensional vector called *factor loadings* (PC vectors). The projection of the data $\boldsymbol{Z}_k = \boldsymbol{X}\boldsymbol{\alpha}_k$ is called the k-th principal component. The technique proves most successful if one can use a small number $k \ll p$ of components to account for most of the variance and thus provide a relatively simple explanation of the underlying data structure. Some algebra shows that the factor loadings can be obtained by solving the following optimization problem

$$\widehat{\boldsymbol{\alpha}}_k = \arg\max_{\boldsymbol{\alpha}\perp\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_{k-1}} \boldsymbol{\alpha}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{\alpha} \tag{7.2}$$

where $\widehat{\boldsymbol{\Sigma}} = 1/n(\boldsymbol{X}^T\boldsymbol{X})$ denotes the sample covariance of the data. The solution of (7.2) is given by the eigenvector corresponding to the k-th largest eigenvalue of $\widehat{\boldsymbol{\Sigma}}$. An alternative way to derive the PC vectors, which proves useful in subsequent developments, is to solve the following constrained least squares problem:

$$\min_{\boldsymbol{A}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{A}\boldsymbol{A}^T\|_F^2, \quad \text{subject to } \ \boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I}_K \ , \tag{7.3}$$

where I_K denotes a $K \times K$ identity matrix, $\|\boldsymbol{M}\|_F$ is the Frobenius norm of a matrix $\boldsymbol{M} = (m_{i,j})_{n\times p}$ ($\|\boldsymbol{M}\|_F^2 = \sum_{i,j} m_{ij}^2$), and $\boldsymbol{A} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K]$ is a $p \times K$ matrix with orthogonal columns. The estimate $\widehat{\boldsymbol{A}}$ contains the first K PC vectors, and $\widehat{\boldsymbol{Z}} = \boldsymbol{X}\widehat{\boldsymbol{A}}$ the first K principal components.

To impose sparsity on the PC vectors, Jollife et al. (2003) proposed SCoTLASS, which adds an ℓ_1-norm constraint to objective function (7.2), i.e., for any $1 \leq k \leq K$, solve:

$$\max_{\boldsymbol{\alpha}\perp\boldsymbol{\alpha}_1,\ldots,\alpha_{k-1}} \boldsymbol{\alpha}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{\alpha} \text{ , subject to } \|\boldsymbol{\alpha}\|_1 \leq t \text{ ,} \tag{7.4}$$

where $\|\boldsymbol{\alpha}\|_1 = \sum_{j=1}^p |\boldsymbol{\alpha}_j|$ is the ℓ_1 norm of the vector $\boldsymbol{\alpha}$. Due to the singularity property of the ℓ_1 norm, the constraint $\|\boldsymbol{\alpha}\|_1 \leq t$ shrinks some components of $\boldsymbol{\alpha}$ to zero for small enough values of t. Therefore, objective function (7.2) produces sparse PC vectors. However, Zou et al. (2006) noted that in many cases, SCoTLASS fails to achieve sufficient sparsity, thus complicating the interpretation of the results. One possible explanation stems from the orthogonality constraint of the PC vectors that is not fully compatible with the desired sparsity condition. Hence, Zou et al. (2006) proposed an alternative way to estimate sparse PC vectors, by relaxing the orthogonality requirement. Their procedure amounts to solving the following regularized regression problem:

$$\begin{aligned} \arg\min_{\boldsymbol{A},\boldsymbol{B}} \quad & \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^T\|_F^2 + \lambda_1 \sum_{k=1}^K \|\boldsymbol{\beta}_k\|_1 + \lambda_2 \sum_{k=1}^K \|\boldsymbol{\beta}_k\|_2^2 \\ \text{subject to} \quad & \boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I}_K \text{ ,} \end{aligned} \tag{7.5}$$

where $\boldsymbol{\beta}_k$ is a p-dimensional column vector and $\boldsymbol{B} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_K]$. The ℓ_2 penalty $\sum_{k=1}^K \|\boldsymbol{\beta}_k\|_2^2$ regularizes the loss function to avoid singular solutions, whenever $n < p$. If $\lambda_1 = 0$, objective function (7.5) reduces to the ordinary PCA problem and the columns of $\widehat{\boldsymbol{B}}$ are proportional to the first K ordinary PC vectors (Zou et al., 2006); otherwise, the ℓ_1 penalty $\|\boldsymbol{\beta}_k\|_1$ imposes sparsity on the elements of $\widehat{\boldsymbol{B}}$, i.e., it shrinks some loadings exactly to zero. In addition, the first term in (7.5) can be written as

$$
\begin{aligned}
\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^T\|_F^2 &= \|\boldsymbol{X}\boldsymbol{A}-\boldsymbol{X}\boldsymbol{B}\|_F^2+\|\boldsymbol{A}_\perp\|_F^2 \\
&= \sum_{k=1}^{K}\|\boldsymbol{X}\boldsymbol{\alpha}_k-\boldsymbol{X}\boldsymbol{\beta}_k\|_F^2+\|\boldsymbol{A}_\perp\|_F^2 \\
&= n\sum_{k=1}^{K}(\boldsymbol{\alpha}_k-\boldsymbol{\beta}_k)^T\widehat{\boldsymbol{\Sigma}}(\boldsymbol{\alpha}_k-\boldsymbol{\beta}_k)+\|\boldsymbol{A}_\perp\|_F^2 \qquad (7.6)
\end{aligned}
$$

where $\boldsymbol{A}_\perp$ is any orthonormal matrix such that $[\boldsymbol{A},\boldsymbol{A}_\perp]$ is a $p\times p$ orthonormal matrix. The quantity $(\boldsymbol{\alpha}_k-\boldsymbol{\beta}_k)^T\widehat{\boldsymbol{\Sigma}}(\boldsymbol{\alpha}_k-\boldsymbol{\beta}_k),\ 1\le k\le p$ measures the difference between $\boldsymbol{\alpha}_k$ and $\boldsymbol{\beta}_k$. Therefore, although there is no direct constraint on the column orthogonality in $\boldsymbol{B}$, the loss function shrinks the difference between $\boldsymbol{A}$ and $\boldsymbol{B}$ and this results in the columns of $\boldsymbol{B}$ becoming closer to orthogonal. Numerical examples in Zou et al. (2006) indicate that sparse PCA produces more zero loadings than SCoTLASS. However, both techniques cannot accommodate block structures in the variables, as the numerical results in Section 7.3 suggest. Next, we introduce a variant of sparse PCA called sparse fused PCA (SFPCA) that addresses this issue.

7.2.2 Sparse Fused Loadings

Our proposal is based on solving the following optimization problem:

$$
\begin{aligned}
\min_{\boldsymbol{A},\boldsymbol{B}} \quad & \|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^T\|_F^2+\lambda_1\sum_{k=1}^{K}\|\boldsymbol{\beta}_k\|_1+\lambda_2\sum_{k=1}^{K}\sum_{s<t}|\rho_{s,t}||\beta_{s,k}-sign(\rho_{s,t})\beta_{t,k}|\ , \\
\text{subject to} \quad & \boldsymbol{A}^T\boldsymbol{A}=\boldsymbol{I}_K\ , \qquad (7.7)
\end{aligned}
$$

where $\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^T\|_F^2=\sum_{i=1}^n\|\boldsymbol{x}_i-\boldsymbol{A}\boldsymbol{B}^T\boldsymbol{x}_i\|_2^2$; $\rho_{s,t}$ denotes the sample correlation between variables X_s and X_t and $sign(\cdot)$ the sign function. The first penalty in (7.7) is the sum of ℓ_1 norms of the K PC vectors. It aims to shrink the elements of the PC vectors to zero, thus ensuring sparsity of the resulting solution. The second penalty is a linear combination of K generalized *fusion* penalties. This penalty shrinks the difference between $\beta_{s,k}$ and $\beta_{t,k}$, if the correlation between variables X_s and X_t is positive; the higher the correlation, the heavier the penalty for on the difference of coefficients. If the correlation is negative, the penalty encourages $\beta_{s,k}$ and $\beta_{t,k}$ to have similar magnitudes, but different signs. It is natural to encourage the loadings of highly correlated variables to be close, since two perfectly correlated variables with the same variance have equal loadings. First, highly correlated variables on the same scale pushing the loadings to the same value has the same effect as setting small regression coefficients to 0 in lasso: fitted model accuracy is not affected much, but interpretation is improved and overfitting avoided. Second, by definition of principal components, the k-th PC vector maximizes the variance of $\sum_{j=1}^{p} \beta_{j,k} X_j$ subject to the orthogonality constraint. Since X_j's are centered, one can show that this variance equals to $\sum_{j=1}^{p} \beta_{j,k}^2 Var(X_j) + 2 \sum_{s<t} \beta_{s,k}\beta_{t,k} Cov(X_s, X_t)$. Thus, in order to maximize the variance, we need the sign of $\beta_{s,k}\beta_{t,k}$ to match the sign of $Cor(X_s, X_t)$ (as far as the orthogonality constraint will allow). Finally, note that if two variables are highly correlated but have substantially different variances, their loadings will have different scales and won't be fused to the same value, which is the correct behavior for PCA on unscaled data. If this behavior is undesirable in a particular application, data should be standardized first (just like in regular PCA, it is the user's decision whether to standardize the data).

The effect of the fusion penalty, due to the singularity property of the ℓ_1 norm, is that some terms in the sum are shrunken exactly to zero, resulting in some loadings having identical magnitudes. Therefore, the penalty aims at blocking the loadings into groups and "fusing" similar variables together for ease of interpretation. Finally, if $\rho_{s,t} = 0$ for any $|t - s| > 1$ and $\rho_{s,s+1}$ is a constant for all s, then the generalized fusion penalty reduces to the fusion penalty (Land and Friedman, 1996; Tibshirani et al., 2005).

Note that one can use other types of weights in the generalized fusion penalty, including partial correlations or other similarity measures Li and Li (2008).

7.2.3 Optimization of the Objective Function

We discuss next how to optimize the posited objective function. It is achieved through alternating optimization over $\boldsymbol{A}$ and $\boldsymbol{B}$, analogously to the sparse PCA algorithm. Overall, the algorithm proceeds as follows.

The Algorithm

Step 1. Initialize $\widehat{\boldsymbol{A}}$ by setting it to the ordinary PCA solution.

Step 2. Given $\boldsymbol{A}$, minimizing the objective function (7.7) over $\boldsymbol{B}$ is equivalent to solving the following K separate problems:

$$\min_{\beta_k} \|\boldsymbol{Y}_k^* - \boldsymbol{X}\boldsymbol{\beta}_k\|^2 + \lambda_1 \|\boldsymbol{\beta}_k\|_1 + \lambda_2 \sum_{s<t} |\rho_{s,t}||\beta_{s,k} - sign(\rho_{s,t})\beta_{t,k}| \tag{7.8}$$

where $\boldsymbol{Y}_k^* = \boldsymbol{X}\boldsymbol{\alpha}_k$. The solution to (7.8) is nontrivial, and is discussed in Section 7.2.4. This step updates the estimate $\widehat{\boldsymbol{B}}$.

Step 3. Given the value of $\boldsymbol{B}$, minimizing (7.7) over $\boldsymbol{A}$ is equivalent to solving

$$\arg\min_{\boldsymbol{A}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^T\|^2 \text{ , subject to } \boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I}_K \text{ .} \tag{7.9}$$

The solution can be derived by a reduced rank Procrustes rotation (Zou et al., 2006). Specifically, we compute the singular value decomposition (SVD) of $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{B} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$ and the solution to (7.9) is given by $\widehat{\boldsymbol{A}} = \boldsymbol{U}\boldsymbol{V}^T$. This step updates the estimate $\widehat{\boldsymbol{A}}$.

Step 4. Repeat Steps 2-3 until convergence.

7.2.4 Estimation of $\boldsymbol{B}$ Given $\boldsymbol{A}$

Objective function (7.8) can be solved by quadratic programming. However, this approach can be inefficient in practice; thus, we propose a more efficient algorithm–local quadratic approximation (LQA) (Fan and Li, 2001). This method has been employed in a number of variable selection procedures for regression and its convergence properties have been studied by Fan and Li (2001) and Hunter and Li (2005). The LQA method approximates the objective function locally via a quadratic form. Notice that

$$\begin{aligned}
& \sum_{s<t} |\rho_{s,t}||\beta_{s,k} - sign(\rho_{s,t})\beta_{t,k}| \\
= & \sum_{s<t} \frac{|\rho_{s,t}|}{|\beta_{s,k} - sign(\rho_{s,t})\beta_{t,k}|}(\beta_{s,k} - sign(\rho_{s,t})\beta_{t,k})^2 \\
= & \sum_{s<t} |w_{s,t}^{(k)}|(\beta_{s,k} - sign(w_{s,t})\beta_{t,k})^2 \qquad (7.10)
\end{aligned}$$

where $w_{s,t}^{(k)} = \rho_{s,t}/|\beta_{s,k} - sign(\rho_{s,t})\beta_{t,k}|$ and consequently $sign(w_{s,t}^{(k)}) = sign(\rho_{s,t})$.

After some algebra, one can show that (7.10) can be written as $\boldsymbol{\beta}^T\boldsymbol{L}^{(k)}\boldsymbol{\beta}$, where $\boldsymbol{L}^{(k)} = \boldsymbol{D}^{(k)} - \boldsymbol{W}^{(k)}$, $\boldsymbol{W}^{(k)} = (w_{s,t})_{p\times p}$ with diagonal elements equal to zero, and $\boldsymbol{D}^{(k)} = diag(\sum_{t\neq 1}|w_{1,t}|, \ldots, \sum_{t\neq p}|w_{p,t}|)$.

Similarly, we have $\|\beta_k\|_1 = \sum_{j=1}^p |\beta_{j,k}| = \sum_{j=1}^p \omega_j^{(k)}\beta_{j,k}^2 = \boldsymbol{\beta}^T\boldsymbol{\Omega}^{(k)}\boldsymbol{\beta}$, where $\boldsymbol{\omega}_j^{(k)} = 1/|\beta_{j,k}|$ and $\boldsymbol{\Omega}^{(k)} = diag(\omega_1^{(k)}, \ldots, \omega_p^{(k)})$. Then, (7.8) can be written as

$$\min_{\boldsymbol{\beta}_k} \|\boldsymbol{Y}_k^* - \boldsymbol{X}\boldsymbol{\beta}_k\|_2^2 + \lambda_1\boldsymbol{\beta}^T\boldsymbol{\Omega}^{(k)}\boldsymbol{\beta} + \lambda_2\boldsymbol{\beta}^T\boldsymbol{L}^{(k)}\boldsymbol{\beta}. \qquad (7.11)$$

Notice that (7.11) takes the form of a least squares problem involving two generalized ridge penalties; hence, its closed form solution is given by

$$\widehat{\boldsymbol{\beta}}_k = (\boldsymbol{X}^T\boldsymbol{X} + \lambda_1\boldsymbol{\Omega}^{(k)} + \lambda_2\boldsymbol{L}^{(k)})^{-1}\boldsymbol{X}^T\boldsymbol{Y}_k^*. \qquad (7.12)$$

Notice that both $\boldsymbol{\Omega}^{(k)}$ and $\boldsymbol{L}^{(k)}$ depend on the unknown parameter $\boldsymbol{\beta}_k$. Specifically, LQA iteratively updates $\boldsymbol{\beta}_k$, $\boldsymbol{L}^{(k)}$ and $\boldsymbol{\Omega}^{(k)}$ as follows, which constitute Step 2 of the algorithm.

Step 2(a). Given $\widehat{\boldsymbol{\beta}}_k$ from the previous iteration, update $\widehat{\boldsymbol{\Omega}}^{(k)}$ and $\widehat{\boldsymbol{L}}^{(k)}$.

Step 2(b). Given $\widehat{\boldsymbol{\Omega}}^{(k)}$ and $\widehat{\boldsymbol{L}}^{(k)}$, update $\widehat{\boldsymbol{\beta}}_k$ by formula (7.12).

Step 2(c). Repeat Steps 2(a) and 2(b) until convergence.

Step 2(d). Scale $\widehat{\boldsymbol{\beta}}_k$ to have unit ℓ_2-norm.

Note that to calculate $\boldsymbol{L}^{(k)}$ in step 2(a), we need to calculate $w_{s,t} = \rho_{s,t}/|\beta_{k,s} - sign(\rho_{s,t})\beta_{k,t}|$. When the values of $\beta_{k,s}$ and $sign(\rho_{s,t})\beta_{k,t}$ are extremely close, $w_{s,t}$ is numerically singular. In this case, we replace $|\beta_{k,s} - sign(\rho_{s,t})\beta_{k,t}|$ by a very small positive number (e.g. 10^{-10}); similarly, we replace $|\beta_{j,k}|$ by a very small positive number if its value is extremely close to 0.

With the new Step 2, the algorithm has two nested loops. However, the inner loop in Step 2 can be effectively approximated by a one step update (Hunter and Li, 2005), i.e., by removing step 2(c). In our numerical experiments, we found that this one step update can lead to significant computational savings without minor sacrifices in terms of numerical accuracy.

7.2.5 Selection of Tuning Parameters

The proposed procedure involves two tuning parameters. One can always use cross-validation to select the optimal values, but it can be computationally expensive. We discuss next an alternative approach for tuning parameter selection based on the Bayesian information criterion (BIC), which we use in simulations in Section 7.3. In general, we found solutions from cross-validation and BIC to be comparable, but BIC solutions tend to be sparser.

Let $\boldsymbol{A}^{\lambda_1,\lambda_2} = [\boldsymbol{\alpha}_1^{\lambda_1,\lambda_2}, \ldots, \boldsymbol{\alpha}_K^{\lambda_1,\lambda_2}]$ and $\boldsymbol{B}^{\lambda_1,\lambda_2} = [\boldsymbol{\beta}_1^{\lambda_1,\lambda_2}, \ldots, \boldsymbol{\beta}_K^{\lambda_1,\lambda_2}]$ be the estimates of $\boldsymbol{A}$ and $\boldsymbol{B}$ in (7.7), obtained using tuning parameters λ_1 and λ_2. Let

$\widehat{\sigma}_\epsilon^2 = 1/n \sum_{i=1}^n \|\boldsymbol{X} - \boldsymbol{X}\widehat{\boldsymbol{A}}\widehat{\boldsymbol{A}}^T\|_F^2$, where the columns of $\widehat{\boldsymbol{A}}$ contain the first K ordinary PC vectors of $\boldsymbol{X}$. We define the BIC for sparse PCA as follows:

$$BIC(\lambda_1, \lambda_2) = \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}^{\lambda_1,\lambda_2}(\boldsymbol{A}^{\lambda_1,\lambda_2})^T\|_F^2/\widehat{\sigma}_\epsilon^2 + \log(n)df^{SPCA} \tag{7.13}$$

and analogously for SFPCA

$$BIC(\lambda_1, \lambda_2) = \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}^{\lambda_1,\lambda_2}(\boldsymbol{A}^{\lambda_1,\lambda_2})^T\|_F^2/\widehat{\sigma}_\epsilon^2 + \log(n)df^{SFPCA} \tag{7.14}$$

where df^{SPCA} and df^{SFPCA} denote the degrees of freedom of sparse and sparse-fused PCA defined as the number of all nonzero/nonzero-distinct elements in B^{λ_1,λ_2}, respectively. These definitions are similar to df defined for Lasso and fused Lasso (Zou et al., 2007; Tibshirani et al., 2005).

7.2.6 Computational Complexity and Convergence

Since $\boldsymbol{X}^T\boldsymbol{X}$ only depends on the data, it is calculated once and requires np^2 operations. The estimation of $\boldsymbol{A}$ by solving an SVD takes $O(pK^2)$. Calculation of $\boldsymbol{\Omega}$ and $\boldsymbol{L}$ in (7.11) requires $O(p^2)$ operations, while the inverse in (7.12) is of order $O(p^3)$. Therefore, each update in LQA is of order $O(p^3K)$, and the total computational cost is $O(np^2) + O(p^3K)$.

The convergence of the algorithm essentially follows from standard results. Note that the loss function is strictly convex in both $\boldsymbol{A}$ and $\boldsymbol{B}$, and the penalties are convex in $\boldsymbol{B}$, and thus the objective function is strictly convex and has a unique global minimum. The integrations between Step 2 and Step 3 of the Algorithm amount to block coordinate descent, which is guaranteed to converge for differentiable convex functions (see, e.g., Bazaraa et al. (1993)). The original objective function has singularities, but the objective function (7.10) obtained from the local quadratic approximation that we are actually optimizing is differentiable everywhere, and thus the convergence of coordinate descent is guaranteed. Thus, we only need to make sure that each step of the coordinate descent is guaranteed to converge. In Step 3, we are optimizing the objective function (7.9) exactly and obtain the solution in closed form. In Step 2, the optimization is iterative, but convergence follows easily by adapting the arguments of Hunter and Li (2005) for local quadratic approximation obtained from general results for minorization-maximization algorithms.

7.3 Numerical illustration of SFPCA

First, we illustrate the performance of the proposed SFPCA method on a number of synthetic datasets described next.

Simulation 1

This simulation scenario is adopted from Zou et al. (2006). Three latent variables are generated as follows:

$$\begin{aligned} V_1 &\sim N(0, 290), \\ V_2 &\sim N(0, 300), \\ V_3 &= -0.3V_1 + 0.6V_2 + \epsilon, \end{aligned}$$

where V_1, V_2 and ϵ are independent, and $\epsilon \sim N(0,1)$. Next, ten observable variables are constructed as follows:

$$X_j = \begin{cases} V_1 + e_j, & \text{if } 1 \leq j \leq 4; \\ V_2 + e_j, & \text{if } 5 \leq j \leq 8; \\ V_3 + e_j, & \text{if } j = 9, 10; \end{cases}$$

where $\epsilon_j, 1 \leq j \leq 10$ are i.i.d. $N(0,1)$. The variance of the three latent variables are 290, 300 and 38, respectively. Notice that by construction, variables X_1 through X_4

form a block with a constant within-block pairwise correlation of .997 ("block 1"), while variables X_5 through X_8 and X_9, X_{10} form another two blocks ("block 2" and "block 3", respectively). Ideally, a sparse first PC should pick up block 2 variables with equal loadings, while a sparse second PC should consist of block 1 variables with equal loadings, since the variance of V_2 is larger than that of V_1.

Zou et al. (2006) compared sparse PCA with ordinary PCA and SCoTLASS using the true covariance matrix. In our simulation, we opted for the more realistic procedure of generating 20 samples according to the above description and repeated the simulation 50 times. PC vectors from ordinary PCA, sparse PCA and SFPCA were computed from these simulated datasets and the results are shown in Table 7.1, along with the ordinary PC vectors computed from the true covariance matrix. The table entries correspond to the median and the median absolute deviation (in parentheses) of the loadings over 50 replications. To measure the variation of the loadings within block 1 and 2, we also calculated the standard deviation among the loadings within these blocks and record their medians and median absolute deviations in rows "Block 1" and "Block 2", respectively. The proportions of adjusted variance and adjusted cumulative variance are reported as "AV (%)" and "ACV (%)". Adjusted variance was defined by Zou et al. (2006) as follows: let $\widehat{B}$ be the first K modified PC vectors. Using the QR decomposition, we have $\boldsymbol{X}\widehat{\boldsymbol{B}} = \boldsymbol{Q}\boldsymbol{R}$, where $\boldsymbol{Q}$ is orthonormal and $\boldsymbol{R}$ is upper triangular. Then the adjusted variance of the k-th PC equals $R^2_{k,k}$.

The tuning parameters were selected by minimizing the Bayesian information criterion (BIC) defined in Section 7.2.5, using a grid search over $\{2^{-10}, 2^{-9}, \ldots, 2^{10}\}$ for λ_1 and $\{10^{-3}, \ldots, 10^{3}\}$ for λ_2, respectively.

Table 7.1 shows that both SFPCA and sparse PCA recover the correct sparse structure of the loadings in the first two PC vectors. The median standard deviations within block 2 in PC 1 and block 1 in PC 2 equal to zero, which implies that SFPCA accurately recovers the loadings within the block. In contrast, the median standard deviations within block 2 in PC 1 and within block 1 in PC 2 reveal that the loadings estimated by sparse PCA exhibit significant variation.

As discussed in Section 2, the PC vectors from both sparse PCA and SFPCA are not exactly orthogonal due to the penalties employed. To study the deviation from orthogonality, the histogram of pairwise angles between the first four PC vectors obtained from SFPCA was obtained (available as supplemental material). It can be seen that the first two PCs are always orthogonal, while the fourth PC is essentially always orthogonal to the remaining three. The third component is the most variable, sometimes being close to the first, and at other times close to the second PC. This distribution of angles is consistent with the structure of the simulation and in general will be dependent on the underlying structure of the data.

Simulation 2

This example is a high-dimensional version ($p > n$) of simulation 1. We define

$$X_j = \begin{cases} V_1 + e_j, & \text{if } 1 \leq j \leq 20; \\ V_2 + e_j, & \text{if } 21 \leq j \leq 40; \\ V_3 + e_j, & \text{if } 41 \leq j \leq 50; \end{cases}$$

where $\epsilon_j, 1 \leq j \leq 50$ are i.i.d. $N(0, 1)$. Then 20 samples were generated in each of the 50 repetitions. The factor loadings estimated from this simulation are illustrated in Figure 7.1. Sparse PCA and SFPCA produce similar sparse structures in the loadings. However, compared with the "jumpy" loadings from sparse PCA, the loadings estimated by SFPCA are smooth and easier for interpretation.

7.4 Application of SFPCA to Real Datasets

Drivers Dataset

This dataset provides information about the physical size and age of 38 drivers along with a response variable, seat position in a car. (Faraway, 2004). For the purposes of PCA, the response variable was excluded from the analysis. The eight available variables on driver characteristics are age, weight, height in shoes, height in bare feet, seated height, lower arm length, thigh length, and lower leg length. All height/length variables are highly correlated (average correlation among these variables is about 0.8) and form a natural block; hence, we expect them to have similar loadings. SFPCA was applied to this dataset and compared its results with those obtained from ordinary PCA and sparse PCA (Table 7.2).

It can be seen that ordinary PCA captures the block structure in the first PC, but the factor loadings exhibit significant variation. Interestingly, the factor loadings from sparse PCA exhibit even greater variability, while the percentage of total variance explained by the first PC is only 55%, as opposed to 70% by ordinary PCA. On the other hand, SPFCA exhibits good performance in terms of goodness of fit (68.7%) and clearly reveals a single block structure in the “size” variables.

Pitprops Dataset

The pitprops dataset, introduced in Jeffers (1967), has become a classic example of the difficulties in interpretation of principal components. In this dataset, the sizes and properties of 180 pitprops (lumbers used to support the roofs of tunnels in coal mines) are recorded. The available variables are: the top diameter of the prop (topdiam), the length of the prop (length), the moisture content of the prop (moist), the specific gravity of the timber at the time of the test (testsg), the oven-dry specific gravity of the timber (ovensg), the number of annual rings at the top of the prop (ringtop), the number of annual rings at the base of the prop (ringbut), the maximum bow (bowmax), the distance of the point of maximum bow from the top of the prop (bowdist), the number of knot whorls (whorls), the length of clear prop from the top of the prop (clear), the average number of knots per whorl (knots) and the average diameter of the knots (diaknot). The first six PCs from regular PCA account for 87% of the total variability (measured by cumulative proportion of total variance explained).

We applied SPFCA and sparse PCA to the dataset and the results are given in Table 7.3. The loadings from SFPCA show a sparse structure similar to that of sparse PCA, but the first three PCs from SFPCA involve fewer variables than those of SPCA. The equal loadings within blocks assigned by SFPCA produce a clear picture for interpretation purposes. Referring to the interpretation in Jeffers (1967), the first PC gives the same loadings to "topdiam", "length", "ringbut", "bowmax", "bowdist" and "whorls" and provides a general measure of size; the second PC assigns equal loadings to "moist" and "testsg" and measures the degree of seasoning; the third PC, giving equal loadings to "ovensg" and "ringtop", accounts for the rate of the growth and the strength of the timber; the following three PCs represent "clear", "knots" and "diaknot", respectively.

Meat Spectrum Data

In this section, we apply SFPCA to a dataset involving spectra obtained from meat analysis (Borggaard and Thodberg, 1992; Thodberg, 1996). In recent decades, spectrometry techniques have been widely used to identify the fat content in pork, because it has proved significantly cheaper and more efficient than traditional analytical chemistry methods. In this dataset, 215 samples were analyzed by a Tecator near-infrared spectrometer which measured the spectrum of light transmitted through a sample of minced pork meat. The spectrum gives the absorbance at 100 wavelength channels in the range of 850 to 1050 nm.

The adjusted cumulative total variances explained by the first two PCs from ordinary PCA, sparse PCA and SFPCA are 99.6%, 98.9% and 98.4%, respectively. Since wavelengths are naturally ordered, a natural way to display the loadings is to plot them against the wavelength. The plot of the first two PCs for the 100 wavelength channels is shown in Figure 7.3.

SFPCA smoothes the ordinary PC vectors producing piece-wise linear curves which are easier to interpret. The SFPCA results show clearly that the first PC represents the overall mean over different wavelengths while the second PC represents a contrast between the low and high frequencies. On the other hand, the high variability in the loadings produces by sparse PCA makes the PC curves difficult to interpret.

USPS Handwritten Digit Data

In this example, the three PCA methods are compared on the USPS handwritten digit data set (Hull, 1994). This data set was collected by the US Postal Service (USPS) and contains 11,000 gray scale digital images of the ten digits at 16 × 16 pixel resolution. We focused on the digit "3" and sampled 20 images at random, thus operating in a large p, small n setting. While BIC gave good results for most data sets we examined, for the USPS data it tended to under shrink the coefficient estimates. However, we found that cross-validation produced good results and was computationally feasible, so we used five- fold cross-validation to select the optimal tuning parameters for SPCA and SFPCA. The optimal tuning parameter for SPCA turned out to be equal to zero, so here SPCA coincides with ordinary PCA. The reconstructed images by the first and second principal components ("eigen-images") arranged in the original spatial order are shown in Figure 7.4. It can be seen that SFPCA achieves a fairly strong fusing effect for the background pixels, thus producing a smoother, cleaner background image. This is confirmed by the results in Table 7.4 that give the proportion of distinct elements in the first two principal components for PCA and SFPCA. Notice that since PCA does not impose any sparsity or fusion, the resulting proportion is 100%, compared to those for SFPCA (35.5% and 22.7% for the first and second PCs, respectively).

7.5 Concluding Remarks

In this paper, a method is developed to estimate principal components that capture block structures in the variables, which aids in the interpretation of the data analysis results. To achieve this goal, the orthogonality requirement is relaxed and an ℓ_1 penalty is imposed on the norm of the PC vectors, as well as a "fusion" penalty driven by variable correlations. Application of the method to both synthetic and real data sets illustrates its advantages when it comes to interpretation.

The idea of sparse fused loadings is also applicable in a number of other unsupervised learning techniques, including canonical correlation and factor analysis, as well as regression analysis, classification techniques (e.g., LDA and SVM) and survival analysis (e.g., Cox model and Buckley-James model). We note that Daye and Jeng (2009) proposed a weighted fusion penalty for variable selection in a regression model. Unlike the generalized fusion penalty which penalizes the pairwise Manhattan distances between the variables, their method penalizes the pairwise Euclidean distances, and thus would not necessarily shrink the coefficients of highly correlated variables to identical values. Similarly, Tutz and Ulbricht (2009) proposed a BlockBoost method, whose penalty also tends to fuse the pairwise difference between the regression coefficients. In particular, when these pairwise correlations are close to ±1, the solution of BlockBoost is closed to that of Daye and Jeng (2009).

Table 7.1: Results for simulation 1. “PCA-T” corresponds to the ordinary PCA estimation from the true covariance matrix. “PCA-S” corresponds to the ordinary PCA estimation from the sample covariance matrix. “SPCA” represents the sparse PCA, and “SFPCA” represents the sparse fused PCA. “AV” is the adjusted variance, and “ACV” is the adjusted cumulative variance. The row“Block 1” shows the standard deviation of the loadings of variables 1 to 4, and “Block 2” shows the same for variables 5 to 8. In each row, the top entry is the median and the bottom entry in parentheses is the median absolute deviation over 50 replications.

Loadings	PC 1				PC 2			
	PCA-T	PCA-S	SPCA	SFPCA	PCA-T	PCA-S	SPCA	SFPCA
1	0.055	-0.123	0	0	0.488	0.447	0.506	0.500
	(—)	(0.162)	(0)	(0)	(—)	(0.032)	(0.072)	(0)
2	0.055	-0.127	0	0	0.488	0.444	0.492	0.500
	(—)	(0.161)	(0)	(0)	(—)	(0.031)	(0.085)	(0)
3	0.055	-0.129	0	0	0.488	0.448	0.491	0.500
	(—)	(0.161)	(0)	(0)	(—)	(0.033)	(0.085)	(0)
4	0.055	-0.125	0	0	0.488	0.442	0.493	0.500
	(—)	(0.159)	(0)	(0)	(—)	(0.032)	(0.089)	(0)
5	-0.453	0.376	0.422	0.487	0.089	0.164	0	0
	(—)	(0.040)	(0.021)	(0.015)	(—)	(0.131)	(0)	(0)
6	-0.453	0.374	0.415	0.487	0.089	0.165	0	0
	(—)	(0.038)	(0.021)	(0.016)	(—)	(0.133)	(0)	(0)
7	-0.453	0.375	0.417	0.487	0.089	0.161	0	0
	(—)	(0.040)	(0.019)	(0.015)	(—)	(0.133)	(0)	(0)
8	-0.453	0.376	0.417	0.487	0.089	0.159	0	0
	(—)	(0.038)	(0.020)	(0.015)	(—)	(0.127)	(0)	(0)
9	-0.289	0.389	0.382	0.155	-0.093	-0.015	0	0
	(—)	(0.025)	(0.021)	(0.122)	(—)	(0.132)	(0)	(0)
10	-0.289	0.389	0.388	0.155	-0.093	-0.009	0	0
	(—)	(0.026)	(0.027)	(0.119)	(—)	(0.127)	(0)	(0)
Block 1	0	0.003	0	0	0	0.002	0.064	0
	(—)	(0.003)	(0)	(0)	(—)	(0.002)	(0.050)	(0)
Block 2	0	0.001	0.014	0	0	0.004	0	0
	(—)	(0.001)	(0.014)	(0)	(—)	(0.003)	(0)	(0)
AV (%)	42.7	61.9	57.6	47.3	40.3	37.7	37.1	36.7
	(—)	(4.4)	(1.0)	(6.3)	(—)	(4.2)	(2.2)	(1.5)
ACV (%)	42.7	61.9	57.6	47.3	83.0	99.5	95.1	83.7
	(—)	(4.4)	(1.0)	(6.3)	(—)	(0.1)	(2.7)	(6.1)

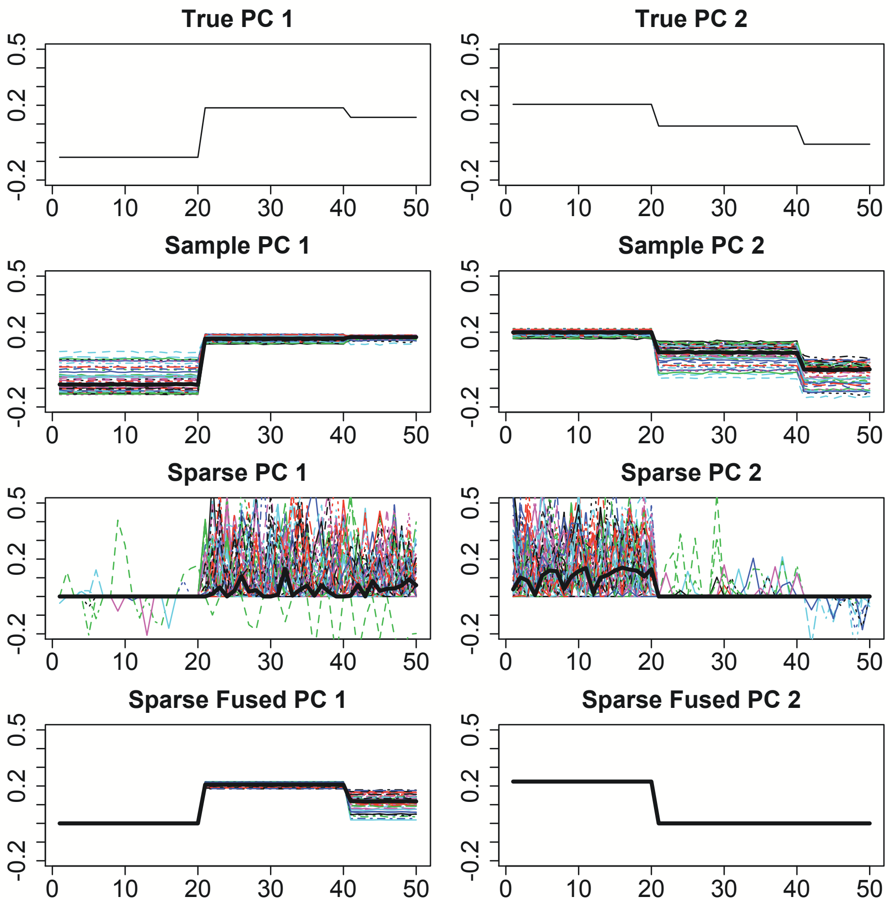


Figure 7.1: Factor loadings of the first (left column) and second (right column) PC vectors estimated by ordinary PCA from the true covariance (first row), ordinary PCA from the sample covariance (second row), sparse PCA (third row) and SFPCA (fourth row). The horizontal axis is the variables and the vertical axis is the value of the loadings. Each colored curve represents the PC vector in one replication. The median loadings over 50 repetitions are represented by the black bold lines.

Table 7.2: Numerical results for the drivers example.

Variables	PC 1			PC 2		
	PCA	SPCA	SFPCA	PCA	SPCA	SFPCA
Age	0.007			0.876	0.970	1.000
Weight	0.367	0.284	0.378	0.045		
HtShoes	0.411	0.139	0.378	-0.106		
Ht	0.412	0.764	0.378	-0.112		
Seated	0.381	0.313	0.378	-0.218		
Arm	0.349	0.208	0.378	0.374	0.242	
Thigh	0.328	0.247	0.378	0.125		
Leg	0.390	0.341	0.378	-0.056		
AV (%)	70.9	55.0	68.7	15.5	14.2	12.2
ACV (%)	70.9	55.0	68.7	86.4	69.2	80.8

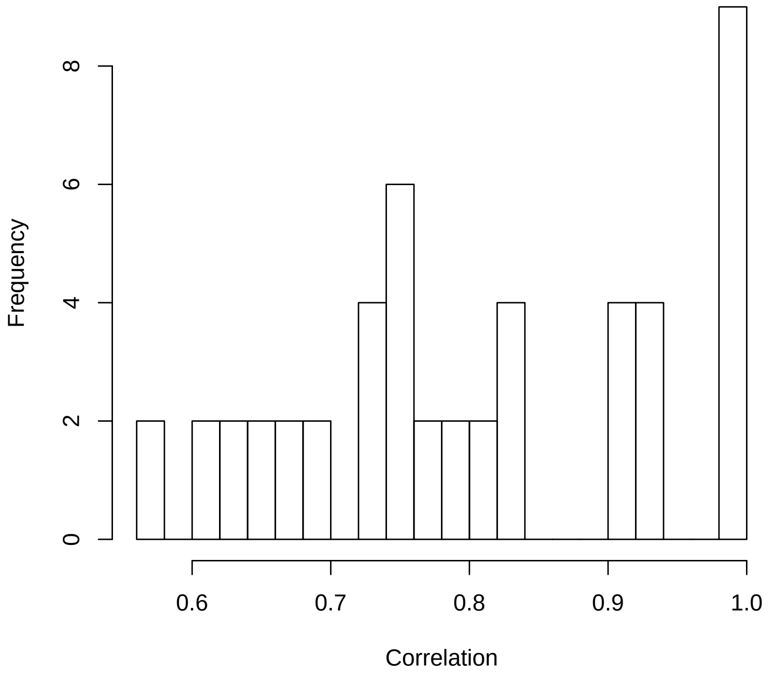


Figure 7.2: The histogram of the pairwise correlations between the height/length variables: weight, height in shoes, height in bare feet, seated height, lower arm length, thigh length, and lower leg length.

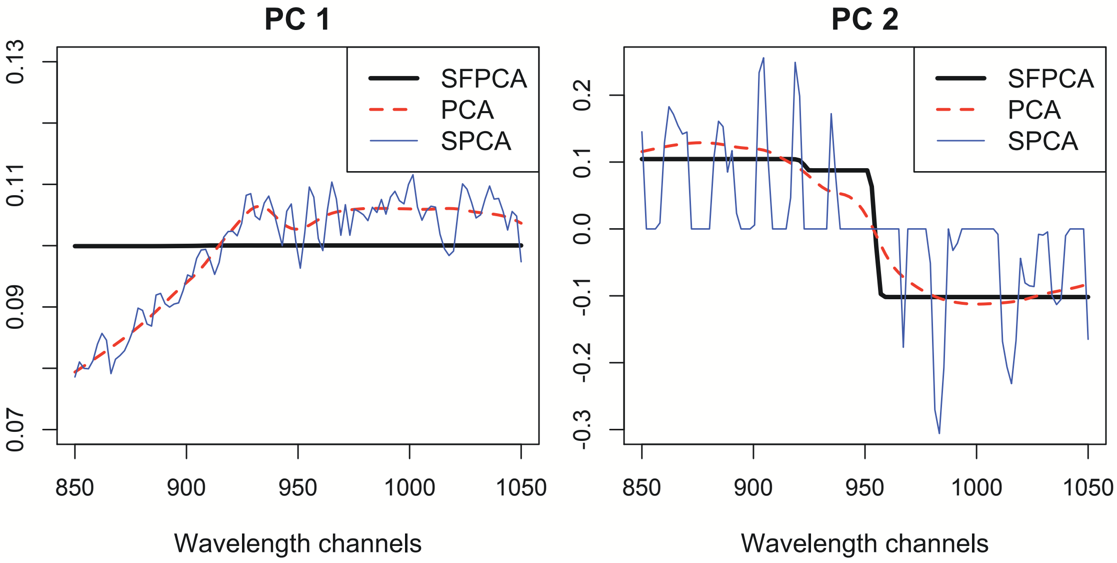


Figure 7.3: Comparison of the first (left panel) and second (right panel) PC vectors from ordinary PCA (dashed line), sparse PCA (dotted line) and SFPCA (solid line).

Table 7.3: Numerical results for the pitprops example.

Variables	PC 1			PC 2			PC 3		
	PCA	SPCA	SFPCA	PCA	SPCA	SFPCA	PCA	SPCA	SFPCA
topdiam	0.404	0.477	0.408	0.218			-0.207		
length	0.406	0.476	0.408	0.186			-0.235		
moist	0.124			0.541	0.785	0.707	0.141		
testsg	0.173			0.456	0.620	0.707	0.352		
ovensg	0.057	-0.177		-0.170			0.481	0.640	0.707
ringtop	0.284		0.052	-0.014			0.475	0.589	0.707
ringbut	0.400	0.250	0.408	-0.190			0.253	0.492	
bowmax	0.294	0.344	0.408	-0.189	-0.021		-0.243		
bowdist	0.357	0.416	0.408	0.017			-0.208		
whorls	0.379	0.400	0.408	-0.248			-0.119		
clear	-0.011			0.205			-0.070		
knots	-0.115			0.343	0.013		0.092	-0.015	
diaknot	-0.113			0.309			-0.326	-0.308519	
AV (%)	32.4	28.0	31.5	18.3	14.4	15.1	14.4	13.3	10.1
ACV (%)	32.4	28.0	31.5	50.7	42.0	46.6	65.1	55.3	56.7

Variables	PC 4			PC 5			PC 6		
	PCA	SPCA	SFPCA	PCA	SPCA	SFPCA	PCA	SPCA	SFPCA
topdiam	-0.091			0.083			0.120		
length	-0.103			0.113			0.163		
moist	0.078			-0.350			-0.276		
testsg	0.055			-0.356			-0.054		
ovensg	0.049			-0.176			0.626		
ringtop	-0.063			0.316			0.052		
ringbut	-0.065			0.215			0.003		
bowmax	0.286			-0.185			-0.055		
bowdist	0.097			0.106			0.034		
whorls	-0.205			-0.156			-0.173		
clear	0.804	1.000	1.000	0.343			0.175		
knots	-0.301			0.600	1.000	1.000	-0.170		
diaknot	-0.303			-0.08			0.626	1.000	1.000
AV (%)	8.5	7.4	8.0	7.0	6.8	7.3	6.3	6.2	7.0
ACV (%)	73.6	62.7	64.7	80.6	69.5	72.0	86.9	75.8	79.0

Table 7.4: The proportion of distinct elements in the eigen-images of digit “3” estimated by PCA and SFPCA, respectively.

PC	PCA (%)	SFPCA (%)
1	100	35.5
2	100	22.7

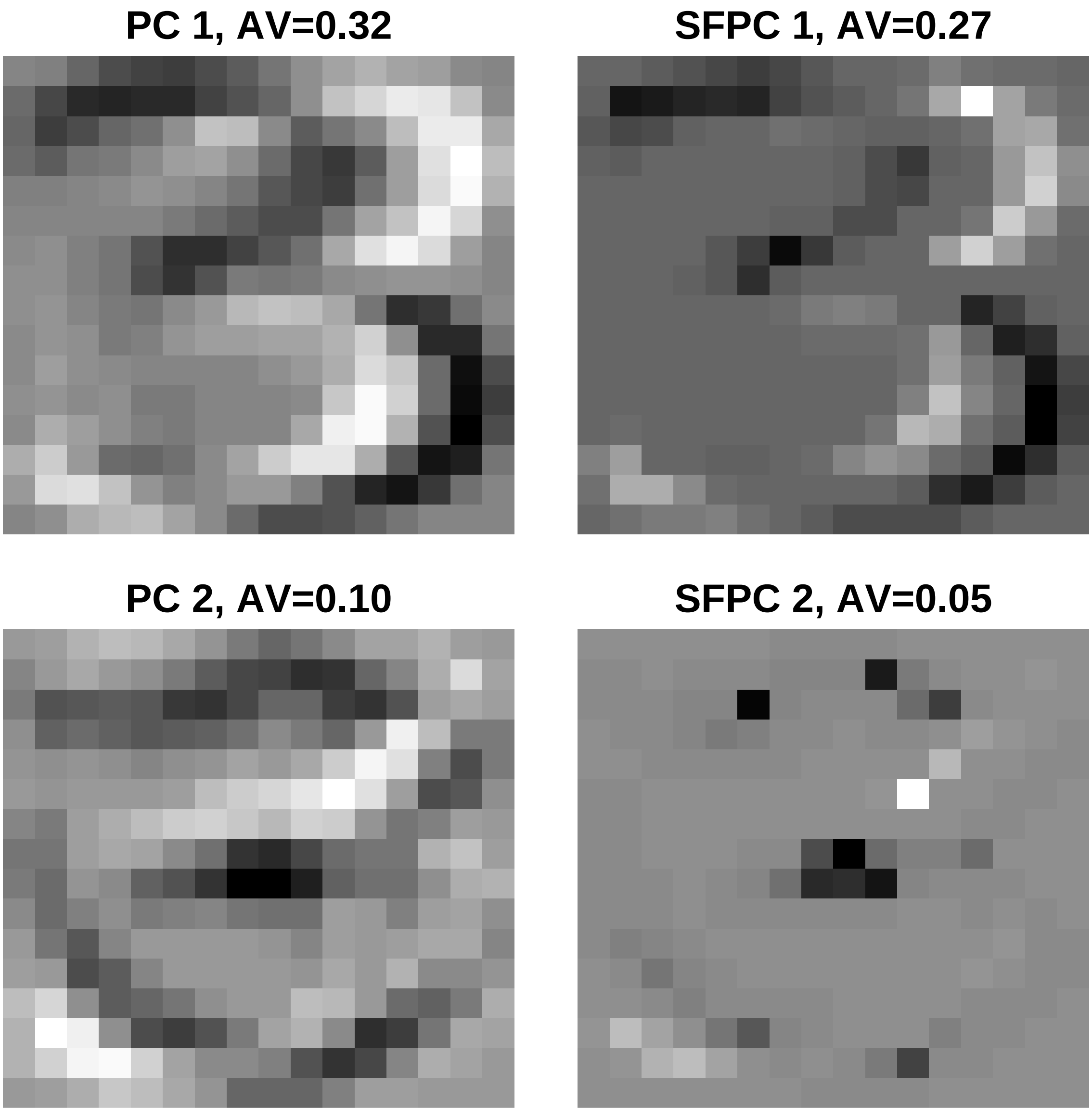


Figure 7.4: The first two eigen-images of digit “3” estimated by PCA and SFPCA, respectively.

APPENDICES

APPENDIX A

Joint Estimation of Multiple Graphical Models

In the beginning, we state some results used in the proof of Theorem II.3 that were established in Theorem 1 of Rothman et al. (2008). We will use the following notation: for a matrix $\boldsymbol{M} = (m_{j,j'})_{p\times p}$, $|\boldsymbol{M}|_1 = \sum_{j,j'} |m_{j,j'}|$, $\boldsymbol{M}^+$ is a diagonal matrix with the same diagonal as $\boldsymbol{M}$, $\boldsymbol{M}^- = \boldsymbol{M} - \boldsymbol{M}^+$, and $\boldsymbol{M}_S$ is $\boldsymbol{M}$ with all elements outside an index set S replaced by zeros. We also write $\widetilde{\boldsymbol{M}}$ for the vectorized $p^2 \times 1$ form of $\boldsymbol{M}$, and $\otimes$ for the Kronecker product of two matrices. In addition, we denote $\boldsymbol{\Sigma}_0^{(k)} = (\boldsymbol{\Omega}_0^{(k)})^{-1}$ as the true covariance matrix of the kth category ($k = 1, \ldots, K$).

Lemma A.1. *Let* $l(\boldsymbol{\Omega}^{(k)}) = \mathrm{trace}(\widehat{\boldsymbol{\Sigma}}^{(k)}\boldsymbol{\Omega}^{(k)}) - \log\{\det(\boldsymbol{\Omega}^{(k)})\}$. *Then for any* $k = 1, \ldots, K$, *the following decomposition holds:*

$$l(\boldsymbol{\Omega}_0^{(k)} + \boldsymbol{\Delta}^{(k)}) - l(\boldsymbol{\Omega}_0^{(k)}) = \mathrm{trace}\{(\widehat{\boldsymbol{\Sigma}}^{(k)} - \boldsymbol{\Sigma}_0^{(k)})\boldsymbol{\Delta}^{(k)}\}$$
$$+(\widetilde{\boldsymbol{\Delta}}^{(k)})^{\mathsf{T}}\{\int_0^1 (1-v)(\boldsymbol{\Omega}_0^{(k)} + v\boldsymbol{\Delta}^{(k)})^{-1} \otimes (\boldsymbol{\Omega}_0^{(k)} + v\boldsymbol{\Delta}^{(k)})^{-1} dv\}\widetilde{\boldsymbol{\Delta}}^{(k)} \quad . \qquad \text{(A.1)}$$

Further, there exist positive constants C_1 *and* C_2 *such that with probability tending to*

1

$$\left|\mathrm{trace}\{(\widehat{\boldsymbol{\Sigma}}^{(k)}-\boldsymbol{\Sigma}_0^{(k)})\boldsymbol{\Delta}^{(k)}\}\right| \;\le\; C_1\Big(\frac{\log p}{n}\Big)^{1/2}|\boldsymbol{\Delta}^{(k)-}|_1+C_2\Big(\frac{p\log p}{n}\Big)^{1/2}\|\boldsymbol{\Delta}^{(k)+}\|_F, \tag{A.2}$$

$$(\widetilde{\boldsymbol{\Delta}}^{(k)})^{\mathsf{T}}\Big\{\int_0^1(1-v)(\boldsymbol{\Omega}_0^{(k)}+v\boldsymbol{\Delta}^{(k)})^{-1}\otimes(\boldsymbol{\Omega}_0^{(k)}+v\boldsymbol{\Delta}^{(k)})^{-1}dv\Big\}\widetilde{\boldsymbol{\Delta}}^{(k)} \;\ge\; \frac{1}{4\tau_2^2}\|\boldsymbol{\Delta}^{(k)}\|_F^2 \tag{A.3}$$

Proof of Theorem II.3

In a slight abuse of notation, we will write $\boldsymbol{\Omega}=(\boldsymbol{\Omega}^{(k)})_{k=1}^K$, $\boldsymbol{\Omega}_0=(\boldsymbol{\Omega}_0^{(k)})_{k=1}^K$, and $\boldsymbol{\Delta}=(\boldsymbol{\Delta}^{(k)})_{k=1}^K$, where $\boldsymbol{\Delta}^{(k)}=(\delta_{j,j'}^{(k)})_{p\times p}$ is defined as $\boldsymbol{\Delta}^{(k)}=\boldsymbol{\Omega}^{(k)}-\boldsymbol{\Omega}_0^{(k)}$ $(k=1,\ldots,K)$. Let $Q(\boldsymbol{\Omega})$ be the objective function of (2.4), and let $G(\boldsymbol{\Delta})=Q(\boldsymbol{\Omega}_0+\boldsymbol{\Delta})-Q(\boldsymbol{\Omega}_0)$. If we take a closed bounded convex set $\mathcal{A}$ which contains 0, and show that G is strictly positive everywhere on the boundary $\partial\mathcal{A}$, then it implies that G has a local minimum inside $\mathcal{A}$, since G is continuous and $G(\mathbf{0})=0$. Specifically, we define $\mathcal{A}=\{\boldsymbol{\Delta}:(\sum_{k=1}^K\|\boldsymbol{\Delta}^{(k)}\|_F)\le Mr_n\}$, with boundary $\partial\mathcal{A}=\{\boldsymbol{\Delta}:(\sum_{k=1}^K\|\boldsymbol{\Delta}^{(k)}\|_F)=Mr_n\}$, where M is a positive constant and $r_n=\{(p+q)(\log p)/n\}^{1/2}$.

By the decomposition (A.1) in Lemma A.1, we can write $G(\boldsymbol{\Delta})=I_1+I_2+I_3+I_4$, where

$$\begin{aligned}
I_1 &= \sum_{k=1}^K\mathrm{trace}\{(\widehat{\boldsymbol{\Sigma}}^{(k)}-\boldsymbol{\Sigma}_0^{(k)})\boldsymbol{\Delta}^{(k)}\}\\
I_2 &= \sum_{k=1}^K(\widetilde{\boldsymbol{\Delta}}^{(k)})^{\mathsf{T}}\Big\{\int_0^1(1-v)(\Omega_0^{(k)}+v\boldsymbol{\Delta}^{(k)})^{-1}\otimes(\Omega_0^{(k)}+v\boldsymbol{\Delta}^{(k)})^{-1}dv\Big\}\widetilde{\boldsymbol{\Delta}}^{(k)}\\
I_3 &= \lambda\sum_{(j,j')\in T^c}(\sum_{k=1}^K|\delta_{j,j'}^{(k)}|)^{1/2}\\
I_4 &= \lambda\sum_{j\neq j':(j,j')\in T}\{(\sum_{k=1}^K|\omega_{j,j'}^{(k)}|)^{1/2}-(\sum_{k=1}^K|\omega_{0,j,j'}^{(k)}|)^{1/2}\}
\end{aligned}$$

We first consider I_1. By applying inequality (A.2) in Lemma A.1, we have $|I_1|\le$

$I_{1,1}+I_{1,2}$, where $I_{1,1}=C_1\{(\log p)/n\}^{1/2}\sum_{k=1}^K|\mathbf{\Delta}_T^{(k)-}|_1+C_2\{(p\log p)/n\}^{1/2}\sum_{k=1}^K\|\mathbf{\Delta}^{(k)+}\|_F$ and $I_{1,2}=C_1\{(\log p)/n\}^{1/2}\sum_{k=1}^K|\mathbf{\Delta}_{T^c}^{(k)-}|_1$. By applying the bound $|\mathbf{\Delta}_T^{(k)-}|_1\le q_k^{1/2}\|\mathbf{\Delta}_T^{(k)-}\|_F$, we have

$$\begin{aligned} I_{1,1} &\le C_1\Big(\frac{q\log p}{n}\Big)^{1/2}\sum_{k=1}^K\|\mathbf{\Delta}_T^{(k)-}\|_F+C_2\Big(\frac{p\log p}{n}\Big)^{1/2}\sum_{k=1}^K\|\mathbf{\Delta}^{(k)+}\|_F \\ &\le (C_1+C_2)\Big\{\frac{(p+q)\log p}{n}\Big\}^{1/2}\sum_{k=1}^K\|\mathbf{\Delta}^{(k)}\|_F\le M(C_1+C_2)\frac{(p+q)\log p}{n}\end{aligned}$$

on the boundary $\partial\mathcal{A}$.

Next, since for r_n small enough we have $I_3\ge\lambda\sum_{k=1}^K|\mathbf{\Delta}_{T^c}^{(k)-}|_1$, the term $I_{1,2}$ is dominated by the positive term I_3:

$$I_3-I_{1,2}\ge\lambda\sum_{k=1}^K|\mathbf{\Delta}_{T^c}^{(k)-}|_1-C_1\Big(\frac{\log p}{n}\Big)^{1/2}\sum_{k=1}^K|\mathbf{\Delta}_{T^c}^{(k)-}|_1\ge(\Lambda_1-C_1)\Big(\frac{\log p}{n}\Big)^{1/2}\sum_{k=1}^K|\mathbf{\Delta}_{T^c}^{(k)-}|_1$$

The last inequality uses the condition $\lambda\ge\Lambda_1\{(\log p)/n\}^{1/2}$. Therefore, $I_3-I_{1,2}\ge0$ when Λ_1 is large enough. Next we consider I_2. By applying inequality (A.3) in Lemma A.1, we have $I_2\ge(1/4\tau_2^2)\sum_{k=1}^K\|\mathbf{\Delta}^{(k)}\|_F^2\ge\{M^2/(8\tau_2^2)\}\{(p+q)(\log p)/n\}$. Finally consider the remaining term I_4. Using condition (B), we have

$$\begin{aligned} |I_4| &\le \lambda\sum_{j\ne j':(j,j')\in T}\frac{\sum_{k=1}^K\Big||\omega_{j,j'}^{(k)}|-|\omega_{0,j,j'}^{(k)}|\Big|}{\big(\sum_{k=1}^K|\omega_{j,j'}^{(k)}|\big)^{1/2}+\big(\sum_{k=1}^K|\omega_{0,j,j'}^{(k)}|\big)^{1/2}} \\ &\le \frac{\lambda}{\tau_3^{1/2}}\sum_{k=1}^K\sum_{j\ne j':(j,j')\in T}|\omega_{j,j'}^{(k)}-\omega_{0,j,j'}^{(k)}|\le\frac{\lambda}{\tau_3^{1/2}}q^{1/2}\sum_{k=1}^K\|\mathbf{\Delta}^{(k)}\|_F\le\frac{M\Lambda_2}{\tau_3^{1/2}}\frac{(p+q)(\log p)}{n}\ .\end{aligned}$$

The last inequality uses the condition $\lambda\le\Lambda_2\{(1+p/q)(\log p)/n\}^{1/2}$. Putting everything together and using $I_2>0$ and $I_3-I_{1,2}>0$, we have

$$G(\mathbf{\Delta})\ge I_2-I_{1,1}-|I_4|\ge M^2\frac{(p+q)\log p}{n}\Big(\frac{1}{8\tau_2^2}-\frac{C_1+C_2+\Lambda_2/\tau_3^{1/2}}{M}\Big)\ .$$

Thus for M sufficiently large, we have $G(\mathbf{\Delta}) > 0$ for any $\mathbf{\Delta} \in \partial\mathcal{A}$.

□

Proof of Theorem II.4

It suffices to show that for all $(j,j') \in T_k^c$ $(k = 1, \ldots, K)$, the derivative $\partial Q/\partial\omega_{j,j'}^{(k)}$ at $\widehat{\omega}_{j,j'}^{(k)}$ has the same sign as $\widehat{\omega}_{j,j'}^{(k)}$ with probability tending to 1. To see that, suppose that for some $(j,j') \in T_k^c$, the estimate $\widehat{\omega}_{j,j'}^{(k)} \neq 0$. Without loss of generality, suppose $\widehat{\omega}_{j,j'}^{(k)} > 0$. Then there exists $\xi > 0$ such that $\widehat{\omega}_{j,j'}^{(k)} - \xi > 0$. Since $\widehat{\mathbf{\Omega}}$ is a local minimizer of $Q(\mathbf{\Omega})$, we have $\partial Q/\partial\omega_{j,j'}^{(k)} < 0$ at $\widehat{\omega}_{j,j'}^{(k)} - \xi$ for ξ small, contradicting the claim $\partial Q/\partial\omega_{j,j'}^{(k)}$ at $\widehat{\omega}_{j,j'}^{(k)}$ has the same sign as $\widehat{\omega}_{j,j'}^{(k)}$.

The derivative of the objective function can be written as

$$\frac{\partial Q}{\partial\omega_{j,j'}^{(k)}} = 2\{\alpha_{j,j'}^{(k)} + \beta_{j,j'}\text{sgn}(\omega_{j,j'}^{(k)})\} \ , \tag{A.4}$$

where $\alpha_{j,j'}^{(k)} = \widehat{\sigma}_{j,j'}^{(k)} - \sigma_{j,j'}^{(k)}$ and $\beta_{j,j'} = \lambda/(\sum_{k=1}^K |\omega_{j,j'}^{(k)}|)^{1/2}$. Arguing as in Theorem 2 of Lam and Fan (2009), one can show that $\max_{k=1,\ldots,K} \max_{j,j'} |\alpha_{j,j'}^{(k)}| = O_{\P}[\{(\log p)/n\}^{1/2} + \eta_n^{1/2}]$. On the other hand, by Theorem II.3, we have $\sum_{k=1}^K |\omega_{j,j'}^{(k)} - \omega_{0,j,j'}^{(k)}| \leq \sum_{k=1}^K \|\mathbf{\Omega}^{(k)} - \mathbf{\Omega}_0^{(k)}\|_F = O_{\P}(\eta_n) = o(1)$. Then for any $\epsilon > 0$ and large enough n we have $\sum_{k=1}^K |\omega_{j,j'}^{(k)}| \leq \sum_{k=1}^K |\omega_{0,j,j'}^{(k)}| + \epsilon$. Then we have $|\beta_{j,j'}| \geq \lambda/(1 + \sum_{k=1}^K |\omega_{0,j,j'}^{(k)}|)^{-1/2}$. By assumption, $\{(\log p)/n\}^{1/2} + \eta_n^{1/2} = O(\lambda)$, and thus the term $\beta_{j,j'}$ dominates $\alpha_{j,j'}^{(k)}$ in (A.4) for any $(j,j') \in T_k^c$ $(k = 1, \ldots, K)$. Therefore, $\text{sgn}\{(\partial Q/\partial\omega_{j,j'}^{(k)})|_{\omega_{j,j'}^{(k)} = \widehat{\omega}_{j,j'}^{(k)}}\} = \text{sgn}(\widehat{\omega}_{j,j'}^{(k)})$.

□

APPENDIX B

Asymptotic Properties of the Joint Neighborhood Selection Method for Estimating Categorical Markov Networks

The proof of our main result is divided into many steps; Appendix I presents the main idea of the proof by listing the important propositions and the proofs of Theorems III.1 and III.2, whereas Appendix II contains additional technical lemmas and proofs of the propositions. The proof bears some similarities to the proof of Ravikumar et al. (2009) for the neighborhood selection method, who in turn adapted the proof from Meinshausen and Buhlmann (2006) to binary data; however, there are also important differences, since all conditions and results are for joint estimation, and many of our bounds need to be more precise than those given by Ravikumar et al. (2009).

The main idea of the proof is as follows. First, we introduce a restricted version of criterion (3.9), where S is assumed known and all parameters in S^c are set to zero:

$$\widetilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}^{[S]}} \; l(\boldsymbol{\theta}^{[S]}) - \lambda \sum_{(j,j')\in S} |\theta_{j,j'}|. \tag{B.1}$$

Further, we introduce sample versions of conditions (A) and (B) as follows (see below for detailed definitions of $\boldsymbol{Q}^n$ and $\boldsymbol{U}^n$, the sample analogues of the population quantities $\boldsymbol{Q}^0$ and $\boldsymbol{U}^0$):

(A′) Dependency (sample): There exist positive constants $\tau_{\min}$ and $\tau_{\max}$ such that

$$\Lambda_{\min}(\boldsymbol{Q}^n_{S,S}) \geq \tau_{\min} \quad \text{and} \quad \Lambda_{\max}(\boldsymbol{U}^n_{S,S}) \leq \tau_{\max}. \tag{B.2}$$

(B′) Incoherence (sample): There exists a constant $\tau \in (0,1)$ such that

$$\|\boldsymbol{Q}^n_{S^c,S}(\boldsymbol{Q}^n_{S,S})^{-1}\|_\infty \leq 1 - \tau. \tag{B.3}$$

The proof consists of the following steps. Proposition 2 and Proposition 3 show that, under sample regularity conditions (A′) and (B′), the conclusions of Theorems 1 and 2 hold for the solution of the restricted problem (B.1), respectively. Next, Proposition 4 and Proposition 5 prove that the population regularity conditions (A) and (B) give rise to their sample counterparts (A′) and (B′) with probability tending to 1. Proposition 6 gives the Karush-Kuhn-Tucker (KKT) conditions for the full problem (3.9), and Proposition 7 shows that, with probability tending to 1, the solution of the restricted problem (B.1) satisfies the KKT conditions of (3.9). Thus, the solution of the restricted problem is also the solution of the original problem with probability tending to 1 and both theorems hold.

We start by introducing additional notation. Denote the log-likelihood for the i-th observation by

$$l_i(\boldsymbol{\theta}) = \sum_{j=1}^{p} x_{i,j}\Big(\sum_{k \neq j} \theta_{j,k} x_{i,k}\Big) - \log\Big\{1 + \exp\Big(\sum_{k \neq j} \theta_{j,k} x_{i,k}\Big)\Big\}\,, \tag{B.4}$$

The first derivative of the log-likelihood is $\nabla l_i(\boldsymbol{\theta}) = (\nabla_{1,2} l_i(\boldsymbol{\theta}), \ldots, \nabla_{p-1,p} l_i(\boldsymbol{\theta}))^\top$, where

$$\begin{aligned}\nabla_{j,j'} l_i(\boldsymbol{\theta}) &= x_{i,j'}\Big\{x_{i,j} - \frac{\exp(\sum_{k\neq j}\theta_{j,k}x_{i,k})}{1+\exp(\sum_{k\neq j}\theta_{j,k}x_{i,k})}\Big\} \\ &\quad + x_{i,j}\Big\{x_{i,j'} - \frac{\exp(\sum_{k\neq j'}\theta_{j',k}x_{i,k})}{1+\exp(\sum_{k\neq j'}\theta_{j',k}x_{i,k})}\Big\} \ . \end{aligned} \tag{B.5}$$

The second derivative of $l_i(\boldsymbol{\theta})$ is given by

$$\nabla^2 l_i(\boldsymbol{\theta}) = -\boldsymbol{\mathcal{X}}^{(i)^\intercal}\boldsymbol{\eta}^{(i)}(\boldsymbol{\theta})\boldsymbol{\mathcal{X}}^{(i)} \ , \tag{B.6}$$

where $\boldsymbol{\eta}^{(i)}(\boldsymbol{\theta}) = \mathrm{diag}(\eta_1^{(i)}(\boldsymbol{\theta}), \dots, \eta_p^{(i)}(\boldsymbol{\theta}))$ is a $p \times p$ diagonal matrix, and

$$\eta_j^{(i)}(\boldsymbol{\theta}) = \frac{\exp(\sum_{k\neq j}\theta_{j,k}x_{i,k})}{\{1+\exp(\sum_{k\neq j}\theta_{j,k}x_{i,k})\}^2} \ . \tag{B.7}$$

The first derivative of $\eta_j^{(i)}(\boldsymbol{\theta})$ is given by $\nabla\eta_j^{(i)}(\boldsymbol{\theta}) = \xi_j^{(i)}(\boldsymbol{\theta})(\boldsymbol{\mathcal{X}}^{(i,j)})^\intercal$, where

$$\xi_j^{(i)}(\boldsymbol{\theta}) = \frac{\exp(\sum_{k\neq j}\theta_{j,k}x_{i,k})[1-\exp(\sum_{k\neq j}\theta_{j,k}x_{i,k})]}{[1+\exp(\sum_{k\neq j}\theta_{j,k}x_{i,k})]^3} \ . \tag{B.8}$$

It is easy to check that $|\nabla_{j,j'}l_i(\boldsymbol{\theta})| \leq 2$, $|\eta_j^{(i)}(\boldsymbol{\theta})| \leq 1$ and $|\xi_j^{(i)}(\boldsymbol{\theta})| \leq 1$. For n observations, the log-likelihood, its first derivative and its second derivative are $l(\boldsymbol{\theta}) = 1/n\sum_{i=1}^n l_i(\boldsymbol{\theta})$, $\nabla l(\boldsymbol{\theta}) = 1/n\sum_{i=1}^n \nabla l_i(\boldsymbol{\theta})$, and $\nabla^2 l(\theta) = 1/n\sum_{i=1}^n \nabla^2 l_i(\theta)$, respectively. Then, the population Fisher information matrix of (3.9) at $\boldsymbol{\theta}^0$ can be represented as $\boldsymbol{Q}^0 = \mathrm{E}[\boldsymbol{\mathcal{X}}^{(i)^\intercal}\boldsymbol{\eta}^{(i)}(\boldsymbol{\theta}^0)\boldsymbol{\mathcal{X}}^{(i)}]$, and its sample counterpart $\boldsymbol{Q}^n = -\nabla^2 l(\boldsymbol{\theta}^0) = 1/n\sum_{i=1}^n \boldsymbol{\mathcal{X}}^{(i)^\intercal}\boldsymbol{\eta}^{(i)}(\boldsymbol{\theta}^0)\boldsymbol{\mathcal{X}}^{(i)}$. We also define $\boldsymbol{U}^n = 1/n\sum_{i=1}^n \boldsymbol{\mathcal{X}}^{(i)^\intercal}\boldsymbol{\mathcal{X}}^{(i)}$ as the sample counterpart of $\boldsymbol{U}^0 = \mathrm{E}(\boldsymbol{\mathcal{X}}^\intercal\boldsymbol{\mathcal{X}})$ defined in Section 2.4. Let $\mathcal{W}$ be any subset of the index set $\{1, 2, \dots, p(p-1)/2\}$. For any vector $\boldsymbol{\gamma}$, we define $\boldsymbol{\gamma}_{\mathcal{W}}$ as the vector consisting of the elements of $\boldsymbol{\gamma}$ associated with $\mathcal{W}$. Similarly, we define $\boldsymbol{\mathcal{X}}_{\mathcal{W}}^{(i)}$ as the columns of $\boldsymbol{\mathcal{X}}^{(i)}$ associated with $\mathcal{W}$, respectively. Finally, we write $\boldsymbol{\delta} = \boldsymbol{\theta} - \boldsymbol{\theta}^0$, $\widetilde{\boldsymbol{\delta}} = \widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0$ and $\widehat{\boldsymbol{\delta}} = \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0$.

Proposition 2. *Suppose the sample conditions (A′) and (B′) hold. If the tuning parameter* $\lambda = C_\lambda\sqrt{(\log p)/n}$ *for some constant* $C_\lambda > 16(2-\tau)/\tau$ *and* $q\sqrt{(\log p)/n} = o(1)$*, then with probability tending to 1, the optimizer of the restricted criterion* $\widetilde{\boldsymbol{\theta}}$ *satisfies*

$$\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\|_2 \leq M\sqrt{\frac{q\log p}{n}} \tag{B.9}$$

for some constant $M > (2C_\lambda/\tau_{min})\{1 + \tau/(8-4\tau)\}$.

Proposition 3. *Under conditions of Proposition 2, if we further assume* $\theta^0_{min} \geq 2M\sqrt{q(\log p)/n}$*, then with probability tending to 1,* $\widetilde{\theta}_{j,j'} \neq 0$ *for all* $(j,j') \in S$ *and* $\widetilde{\theta}_{j,j'} = 0$ *for all* $(j,j') \in S^c$.

Proposition 4. *(Relationship between sample and population dependency) Suppose the regularity conditions (A) hold, then for any* $\epsilon > 0$,

(i) $\mathrm{P}\{\Lambda_{min}(\boldsymbol{Q}^n_{S,S}) \leq \tau_{min} - \epsilon\} \leq 2\exp\{-(\epsilon^2/2)(n/q^2) + 2\log q\}$*;*

(ii) $\mathrm{P}\{\Lambda_{\max}(\boldsymbol{U}^n_{S,S}) \geq \tau_{\max} + \epsilon\} \leq 2\exp\{-(\epsilon^2/2)(n/q^2) + 2\log q\}$.

Proposition 5. *(Relationship between sample and population incoherence) Suppose the regularity conditions (A) and (B) hold, then for any* $\epsilon > 0$*, there exists a constant* $C = \min\{\tau^2_{min}\tau^2/288(1-\tau)^2, \tau^2_{min}\tau^2/72, \tau_{min}\tau/48\}$*, such that*

$$\mathrm{P}[\|\boldsymbol{Q}^n_{S^c,S}(\boldsymbol{Q}^n_{S,S})^{-1}\|_\infty \geq 1 - \frac{\tau}{2}] \leq 12\exp\Big(-C\frac{n}{q^3} + 4\log p\Big)\ . \tag{B.10}$$

Proposition 6. *(KKT conditions) The sufficient and necessary condition for* $\widehat{\boldsymbol{\theta}}$ *to be a solution of problem* (3.9) *is*

$$\begin{array}{rcll} \nabla_{j,j'}l(\widehat{\boldsymbol{\theta}}) & = & \lambda\mathrm{sgn}(\widehat{\theta}_{j,j'}), & \textit{if } \widehat{\theta}_{j,j'} \neq 0; \\ |\nabla_{j,j'}l(\widehat{\boldsymbol{\theta}})| & < & \lambda, & \textit{if } \widehat{\theta}_{j,j'} = 0. \end{array} \tag{B.11}$$

Moreover, this solution is unique due to the strict convexity of problem (3.9).

Proposition 7. *(The restricted solution satisfies KKT conditions) Under all conditions of Proposition 3, with probability tending to 1, we have,*

(i) $\nabla_{j,j'} l(\widetilde{\boldsymbol{\theta}}) = \lambda \mathrm{sgn}(\widetilde{\theta}_{j,j'})$, *for all* $(j, j') \in S$*;*

(ii) $|\nabla_{j,j'} l(\widetilde{\boldsymbol{\theta}})| < \lambda$, *for all* $(j, j') \in S^c$*.*

Proof of Theorem III.1. The condition $n > (4/C)q^3 \log(p)$ implies $q\sqrt{(\log p)/n} = o(1)$. In addition, since $n > (4/C)q^3 \log(p)$, we have $-(\epsilon^2/2)(n/q^2) + 2\log q \to -\infty$ and $-Cn/q^3 + 4\log(p)] \to -\infty$. Thus, by Propositions 4 and 5, the sample dependency and incoherence conditions (A$'$) and (B$'$) hold with probability 1. Therefore, Proposition 2 holds and, with probability tending to 1, the solution of the restricted problem (B.1) satisfies parameter estimation consistency.

On the other hand, Proposition 7 shows that, with probability tending to 1, the solution of the restricted problem $\widetilde{\boldsymbol{\theta}}$ satisfies the KKT conditions in Proposition 6. Since the criterion (3.9) is strictly convex, we conclude $\widetilde{\boldsymbol{\theta}}$ is the unique solution of (3.9), i.e., $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}$. This proves Theorem III.1.

□

Proof of Theorem III.2 is analogous to Proof of Theorem III.1 and is omitted.

Appendix II: Proofs of Propositions

This appendix contains several additional technical lemmas and proofs of Propositions 1-6.

Lemma B.1. *[Bound on $\nabla l(\boldsymbol{\theta}^0)$] With probability tending to 1,* $\|\nabla l(\boldsymbol{\theta}^0)\|_\infty \le C_\nabla \sqrt{(\log p)/n}$ *for some constant* $C_\nabla > 4$*.*

Proof of Lemma B.1: Note that $E[\nabla l_i(\boldsymbol{\theta}^0)] = \mathbf{0}$, $1 \le i \le n$ and $|\nabla_{j,j'} l_i(\theta_0)| \le 2$, $1 \le i \le n, 1 \le j < j' \le p$. By applying the Azuma-Hoeffding inequality (Hoeffding, 1963), we get

$$P[|\nabla_{j,j'} l(\boldsymbol{\theta}^0)| \geq t] \leq 2\exp(-nt^2/8). \tag{B.12}$$

Letting $t = C_\nabla \sqrt{(\log p)/n}$ for some constant $C_\nabla > 0$, we obtain

$$P\left[|\nabla_{j,j'} l(\boldsymbol{\theta}^0)| \geq C_\nabla \sqrt{\frac{\log p}{n}}\right] \leq 2\exp(-C_\nabla^2 \log p/8)\ . \tag{B.13}$$

Then, by the union-sum inequality we have

$$P[\|\nabla l(\boldsymbol{\theta}^0)\|_\infty \geq C_\nabla \sqrt{\frac{\log p}{n}}] \leq 2\exp(-C_\nabla^2 \log p\,/8 + 2\log p). \tag{B.14}$$

Setting $C_\nabla > 4$ establishes the lemma. □

Lemma B.2. *[Bound on $-\boldsymbol{\delta}_S{}^\intercal[\nabla^2 l(\boldsymbol{\theta}^0 + \alpha\boldsymbol{\delta}^{[S]})]_{S,S}\boldsymbol{\delta}_S$]m If the sample dependency condition (A′) holds and $q\sqrt{(\log p)/n} = o(1)$, then for any $\alpha \in [0,1]$, with probability tending to 1,*

$$-\boldsymbol{\delta}_S{}^\intercal[\nabla^2 l(\boldsymbol{\theta}^0 + \alpha\boldsymbol{\delta}^{[S]})]_{S,S}\boldsymbol{\delta}_S \geq \frac{1}{2}\tau_{\min}\|\boldsymbol{\delta}_S\|_2^2\ . \tag{B.15}$$

Proof of Lemma B.2: Applying the mean value theorem, we have $\eta_j(\boldsymbol{\theta}^0 + \alpha\boldsymbol{\delta}^{[S]}) =$

$\eta_j(\boldsymbol{\theta}^0) + \alpha \nabla \eta_j(\boldsymbol{\theta}^0 + \alpha^* \boldsymbol{\delta}^{[S]})^\intercal \boldsymbol{\delta}^{[S]}$, for some constant $\alpha^* \in (0, \alpha)$. Then, we have

$$\begin{aligned}
& -\boldsymbol{\delta}_S{}^\intercal [\nabla^2 l(\boldsymbol{\theta}^0 + \alpha \boldsymbol{\delta}^{[S]})]_{S,S} \boldsymbol{\delta}_S \\
= \;& \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{\mathcal{X}}_S^{(i)} \boldsymbol{\delta}_S)^\intercal \boldsymbol{\eta}(\boldsymbol{\theta}^0 + \alpha \boldsymbol{\delta}^{[S]}) (\boldsymbol{\mathcal{X}}_S^{(i)} \boldsymbol{\delta}_S) \\
= \;& \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{p} \eta_j(\boldsymbol{\theta}^0) (\boldsymbol{\mathcal{X}}_S^{(i,j)} \boldsymbol{\delta}_S)^2 \\
& + \frac{\alpha}{n} \sum_{i=1}^{n} \sum_{j=1}^{p} \nabla \eta_j(\boldsymbol{\theta}^0 + \alpha^* \boldsymbol{\delta}^{[S]})^\intercal \boldsymbol{\delta}^{[S]} (\boldsymbol{\mathcal{X}}_S^{(i,j)} \boldsymbol{\delta}_S)^2 \\
\geq \;& -\boldsymbol{\delta}_S{}^\intercal [\nabla^2 l(\boldsymbol{\theta}^0)]_{S,S} \boldsymbol{\delta}_S \\
& - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{p} |\xi_j^{(i)}(\boldsymbol{\theta}^0 + \alpha^* \boldsymbol{\delta}^{[S]})| |\boldsymbol{\mathcal{X}}_S^{(i,j)} \boldsymbol{\delta}_S| (\boldsymbol{\mathcal{X}}_S^{(i,j)} \boldsymbol{\delta}_S)^2 \; . \qquad \text{(B.16)}
\end{aligned}$$

The first term is bounded from below by

$$-\boldsymbol{\delta}_S{}^\intercal [\nabla^2 l(\boldsymbol{\theta}^0)]_{S,S} \boldsymbol{\delta}_S \geq \Lambda_{min}(\boldsymbol{Q}_{S,S}^n) \|\boldsymbol{\delta}_S\|_2^2 \geq \tau_{min} \|\boldsymbol{\delta}_S\|_2^2 \; . \qquad \text{(B.17)}$$

To bound the second term, notice that $|\boldsymbol{\mathcal{X}}_S^{(i,j)} \boldsymbol{\delta}_S| \leq \|\boldsymbol{\mathcal{X}}_S^{(i,j)}\|_\infty \|\boldsymbol{\delta}_S\|_1 \leq \|\boldsymbol{\delta}_S\|_1$ and recall that $|\xi_j^{(i)}| \leq 1$. Then the second term is bounded from above by

$$\|\boldsymbol{\delta}_S\|_1 \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{p} (\boldsymbol{\mathcal{X}}_S^{(i,j)} \boldsymbol{\delta}_S)^2 \leq \tau_{\max} \|\boldsymbol{\delta}_S\|_1 \|\boldsymbol{\delta}_S\|_2^2 \leq (\tau_{min}/2) \|\boldsymbol{\delta}_S\|_2^2 \; , \qquad \text{(B.18)}$$

since $\|\boldsymbol{\delta}_S\|_1 \leq \sqrt{q} \|\boldsymbol{\delta}_S\|_2 = Mq\sqrt{(\log p)/n} = o(1)$ and thus when n is large enough, $\|\boldsymbol{\delta}_S\|_1 \leq \tau_{min}/(2\tau_{\max})$. Putting (B.17) and (B.18) together establishes the lemma. □

Proof of Proposition 2: The proof relies on the convex function proof method from Rothman et al. (2008). Define

$$\mathrm{G}(\boldsymbol{\delta}_S) = -[l(\boldsymbol{\theta}^0 + \boldsymbol{\delta}^{[S]}) - l(\boldsymbol{\theta}^0)] + \lambda(\|\boldsymbol{\theta}^0 + \boldsymbol{\delta}^{[S]}\|_1 - \|\boldsymbol{\theta}^0\|_1). \qquad \text{(B.19)}$$

It can be seen from (B.1) that $\widetilde{\boldsymbol{\delta}}_S = \widetilde{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_S^0$ minimizes $\mathrm{G}(\boldsymbol{\delta}_S)$. Moreover, $\mathrm{G}(\mathbf{0}_S) = 0$, thus we must have $\mathrm{G}(\widetilde{\boldsymbol{\delta}}_S) \leq 0$. If we take a ball $\mathcal{A}$ which contains $\mathbf{0}_S$, and show that G is strictly positive everywhere on the boundary $\partial\mathcal{A}$, then it implies that G has a local minimum inside $\mathcal{A}$, since G is continuous and $\mathrm{G}(\mathbf{0}_S) = 0$. Specifically, we define $\mathcal{A} = \{\boldsymbol{\delta}_S : \|\boldsymbol{\delta}_S\|_2 \leq Ma_n\}$, with boundary $\partial\mathcal{A} = \{\boldsymbol{\delta}_S : \|\boldsymbol{\delta}_S\|_2 = Ma_n\}$, for some constant $M > (2/\tau_{\min})[1 + \tau/(8-4\tau)]C_\lambda$ and $a_n = \sqrt{q(\log p)/n}$. For any $\boldsymbol{\delta}_S \in \partial\mathcal{A}$, the Taylor series expansion gives $\mathrm{G}(\boldsymbol{\delta}_S) = I_1 + I_2 + I_3$, where

$$\begin{aligned} I_1 &= -[\nabla l(\boldsymbol{\theta}^0)]_S{}^{\mathsf{T}}\boldsymbol{\delta}_S \ , \\ I_2 &= -\boldsymbol{\delta}_S{}^{\mathsf{T}}[\nabla^2 l(\boldsymbol{\theta}^0 + \alpha\boldsymbol{\delta}^{[S]})]_{S,S}\boldsymbol{\delta}_S, \text{ for some } \alpha \in [0,1] \ , \\ I_3 &= \lambda(\|\boldsymbol{\theta}^0 + \boldsymbol{\delta}^{[S]}\|_1 - \|\boldsymbol{\theta}^0\|_1) = \lambda(\|\boldsymbol{\theta}_S^0 + \boldsymbol{\delta}_S\|_1 - \|\boldsymbol{\theta}_S^0\|_1) \ . \end{aligned} \tag{B.20}$$

Since $C_\lambda > 16(2-\tau)/\tau$, we have $[\tau/(8-4\tau)]C_\lambda > 4$. By Lemma B.1,

$$|I_1| \leq \|[\nabla l(\boldsymbol{\theta}^0)]_S\|_\infty \|\boldsymbol{\delta}_S\|_1 \leq \|[\nabla l(\boldsymbol{\theta}^0)]_S\|_\infty \sqrt{q}\|\boldsymbol{\delta}_S\|_2 \leq \frac{\tau}{8-4\tau} C_\lambda M q \frac{\log p}{n} \ .$$

By Lemma B.2, $I_2 \geq (\tau_{\min}/2)\|\boldsymbol{\delta}_S\|_2^2 = (\tau_{\min}/2)M^2 q(\log p)/n$. Finally, by the triangular inequality $|I_3| \leq \lambda\|\boldsymbol{\delta}_S\|_1 \leq \lambda\sqrt{q}\|\boldsymbol{\delta}_S\|_2 = C_\lambda M q(\log p)/n$. Then we have

$$\mathrm{G}(\boldsymbol{\delta}_S) \geq M^2 \frac{q \log p}{n}\Big(\frac{\tau_{\min}}{2} - \frac{\tau C_\lambda}{4(2-\tau)M} - \frac{C_\lambda}{M}\Big) > 0. \tag{B.21}$$

The last inequality uses the condition $M > 2C_\lambda[1 + \tau/(8-4\tau)]/\tau_{\min}$. Therefore, with probability tending to 1, we have $\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\|_F = \|\widetilde{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_S^0\|_F \leq M\sqrt{(q\log p)/n}$.

□

Proof of Proposition 3: Since $\widetilde{\boldsymbol{\theta}}$ is the solution of the restricted problem (B.1), we have $\widetilde{\boldsymbol{\theta}}_{j,j'} = 0$ for all $(j,j') \in S^c$. To show $\widetilde{\boldsymbol{\theta}}_{j,j'} \neq 0$ for all $(j,j') \in S$, it is sufficient to show

$$\|\widetilde{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_S^0\|_\infty \leq \frac{\theta_{\min}^0}{2}\ , \tag{B.22}$$

because then $|\widetilde{\theta}_{j,j'}| \geq |\widetilde{\theta}_{j,j'}^0| - |\widetilde{\theta}_{j,j'} - \widetilde{\theta}_{j,j'}^0| \geq \theta_{\min}^0/2$ for all $(j,j') \in S$. With probability tending to 1, by Proposition 2 we have

$$\|\widetilde{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_S^0\|_\infty \leq \|\widetilde{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_S^0\|_2 \leq M\sqrt{\frac{q(\log p)}{n}}\ .$$

The additional condition $\theta_{\min}^0 \geq 2M\sqrt{q(\log p)/n}$ implies (B.22). □

Lemma B.3. *For any $\epsilon > 0$,*

(i) $P[\|\boldsymbol{Q}_{S^c,S}^n - \boldsymbol{Q}_{S^c,S}^0\|_\infty \geq \epsilon] \leq 2\exp\{-(\epsilon^2/2)(n/q^2) + \log(q) + \log[p(p-1)/2 - q]\}$,

(ii) $P[\|\boldsymbol{Q}_{S,S}^n - \boldsymbol{Q}_{S,S}^0\|_\infty \geq \epsilon] \leq 2\exp\{-(\epsilon^2/2)(n/q^2) + 2\log(q)\}$.

Proof of Lemma B.3: We first prove claim (i). Let $v_{(j,j'),(h,h')}^{(i)}$ be the $[(j,j'),(h,h')]$-th element of matrix $\boldsymbol{\mathcal{X}}^{(i)\mathsf{T}}\boldsymbol{\eta}\boldsymbol{\mathcal{X}}^{(i)} - \boldsymbol{Q}^0$. Note $\mathrm{E}(v_{(j,j'),(h,h')}^{(i)}) = 0$ and $|v_{(j,j'),(h,h')}^{(i)}| \leq 1$, and let $v_{(j,j'),(h,h')} = 1/n\sum_{i=1}^n v_{(j,j'),(h,h')}^{(i)}$. Then

$$\begin{aligned} P[\sum_{(h,h')\in S} |v_{(j,j'),(h,h')}| \geq \epsilon] &\leq \sum_{(h,h')\in S} P[|v_{(j,j'),(h,h')}| \geq \epsilon/q] \\ &\leq q \max_{(h,h')\in S} P[|v_{(j,j'),(h,h')}| \geq \epsilon/q]. \end{aligned} \tag{B.23}$$

Combining the union-sum inequality with (B.23), we have

$$P[\|\boldsymbol{Q}_{S^c,S}^n - \boldsymbol{Q}_{S^c,S}^0\|_\infty \geq \epsilon] \leq q\Big(\frac{p(p-1)}{2} - q\Big) \max_{(h,h')\in S} P[|v_{(j,j'),(h,h')}| \geq \epsilon/q]. \tag{B.24}$$

Then, by the Azuma-Hoeffding inequality (Hoeffding, 1963), we have $P[|v_{(j,j'),(h,h')}| \geq \epsilon/q] \leq 2\exp\{-(\epsilon^2/2)(n/q^2)\}$, and (i) follows. The proof of (ii) is similar. □

Proof of Proposition 4: Note that

$$\begin{aligned}\Lambda_{\min}(\boldsymbol{Q}^n_{S,S}) &= \min_{\|y\|_2=1}[y^\intercal \boldsymbol{Q}^0_{S,S} y + y^\intercal(\boldsymbol{Q}^n_{S,S} - \boldsymbol{Q}^0_{S,S})y] \\ &\geq \Lambda_{\min}(\boldsymbol{Q}^0_{S,S}) - \|\boldsymbol{Q}^n_{S,S} - \boldsymbol{Q}^0_{S,S}\|_2 \geq \tau_{\min} - \|\boldsymbol{Q}^n_{S,S} - \boldsymbol{Q}^0_{S,S}\|_\infty \ .\end{aligned}$$

Now claim (i) follows from Lemma B.3 (ii). The proof of claim (ii) is similar. □

Lemma B.4. *Suppose conditions (A) and (B) hold. Then for any $\epsilon > 0$,*

$$P[\|(\boldsymbol{Q}^n_{S,S})^{-1} - (\boldsymbol{Q}^0_{S,S})^{-1}\|_\infty \geq \epsilon] \leq 4\exp\{-(\tau_{\min}\epsilon^2/8)(n/q^3) + 2\log(q)\}. \qquad \text{(B.25)}$$

Proof of Lemma B.4: Writing $(\boldsymbol{Q}^n_{S,S})^{-1} - (\boldsymbol{Q}^0_{S,S})^{-1} = (\boldsymbol{Q}^0_{S,S})^{-1}(\boldsymbol{Q}^0_{S,S} - \boldsymbol{Q}^n_{S,S})(\boldsymbol{Q}^n_{S,S})^{-1}$ and applying norm inequalities, we have

$$\begin{aligned}\|(\boldsymbol{Q}^n_{S,S})^{-1} - (\boldsymbol{Q}^0_{S,S})^{-1}\|_\infty &\leq \sqrt{q}\|(\boldsymbol{Q}^0_{S,S})^{-1}(\boldsymbol{Q}^0_{S,S} - \boldsymbol{Q}^n_{S,S})(\boldsymbol{Q}^n_{S,S})^{-1}\|_2 \\ &\leq \sqrt{q}\|(\boldsymbol{Q}^0_{S,S})^{-1}\|_2\|\boldsymbol{Q}^0_{S,S} - \boldsymbol{Q}^n_{S,S}\|_\infty\|(\boldsymbol{Q}^n_{S,S})^{-1}\|_2 \\ &\leq \frac{\sqrt{q}}{\tau_{\min}}\|\boldsymbol{Q}^0_{S,S} - \boldsymbol{Q}^n_{S,S}\|_\infty\|(\boldsymbol{Q}^n_{S,S})^{-1}\|_2 \ . \qquad \text{(B.26)}\end{aligned}$$

The last inequality holds because $\|(\boldsymbol{Q}^0_{S,S})^{-1}\|_2 = \{\Lambda_{\min}(\boldsymbol{Q}^0_{S,S})\}^{-1}$. In addition, we have $\|(\boldsymbol{Q}^n_{S,S})^{-1}\|_2 = \{\Lambda_{\min}(\boldsymbol{Q}^n_{S,S})\}^{-1}$. Then by setting $\epsilon = \tau_{\min}/2$ in Proposition 4 (i), we have

$$\begin{aligned}\mathrm{P}\left[\frac{\|(\boldsymbol{Q}^n_{S,S})^{-1}\|_2}{\tau_{\min}} \geq \frac{2}{\tau^2_{\min}}\right] &= \mathrm{P}[\Lambda_{\min}(\boldsymbol{Q}^n_{S,S}) \\ &\leq \frac{\tau_{\min}}{2}] \leq 2\exp(-\frac{\tau^2_{\min}}{8}\frac{n}{q^2} + 2\log q). \qquad \text{(B.27)}\end{aligned}$$

By replacing ϵ in Lemma B.3 (ii) with $\tau_{\min}^2\epsilon/(2\sqrt{q})$, we have

$$\mathrm{P}[\|\boldsymbol{Q}_{S,S}^0 - \boldsymbol{Q}_{S,S}^n\|_\infty \geq \frac{\tau_{\min}^2\epsilon}{2\sqrt{q}}] \leq 2\exp(-\frac{\tau_{\min}^4\epsilon^2}{8}\frac{n}{q^3} + 2\log q) \ . \qquad \text{(B.28)}$$

Finally,

$$\mathrm{P}[\|(\boldsymbol{Q}_{S,S}^n)^{-1} - (\boldsymbol{Q}_{S,S}^0)^{-1}\|_\infty \geq \epsilon] \leq \mathrm{P}[\frac{\|\boldsymbol{Q}_{S,S}^n\|_2}{\tau_{\min}} \geq \frac{2}{\tau_{\min}^2}] + \mathrm{P}[\sqrt{q}\|\boldsymbol{Q}_{S,S}^0 - \boldsymbol{Q}_{S,S}^n\|_\infty \geq \frac{\tau_{\min}^2\epsilon}{2}] \ ,$$

and the lemma follows. □

Proof of Proposition 5: we write $\boldsymbol{Q}_{S^c,S}^n(\boldsymbol{Q}_{S,S}^n)^{-1} = \boldsymbol{T}_1 + \boldsymbol{T}_2 + \boldsymbol{T}_3 + \boldsymbol{T}_4$, where

$$\begin{aligned}
\boldsymbol{T}_1 &= \boldsymbol{Q}_{S^c,S}^0[(\boldsymbol{Q}_{S,S}^n)^{-1} - (\boldsymbol{Q}_{S,S}^0)^{-1}] \ , \\
\boldsymbol{T}_2 &= (\boldsymbol{Q}_{S^c,S}^n - \boldsymbol{Q}_{S^c,S}^0)(\boldsymbol{Q}_{S,S}^0)^{-1} \ , \\
\boldsymbol{T}_3 &= (\boldsymbol{Q}_{S^c,S}^n - \boldsymbol{Q}_{S^c,S}^0)[(\boldsymbol{Q}_{S,S}^n)^{-1} - (\boldsymbol{Q}_{S,S}^0)^{-1}] \ , \\
\boldsymbol{T}_4 &= \boldsymbol{Q}_{S^c,S}^0(\boldsymbol{Q}_{S,S}^0)^{-1} \ .
\end{aligned}$$

To bound $\boldsymbol{T}_1$, we write $\boldsymbol{T}_1 = \boldsymbol{Q}_{S^c,S}^0(\boldsymbol{Q}_{S,S}^0)^{-1}(\boldsymbol{Q}_{S,S}^0 - \boldsymbol{Q}_{S,S}^n)(\boldsymbol{Q}_{S,S}^n)^{-1}$. Thus,

$$\|\boldsymbol{T}_1\|_\infty \leq \|\boldsymbol{Q}_{S^c,S}^0(\boldsymbol{Q}_{S,S}^0)^{-1}\|_\infty\|\boldsymbol{Q}_{S,S}^n - \boldsymbol{Q}_{S,S}^0\|_\infty(\sqrt{q}\|(\boldsymbol{Q}_{S,S}^n)^{-1}\|_2) \ .$$

By condition (B), we have $\|\boldsymbol{Q}_{S^c,S}^0(\boldsymbol{Q}_{S,S}^0)^{-1}\|_\infty \leq 1-\tau$. By setting $\epsilon = \tau_{\min}/2$ in Proposition 4(i), and $\epsilon = \tau_{\min}\tau/(12(1-\tau)\sqrt{q})$ in Lemma B.3(ii), we have

$$\begin{aligned}
& \mathrm{P}[\|\boldsymbol{T}_1\|_\infty \geq \frac{\tau}{6}] \\
\leq \ & \mathrm{P}\left[\|\boldsymbol{Q}_{S,S}^n - \boldsymbol{Q}_{S,S}\|_\infty \geq \frac{\tau_{\min}\tau}{12(1-\tau)\sqrt{q}}\right] + \mathrm{P}\left[\|(\boldsymbol{Q}_{S,S}^n)^{-1}\|_2 \geq \frac{2}{\tau_{\min}}\right] \\
\leq \ & 2\exp\left(-\frac{\tau_{\min}^2\tau^2}{288(1-\tau)^2}\frac{n}{q^3} + 2\log q\right) + 2\exp\left(-\frac{\tau_{\min}^2}{8}\frac{n}{q^2} + 2\log q\right) . \qquad \text{(B.29)}
\end{aligned}$$

To bound $\boldsymbol{T}_2$, we write

$$\|\boldsymbol{T}_2\|_\infty \le \|\boldsymbol{Q}^n_{S^c,S} - \boldsymbol{Q}^0_{S^c,S}\|_\infty \sqrt{q} \|(\boldsymbol{Q}^0_{S,S})^{-1}\|_2 \le \frac{\sqrt{q}}{\tau_{\min}} \|\boldsymbol{Q}^n_{S^c,S} - \boldsymbol{Q}^0_{S^c,S}\|_\infty \ .$$

By setting $\epsilon = \tau_{\min}\tau/(6\sqrt{q})$ in Lemma B.3 (i), we have

$$\begin{aligned} \mathrm{P}[\|\boldsymbol{T}_2\|_\infty \ge \frac{\tau}{6}] &\le \mathrm{P}(\|\boldsymbol{Q}^n_{S^c,S} - \boldsymbol{Q}^0_{S^c,S}\|_\infty \ge \frac{\tau_{\min}\tau}{6\sqrt{q}}) \\ &\le 2\exp\{-\frac{\tau^2_{\min}\tau^2}{72}\frac{n}{q^3} + \log q + \log[p(p-1)/2 - q]\}. \quad \text{(B.30)} \end{aligned}$$

To bound $\boldsymbol{T}_3$, we set $\epsilon = \sqrt{\tau/6}$ in both Lemma B.3 (i) and Lemma B.4, so that

$$\begin{aligned} \mathrm{P}[\|\boldsymbol{T}_3\|_\infty \ge \frac{\tau}{6}] &\le \mathrm{P}[\|\boldsymbol{Q}^n_{S^c,S} - \boldsymbol{Q}^0_{S^c,S}\|_\infty \ge \sqrt{\frac{\tau}{6}}] \\ &\quad + \mathrm{P}[\|(\boldsymbol{Q}^n_{S,S})^{-1} - (\boldsymbol{Q}^0_{S,S})^{-1}\|_\infty \ge \sqrt{\frac{\tau}{6}}] \\ &\le 2\exp\{-\frac{\tau}{12}\frac{n}{q^2} + \log q + \log[p(p-1)/2 - q]\} \\ &\quad + 4\exp\{-\frac{\tau_{\min}\tau}{48}\frac{n}{q^3} + 2\log q\}. \quad \text{(B.31)} \end{aligned}$$

Finally, $\|\boldsymbol{T}_4\|_\infty \le 1-\tau$ by condition (B). Since $\log q \le 2\log p$ and $\log[p(p-1)/2-q] \le 2\log p$, we have

$$\begin{aligned} \mathrm{P}[\|\boldsymbol{Q}^n_{S^c,S}(\boldsymbol{Q}^n_{S,S})^{-1}\|_\infty \ge 1 - \frac{\tau}{2}] &\le \mathrm{P}[\|\boldsymbol{T}_1\|_\infty \ge \frac{\tau}{6}] + \mathrm{P}[\|\boldsymbol{T}_2\|_\infty \ge \frac{\tau}{6}] + \mathrm{P}[\|\boldsymbol{T}_3\|_\infty \ge \frac{\tau}{6}] \\ &\le 12\exp\Big(-C\frac{n}{q^3} + 4\log p\Big) \ , \quad \text{(B.32)} \end{aligned}$$

where $C = \min\{\tau^2_{\min}\tau^2/288(1-\tau)^2, \tau^2_{\min}\tau^2/72, \tau_{\min}\tau/48\}$. □

Lemma B.5. *[Bound on $[\nabla^2 l(\boldsymbol{\theta}^0 + \alpha\boldsymbol{\delta}) - \nabla^2 l(\boldsymbol{\theta}^0)]\boldsymbol{\delta}$] Suppose (A) holds. For any $\alpha \in [0,1]$,*

$$\|[\nabla^2 l(\boldsymbol{\theta}^0 + \alpha\boldsymbol{\delta}) - \nabla^2 l(\boldsymbol{\theta}^0)]\boldsymbol{\delta}\|_\infty \le \tau_{\max}\|\boldsymbol{\delta}_S\|^2_2 \ . \quad \text{(B.33)}$$

Proof of Lemma B.5: We have

$$\begin{aligned}
& |\{[\nabla^2 l(\boldsymbol{\theta}^0 + \alpha\boldsymbol{\delta})]_{(j,j'),S} - [\nabla^2 l(\boldsymbol{\theta}^0)]_{(j,j'),S}\}\boldsymbol{\delta}_S| \\
\leq \; & \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} |\boldsymbol{\mathcal{X}}_{j,j'}^{(i,j)\intercal}||[\eta_j(\boldsymbol{\theta}^0 + \alpha\boldsymbol{\delta}_S) - \eta_j(\boldsymbol{\theta}^0)](\boldsymbol{\mathcal{X}}_S^{(i,j)}\boldsymbol{\delta}_S)| \\
\leq \; & \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} |\xi_j^{(i)}(\boldsymbol{\theta}^0 + \alpha^*\boldsymbol{\delta}_S)|(\boldsymbol{\mathcal{X}}_S^{(i,j)}\boldsymbol{\delta}_S)^2 \leq \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}(\boldsymbol{\mathcal{X}}_S^{(i,j)}\boldsymbol{\delta}_S)^2 \\
\leq \; & \Lambda_{\max}(\boldsymbol{U}^n)\|\boldsymbol{\delta}_S\|_2^2 \leq \tau_{\max}\|\boldsymbol{\delta}_S\|_2^2. \qquad \text{(B.34)}
\end{aligned}$$

Since $\|[\nabla^2 l(\boldsymbol{\theta}^0+\alpha\boldsymbol{\delta})-\nabla^2 l(\boldsymbol{\theta}^0)]\boldsymbol{\delta}\|_\infty = \max_{j<j'} |\{[\nabla^2 l(\boldsymbol{\theta}^0+\alpha\boldsymbol{\delta})]_{(j,j'),S}-[\nabla l(\boldsymbol{\theta}^0)]_{(j,j'),S}\}\boldsymbol{\delta}_S|$, the lemma follows. □

Proof of Proposition 7: By Proposition 3, with probability tending to 1 $\widetilde{\theta}_{j,j'} \neq 0$ for all $(j,j') \in S$. Since $\widetilde{\boldsymbol{\theta}}$ is the maximizer of the restricted problem (B.1), with probability tending to 1, $\nabla_{j,j'} l(\widetilde{\boldsymbol{\theta}}) = \lambda \mathrm{sgn}(\widetilde{\theta}_{j,j'})$ for all $(j,j') \in S$, and claim (i) follows.

To show (ii), let $\boldsymbol{u} = \nabla l(\widetilde{\boldsymbol{\theta}})/\lambda$. By (i), $\|\boldsymbol{u}_S\|_\infty = 1$. In addition, by the mean value theorem we have

$$\lambda\boldsymbol{u} - \nabla l(\boldsymbol{\theta}^0) = \nabla^2 l(\boldsymbol{\theta}^0)\widetilde{\boldsymbol{\delta}} = -\boldsymbol{Q}^n\widetilde{\boldsymbol{\delta}} + \boldsymbol{r}^n \;, \qquad \text{(B.35)}$$

where $\alpha \in (0,1)$ and $\boldsymbol{r}^n = [\nabla^2 l(\boldsymbol{\theta}^0 + \alpha\widetilde{\boldsymbol{\delta}}) - \nabla^2 l(\boldsymbol{\theta}^0)]\widetilde{\boldsymbol{\delta}}$. Decomposing $\boldsymbol{Q}^n$ and using $\widetilde{\boldsymbol{\delta}}_{S^c} = \boldsymbol{0}$, we have

$$\begin{aligned}
\boldsymbol{Q}^n_{S,S}\widetilde{\boldsymbol{\delta}}_S &= -\lambda\boldsymbol{u}_S + [\nabla l(\boldsymbol{\theta}^0)]_S + \boldsymbol{r}^n_S; \qquad \text{(B.36)} \\
\boldsymbol{Q}^n_{S^c,S}\widetilde{\boldsymbol{\delta}}_S &= -\lambda\boldsymbol{u}_{S^c} + [\nabla l(\boldsymbol{\theta}^0)]_{S^c} + \boldsymbol{r}^n_{S^c}. \qquad \text{(B.37)}
\end{aligned}$$

The sample dependency condition implies $\boldsymbol{Q}^n_{S,S}$ is invertible. Thus we can plug (B.36) into (B.37) to obtain

$$\boldsymbol{Q}^n_{S^c,S}(\boldsymbol{Q}^n_{S,S})^{-1}(-\lambda\boldsymbol{u}_S + [\nabla l(\boldsymbol{\theta}^0)]_S + \boldsymbol{r}^n_S) = -\lambda\boldsymbol{u}_{S^c} + [\nabla l(\boldsymbol{\theta}^0)]_{S^c} + \boldsymbol{r}^n_{S^c}\ . \tag{B.38}$$

Extracting $\boldsymbol{u}_{S^c}$, we have

$$\begin{aligned} \|\boldsymbol{u}_{S^c}\|_\infty &\leq \frac{\|[\nabla l(\boldsymbol{\theta}^0)]_{S^c}\|_\infty}{\lambda} + \frac{\|\boldsymbol{r}^n_{S^c}\|_\infty}{\lambda} \\ &\quad + \|\boldsymbol{Q}^n_{S^c,S}(\boldsymbol{Q}^n_{S,S})^{-1}\|_\infty\left(\|\boldsymbol{u}_S\|_\infty + \frac{\|[\nabla l(\boldsymbol{\theta}^0)]_S\|_\infty}{\lambda} + \frac{\|\boldsymbol{r}^n_S\|_\infty}{\lambda}\right) \\ &\leq \frac{\|\nabla l(\boldsymbol{\theta}^0)\|_\infty}{\lambda} + \frac{\|\boldsymbol{r}^n\|_\infty}{\lambda} \\ &\quad + \|\boldsymbol{Q}^n_{S^c,S}(\boldsymbol{Q}^n_{S,S})^{-1}\|_\infty(\|\boldsymbol{u}\|_\infty + \frac{\|\nabla l(\boldsymbol{\theta}^0)\|_\infty}{\lambda} + \frac{\|\boldsymbol{r}^n\|_\infty}{\lambda}) \\ &\leq 1 - \tau + (2-\tau)(\frac{\|\nabla l(\boldsymbol{\theta}^0)\|_\infty}{\lambda} + \frac{\|\boldsymbol{r}^n\|_\infty}{\lambda}). \end{aligned} \tag{B.39}$$

By setting $C_\nabla = \tau(8-4\tau)C_\lambda$ in Lemma B.1, $\|\nabla l(\boldsymbol{\theta}^0)\|_\infty/\lambda \leq \tau/(8-4\tau)$. By Lemma B.5, we have $\|\boldsymbol{r}^n\|_\infty/\lambda \leq \tau_{\max}\|\widetilde{\delta}_S\|_2^2/\lambda \leq (\tau_{\max}M^2/C_\lambda)q\sqrt{\log p/n} \leq \tau/(8-4\tau)$, where the last inequality holds by the condition $q\sqrt{(\log p)/n} = o(1)$ when n is sufficiently large. Thus

$$\|\boldsymbol{u}_{S^c}\|_\infty \leq 1 - \frac{\tau}{2} < 1\ , \tag{B.40}$$

and we have $\|[\nabla l(\widetilde{\boldsymbol{\theta}})]_{S^c}\|_\infty = \lambda\|\boldsymbol{u}_{S^c}\|_\infty < \lambda$. □

APPENDIX C

Estimating Heterogeneous Graphical Models for Discrete Data with an Application to Roll Call Voting

Appendix

The appendix presents the proofs of Theorems III.1 and III.2. The main idea of the proof is closely related to Guo et al. (2010), and some strategies for dealing with the joint estimation are borrowed from Guo et al. (2011).

We introduce notation first. For the k-th category, we define the log-likelihood as

$$l(\boldsymbol{\theta}^{(k)}) = \frac{1}{n_k}\sum_{i=1}^{n_k}\sum_{j=1}^{p}[x_{i,j}^{(k)}(\sum_{j'\neq j}\theta_{j,j'}^{(k)}x_{i,j'}^{(k)}) - \log\{1+\exp(\sum_{j'\neq j}\theta_{j,j'}^{(k)}x_{i,j'}^{(k)})\}] \ ,$$

whose first derivative and second derivative are denoted by $\nabla l(\boldsymbol{\theta}^{(k)})$ and $\nabla^2 l(\boldsymbol{\theta}^{(k)})$, respectively. Note that $\nabla l(\boldsymbol{\theta}^{(k)})$ is a $p(p-1)/2$-dimensional vector and $\nabla^2 l(\boldsymbol{\theta}^{(k)})$ is a $p(p-1)/2 \times p(p-1)/2$ matrix. Then, the population Fisher information matrix of the model in (4.10) at $\overline{\boldsymbol{\theta}}$ can be defined as $\overline{\boldsymbol{Q}}^{(k)} = -\mathrm{E}[\nabla^2 l(\overline{\boldsymbol{\theta}}^{(k)})]$, and its sample counterpart is $\widehat{\boldsymbol{Q}}^{(k)} = -\nabla^2 l(\overline{\boldsymbol{\theta}}^{(k)})$. We also write $\widehat{\boldsymbol{U}}^{(k)} = 1/n\sum_{i=1}^{n}\boldsymbol{\mathcal{X}}_{(i)}^{(k)\top}\boldsymbol{\mathcal{X}}_{(i)}^{(k)}$ for the sample counterpart of $\overline{\boldsymbol{U}}^{(k)}$ defined in Section 4.5. Let $\underline{\boldsymbol{\theta}}^{(k)} = (\underline{\theta}_{1,2}^{(k)}, \ldots, \underline{\theta}_{j,j'}^{(k)}, \ldots, \underline{\theta}_{p-1,p}^{(k)})$ be the same as $\boldsymbol{\theta}^{(k)}$ except that all elements in S_k^c are set to zero and write $\boldsymbol{\delta}^{(k)} = \boldsymbol{\theta}^{(k)} - \overline{\boldsymbol{\theta}}^{(k)}$ and $\underline{\boldsymbol{\delta}}^{(k)} = \underline{\boldsymbol{\theta}}^{(k)} - \overline{\boldsymbol{\theta}}^{(k)}$. Finally, let $\mathcal{W}$ be a subset of the index set $\{1, 2, \ldots, p(p-1)/2\}$. For a $p(p-1)/2$-dimensional vector $\boldsymbol{\beta}$, we define $\boldsymbol{\beta}_{\mathcal{W}}$ as the vector consisting of the elements of $\boldsymbol{\beta}$ associated with $\mathcal{W}$.

Next, we introduce a variant of criterion (4.10) by restricting all true zeros in $\{\boldsymbol{\theta}^{(k)}\}_{k=1}^K$ to be estimated as zero. Specifically, the restricted criterion is formulated as follows:

$$\max_{\{\underline{\boldsymbol{\theta}}^{(k)}\}_{k=1}^K} \sum_{k=1}^K l(\underline{\boldsymbol{\theta}}^{(k)}) - \lambda \sum_{1 \le j < j' \le p} \sqrt{\sum_{k=1}^K |\underline{\theta}_{j,j'}^{(k)}|} \ , \tag{C.1}$$

and its maximizer is denoted by $\{\widehat{\underline{\boldsymbol{\theta}}}^{(k)}\}_{k=1}^K$. In addition, we consider the sample versions of regularity conditions (B) and (C).

(B′) Sample dependency: There exist positive constants $\tau_{\min}$ and $\tau_{\max}$ such that for any $k = 1, \ldots, K$,

$$\Lambda_{\min}(\widehat{\boldsymbol{Q}}_{S_k,S_k}^{(k)}) \ge \tau_{\min} \quad \text{and} \quad \Lambda_{\max}(\widehat{\boldsymbol{U}}_{S_k,S_k}^{(k)}) \le \tau_{\max} \ . \tag{C.2}$$

(C′) Sample incoherence: There exists a constant $\tau \in (1 - \sqrt{\gamma_{\min}/4\gamma_{\max}}, 1)$ such that for any $k = 1, \ldots, K$,

$$\|\widehat{\boldsymbol{Q}}_{S_k^c,S_k}^{(k)} (\widehat{\boldsymbol{Q}}_{S_k,S_k}^{(k)})^{-1}\|_\infty \le 1 - \tau \ . \tag{C.3}$$

For convenience of the readers, the proof of our main result is divided into two parts: Part I presents the main idea of the proof by listing the important propositions and the proofs of Theorems III.1 and III.2, whereas Part II contains additional technical details and proofs of propositions in Part I.

Part I: Propositions and Proof of Theorems III.1 and III.2

The proof consists of the following steps. Proposition 8 shows that, under sample regularity conditions (B$'$) and (C$'$), the conclusions of Theorems III.1 and III.2 hold for the local maximizer of the restricted problem (C.1). Next, Proposition 9 proves that the population regularity conditions (B) and (C) give rise to their sample counterparts (B$'$) and (C$'$) with probability tending to one; hence the conclusions of Proposition 8 also hold with the population regularity conditions. Lastly, we show that the local maximizer of (C.1) is also a local maximizer of the original model (4.10). This is established via Proposition 10, which sets out the Karush-Kuhn-Tucker (KKT) conditions for the local maximizer of criterion (4.10), and Proposition 11, which shows that, with probability tending to one, the local maximizer of (C.1) satisfies these KKT conditions.

Proposition 8. *Suppose condition (A) and the sample conditions (B$'$) and (C$'$) hold. If the tuning parameter $\lambda = C_\lambda\sqrt{(\log p)/n}$ for some constant $C_\lambda > (8-4\tau)\sqrt{\gamma_{\min}}/(1-\tau)$ and $q\sqrt{(\log p)/n} = o(1)$, then with probability tending to one, there exists a local maximizer of the restricted criterion, $\{\widehat{\underline{\boldsymbol{\theta}}}^{(k)}\}_{k=1}^{K}$, satisfying*

(i) $\sum_{k=1}^{K} \|\widehat{\underline{\boldsymbol{\theta}}}^{(k)} - \overline{\boldsymbol{\theta}}^{(k)}\|_2 \leq M\sqrt{q(\log p)/n}$ *for some constant* $M > (2KC_\lambda/\tau_{\min}\sqrt{\gamma_{\min}})[(3-2\tau)/(2-\tau)]$*;*

(ii) *For each $k = 1, \ldots, K$, $\widehat{\underline{\theta}}_{j,j'}^{(k)} \neq 0$ for all $(j, j') \in S_k$ and $\widehat{\underline{\theta}}_{j,j'}^{(k)} = 0$ for all $(j, j') \in S_k^c$.*

Proposition 9. *Suppose the regularity conditions (B) and (C) hold, then for any $\epsilon > 0$, the following inequalities hold with probability tending to one for all $k = 1, \ldots, K$:*

(i) $\mathrm{P}\{\Lambda_{min}(\widehat{\boldsymbol{Q}}_{S_k,S_k}^{(k)}) \leq \tau_{min} - \epsilon\} \leq 2\exp\{-(\epsilon^2/2)(n_k/q_k^2) + 2\log q_k\}$*;*

(ii) $\mathrm{P}\{\Lambda_{\max}(\widehat{\boldsymbol{U}}_{S_k,S_k}^{(k)}) \geq \tau_{\max} + \epsilon\} \leq 2\exp\{-(\epsilon^2/2)(n_k/q_k^2) + 2\log q_k\}$*;*

(iii) $\mathrm{P}[\|\widehat{\boldsymbol{Q}}_{S_k^c,S_k}^{(k)}(\widehat{\boldsymbol{Q}}_{S_k,S_k}^{(k)})^{-1}\|_\infty \geq 1 - \tau/2] \leq 12\exp(-Cn_k/q_k^3 + 4\log p)$*, for some constant* $C = \min\{\tau_{min}^2\tau^2/288(1-\tau)^2, \tau_{min}^2\tau^2/72, \tau_{min}\tau/48\}$*.*

Proposition 10. $\{\widehat{\boldsymbol{\theta}}\}_{k=1}^{K}$ *is a local maximizer of problem* (4.10) *if and only if the following conditions hold for all* $k = 1, \ldots, K$*:*

$$\begin{array}{rcll} \nabla_{j,j'} l(\widehat{\boldsymbol{\theta}}^{(k)}) & = & \lambda \mathrm{sgn}(\widehat{\theta}^{(k)}_{j,j'})/(\sum_{k=1}^{K} |\widehat{\theta}^{(k)}_{j,j'}|)^{1/2}, & \textit{if } \widehat{\theta}^{(k)}_{j,j'} \neq 0; \\ |\nabla_{j,j'} l(\widehat{\boldsymbol{\theta}}^{(k)})| & < & \lambda/(\sum_{k=1}^{K} |\widehat{\theta}^{(k)}_{j,j'}|)^{1/2}, & \textit{if } \widehat{\theta}^{(k)}_{j,j'} = 0. \end{array} \tag{C.4}$$

Proposition 11. *Under all conditions of Proposition 8, with probability tending to one, we have, for each* $k = 1, \ldots, K$*,*

$$\begin{array}{rcll} \nabla_{j,j'} l(\underline{\widehat{\boldsymbol{\theta}}}^{(k)}) & = & \lambda \mathrm{sgn}(\underline{\widehat{\theta}}^{(k)}_{j,j'})/(\sum_{k=1}^{K} |\underline{\widehat{\theta}}^{(k)}_{j,j'}|)^{1/2}, & \textit{for all } (j,j') \in S_k; \\ |\nabla_{j,j'} l(\underline{\widehat{\boldsymbol{\theta}}}^{(k)})| & < & \lambda/(\sum_{k=1}^{K} |\underline{\widehat{\theta}}^{(k)}_{j,j'}|)^{1/2}, & \textit{for all } (j,j') \in S_k^c. \end{array} \tag{C.5}$$

Proof of Theorems III.1 and III.2

The condition $\min\{n/q^3, n_1/q_1^3, \ldots, n_K/q_K^3\} > (4/C)\log p$ implies that, for each $k = 1, \ldots, K$, we have $-Cn_k/q_k^3 + 4\log p < 0$ and $-(\epsilon^2/2)(n_k/q_k^2) + 2\log q_k < 0$ when q_k is large enough. This condition also implies $q\sqrt{(\log p)/n} = o(1)$. In addition, by Proposition 9, the sample conditions (B′) and (C′) hold with probability tending to one when regularity conditions (B) and (C) hold. Therefore, by Proposition 8, with probability tending to one, the solution of the restricted problem $\{\underline{\widehat{\boldsymbol{\theta}}}^{(k)}\}_{k=1}^{K}$ satisfies both parameter estimation consistency and structure selection consistency. On the other hand, by Proposition 11, with probability tending to one, $\{\underline{\widehat{\boldsymbol{\theta}}}^{(k)}\}_{k=1}^{K}$ also satisfies the KKT conditions in Proposition 10, thus it is a local maximizer of criterion (4.10). This proves Theorems III.1 and III.2. □

Part II: Proofs of Propositions

Before proving the propositions, we state a few lemmas which will be used in the proofs. These lemmas are variants of Lemmas 1, 2 and 5 in Guo et al. (2010), adapted to the settings of the heterogenous model and thus the proofs are omitted here. Likewise, the proof of Proposition 9 is very similar to the proof of Propositions 3 and 4 in Guo et al. (2010) and is omitted.

Lemma C.1. *For each $k = 1, \ldots, K$, with probability tending to 1, we have $\|\nabla l(\overline{\boldsymbol{\theta}}^{(k)})\|_\infty \le C_\nabla\sqrt{(\log p)/n}$ for some constant $C_\nabla > 4$.*

Lemma C.2. *If the sample dependency condition (B′) holds and $q\sqrt{(\log p)/n} = o(1)$, then for any $\alpha_k \in [0, 1]$, $k = 1, \ldots, K$, the following inequality holds with probability tending to 1:*

$$-\sum_{k=1}^{K} \boldsymbol{\delta}_{S_k}^{(k)\intercal}[\nabla^2 l(\overline{\boldsymbol{\theta}}^{(k)} + \alpha_k \underline{\boldsymbol{\delta}}^{(k)})]_{S_k,S_k}\boldsymbol{\delta}_{S_k}^{(k)} \ge \frac{1}{2}\tau_{\min}\sum_{k=1}^{K}\|\underline{\boldsymbol{\delta}}^{(k)}\|_2^2\ . \tag{C.6}$$

Lemma C.3. *Suppose the sample dependency condition (B) holds. For any $\alpha_k \in [0, 1]$, $k = 1, \ldots, K$, the following inequality holds with probability tending to one:*

$$\|[\nabla^2 l(\overline{\boldsymbol{\theta}}^{(k)} + \alpha_k \underline{\boldsymbol{\delta}}^{(k)}) - \nabla^2 l(\overline{\boldsymbol{\theta}}^{(k)})]\underline{\boldsymbol{\delta}}^{(k)}\|_\infty \le \tau_{\max}\|\underline{\boldsymbol{\delta}}^{(k)}\|_2^2\ . \tag{C.7}$$

Proof of Proposition 8

The main idea of the proof was first introduced in this context in Rothman et al. (2008) and has since been used by many authors. Define

$$\mathrm{G}(\{\underline{\boldsymbol{\delta}}^{(k)}\}_{k=1}^K) = -\sum_{k=1}^{K}[l(\overline{\boldsymbol{\theta}}^{(k)}+\underline{\boldsymbol{\delta}}^{(k)})-l(\overline{\boldsymbol{\theta}}^{(k)})]+\lambda\sum_{1\le j<j'\le p}\{(\sum_{k=1}^{K}|\overline{\theta}_{j,j'}^{(k)}+\underline{\delta}_{j,j'}^{(k)}|)^{1/2}-(\sum_{k=1}^{K}|\overline{\theta}_{j,j'}^{(k)}|)^{1/2}\}. \tag{C.8}$$

It can be seen from (C.1) that, $\{\widehat{\underline{\boldsymbol{\delta}}}^{(k)}\}_{k=1}^K$ minimizes $\mathrm{G}(\{\underline{\boldsymbol{\delta}}^{(k)}\}_{k=1}^K)$ and $\mathrm{G}(\{\mathbf{0}\}_{k=1}^K) = 0$. Thus we must have $\mathrm{G}(\{\widehat{\underline{\boldsymbol{\delta}}}^{(k)}\}_{k=1}^K) \le 0$. If we take a closed set $\mathcal{A}$ which contains $\{\mathbf{0}\}_{k=1}^K$, and show that G is strictly positive everywhere on the boundary $\partial\mathcal{A}$, then it implies that G has a local minimum inside $\mathcal{A}$, since G is continuous and $\mathrm{G}(\{\mathbf{0}\}_{k=1}^K) = 0$. Specifically, we define $\mathcal{A} = \{\{\underline{\boldsymbol{\delta}}^{(k)}\}_{k=1}^K : \sum_{k=1}^K \|\underline{\boldsymbol{\delta}}^{(k)}\|_2 \le Ma_n\}$, with boundary $\partial\mathcal{A} = \{\{\underline{\boldsymbol{\delta}}^{(k)}\}_{k=1}^K : \sum_{k=1}^K \|\underline{\boldsymbol{\delta}}^{(k)}\|_2 = Ma_n\}$, for some constant $M > (2KC_\lambda/\tau_{\min}\sqrt{\gamma_{\min}})[(3 -$

$2\tau)/(2-\tau)]$ and $a_n = \sqrt{q(\log p)/n}$. For any $\{\underline{\boldsymbol{\delta}}^{(k)}\}_{k=1}^K \in \partial\mathcal{A}$, the Taylor series expansion gives $\mathrm{G}(\{\underline{\boldsymbol{\delta}}^{(k)}\}_{k=1}^K) = I_1 + I_2 + I_3$, where

$$\begin{aligned}
I_1 &= -\sum_{k=1}^K [\nabla l(\overline{\boldsymbol{\theta}}^{(k)})]_{S_k}{}^{\mathsf{T}} \boldsymbol{\delta}_{S_k}^{(k)} \ , \\
I_2 &= -\sum_{k=1}^K \boldsymbol{\delta}_{S_k}^{(k)\mathsf{T}} [\nabla^2 l(\overline{\boldsymbol{\theta}}^{(k)} + \alpha_k \underline{\boldsymbol{\delta}}^{(k)})]_{S_k,S_k} \boldsymbol{\delta}_{S_k}^{(k)}, \text{ for some } \alpha_k \in [0,1] \ , \\
I_3 &= \lambda \sum_{(j,j')\in S_\cup} \{(\sum_{k=1}^K |\overline{\theta}_{j,j'}^{(k)} + \underline{\delta}_{j,j'}^{(k)}|)^{1/2} - (\sum_{k=1}^K |\overline{\theta}_{j,j'}^{(k)}|)^{1/2}\} \ . \qquad \text{(C.9)}
\end{aligned}$$

Since $C_\lambda > (8-4\tau)\sqrt{\gamma_{\min}}/(1-\tau)$, we have $[(1-\tau)/(2-\tau)]C_\lambda/\sqrt{\gamma_{\min}} > 4$. By Lemma C.1,

$$|I_1| \leq \sum_{k=1}^K \|[\nabla l(\overline{\boldsymbol{\theta}}^{(k)})]_{S_k}\|_\infty \|\boldsymbol{\delta}_{S_k}^{(k)}\|_1 \leq [(1-\tau)C_\lambda M \gamma_{\min}^{-1/2}/(2-\tau)](q\log p)/n \ . \quad \text{(C.10)}$$

In addition, by condition $q\sqrt{(\log p)/n} = o(1)$, Lemma C.2 holds and thus

$$I_2 \geq (\tau_{\min}/2) \sum_{k=1}^K \|\underline{\boldsymbol{\delta}}^{(k)}\|_2^2 \geq [\tau_{\min}/(2K)] M^2 q(\log p)/n \ . \qquad \text{(C.11)}$$

Finally, by the triangular inequality and regularity condition (A),

$$\begin{aligned}
|I_3| &\leq \lambda \sum_{(j,j')\in S_\cup} \sum_{k=1}^K \frac{||\overline{\theta}_{j,j'}^{(k)} + \underline{\delta}_{j,j'}^{(k)}| - |\overline{\theta}_{j,j'}^{(k)}||}{(\sum_{k=1}^K |\overline{\theta}_{j,j'}^{(k)} + \underline{\delta}_{j,j'}^{(k)}|)^{1/2} + (\sum_{k=1}^K |\overline{\theta}_{j,j'}^{(k)}|)^{1/2}} \\
&\leq (\lambda \gamma_{\min}^{-1/2}) \sum_{k=1}^K \sum_{(j,j')\in S_\cup} |\underline{\delta}_{j,j'}^{(k)}| \leq (\lambda q^{1/2} \gamma_{\min}^{-1/2}) \sum_{k=1}^K \|\underline{\boldsymbol{\delta}}^{(k)}\|_2 \\
&\leq (M C_\lambda \gamma_{\min}^{-1/2})\{q(\log p)/n\} \qquad \text{(C.12)}
\end{aligned}$$

Then we have

$$\mathrm{G}(\{\underline{\boldsymbol{\delta}}^{(k)}\}_{k=1}^K) \geq M^2 \frac{q\log p}{n}\Big(\frac{\tau_{\min}}{2K} - \frac{(1-\tau)C_\lambda}{(2-\tau)M\gamma_{\min}^{1/2}} - \frac{C_\lambda}{M\gamma_{\min}^{1/2}}\Big) > 0. \qquad \text{(C.13)}$$

The last inequality uses the condition $M > (2KC_{\lambda}/\tau_{\min}\sqrt{\gamma_{\min}})[(3-2\tau)/(2-\tau)]$. Therefore, with probability tending to 1, we have $\sum_{k=1}^{K}\|\underline{\widehat{\boldsymbol{\theta}}}^{(k)} - \overline{\boldsymbol{\theta}}^{(k)}\|_2 \leq M\sqrt{q(\log p)/n}$, and consequently claim (i) in Proposition 8 holds.

On the other hand, by the definition of $\underline{\widehat{\boldsymbol{\theta}}}^{(k)}$, we have $\underline{\widehat{\theta}}_{j,j'}^{(k)} = 0$ for all $(j,j') \in S_k^c$. By regularity condition (A) and Proposition 8 (i), for any $(j,j') \in S_k$, $k = 1,\ldots,K$, we have $|\underline{\widehat{\theta}}_{j,j'}^{(k)}| \geq |\overline{\theta}_{j,j'}^{(k)}| - |\underline{\widehat{\theta}}_{j,j'}^{(k)} - \overline{\theta}_{j,j'}^{(k)}| \geq \gamma_{\min}/2 > 0$, when n is large enough. □

Proof of Proposition 11

By Proposition 8, with probability tending to one, we have $\underline{\widehat{\theta}}_{j,j'} \neq 0$ for all $(j,j') \in S_k$. Since $\{\underline{\widehat{\boldsymbol{\theta}}}^{(k)}\}_{k=1}^{K}$ is a local maximizer of the restricted problem (C.1), with probability tending to one, $\nabla_{j,j'} l(\underline{\widehat{\boldsymbol{\theta}}}^{(k)}) = \lambda \mathrm{sgn}(\underline{\widehat{\theta}}_{j,j'}^{(k)})/(\sum_{k=1}^{K}|\underline{\widehat{\theta}}_{j,j'}^{(k)}|)^{1/2}$, for all $(j,j') \in S_k$.

To show the second claim, we apply the mean value theorem and write $\nabla l(\underline{\widehat{\boldsymbol{\theta}}}^{(k)}) = \nabla l(\overline{\boldsymbol{\theta}}^{(k)}) + \boldsymbol{r}^{(k)} - \widehat{\boldsymbol{Q}}^{(k)}\underline{\widehat{\boldsymbol{\delta}}}^{(k)}$, where $\boldsymbol{r}^{(k)} = \{\nabla^2 l(\overline{\boldsymbol{\theta}}^{(k)} + \alpha_k \underline{\widehat{\boldsymbol{\delta}}}^{(k)}) - \nabla^2 l(\overline{\boldsymbol{\theta}}^{(k)})\}\underline{\widehat{\boldsymbol{\delta}}}^{(k)}$. After some simplifications, we have

$$[\nabla l(\underline{\widehat{\boldsymbol{\theta}}}^{(k)})]_{S_k^c} = [\nabla l(\overline{\boldsymbol{\theta}}^{(k)})]_{S_k^c} + \boldsymbol{r}_{S_k^c}^{(k)} - [\widehat{\boldsymbol{Q}}_{S_k^c,S_k}^{(k)}(\widehat{\boldsymbol{Q}}_{S_k,S_k}^{(k)})^{-1}]\{[\nabla l(\overline{\boldsymbol{\theta}}^{(k)})]_{S_k} + \boldsymbol{r}_{S_k}^{(k)} - [\nabla l(\underline{\widehat{\boldsymbol{\theta}}}^{(k)})]_{S_k}\} \tag{C.14}$$

and thus,

$$\begin{aligned}
\|[\nabla l(\underline{\widehat{\boldsymbol{\theta}}}^{(k)})]_{S_k^c}\|_\infty &\leq \|[\nabla l(\overline{\boldsymbol{\theta}}^{(k)})]_{S_k^c}\|_\infty + \|\boldsymbol{r}_{S_k^c}^{(k)}\|_\infty \\
&\quad + \|\widehat{\boldsymbol{Q}}_{S_k^c,S_k}^{(k)}(\widehat{\boldsymbol{Q}}_{S_k,S_k}^{(k)})^{-1}\|_\infty\{\|[\nabla l(\overline{\boldsymbol{\theta}}^{(k)})]_{S_k}\|_\infty + \|\boldsymbol{r}_{S_k}^{(k)}\|_\infty + \|[\nabla l(\underline{\widehat{\boldsymbol{\theta}}}^{(k)})]_{S_k}\|_\infty\} \\
&\leq (2-\tau)\|\nabla l(\overline{\boldsymbol{\theta}}^{(k)})\|_\infty + (2-\tau)\|\boldsymbol{r}^{(k)}\|_\infty + (1-\tau)\|[\nabla l(\underline{\widehat{\boldsymbol{\theta}}}^{(k)})]_{S_k}\|_\infty \\
&\leq [(1-\tau)C_\lambda/\sqrt{\gamma_{\min}}]\sqrt{(\log p)/n} + (2-\tau)\tau_{\max}M^2 q(\log p)/n \\
&\quad + (1-\tau)\lambda/\min_{(j,j')\in S_k}[\sum_{k=1}^{K}|\underline{\widehat{\theta}}_{j,j'}|]^{1/2} \\
&\leq [2(1-\tau)/\sqrt{\gamma_{\min}}]\lambda + o_p(\lambda).
\end{aligned} \tag{C.15}$$

On the other hand, $\lambda/[\sum_{k=1}^{K} |\underline{\widehat{\theta}}^{(k)}_{j,j'}|]^{1/2} = +\infty$ when $(j,j') \in S^c_{\cup}$. Otherwise, if $(j,j') \in S_{\cup} \backslash S_k$, then

$$\lambda/(\sum_{k=1}^{K} |\underline{\widehat{\theta}}_{j,j'}|)^{1/2} \geq \lambda/\{\sum_{k=1}^{K} (|\underline{\widehat{\theta}}_{j,j'} - \overline{\theta}_{j,j'}| + |\overline{\theta}_{j,j'}|\}^{1/2} \geq \lambda/\sqrt{\gamma_{\max}} \geq (2-2\tau)\lambda/\sqrt{\gamma_{\min}}\ .$$

Thus, for any $(j,j') \in S^c_k$ $(k = 1, \ldots, K)$, we have

$$\begin{aligned} |\nabla_{j,j'} l(\underline{\widehat{\boldsymbol{\theta}}}^{(k)})| &\leq \max_{1 \leq k \leq K} \max_{(j,j') \in S^c_k} |\nabla_{j,j'} l(\underline{\widehat{\boldsymbol{\theta}}}^{(k)})| \\ &< \min_{1 \leq k \leq K} \min_{(j,j') \in S^c_k} \lambda/\sqrt{\sum_{k=1}^{K} |\underline{\widehat{\theta}}^{(k)}_{j,j'}|} \ \leq \ \lambda/\sqrt{\sum_{k=1}^{K} |\underline{\widehat{\theta}}^{(k)}_{j,j'}|}\ . \end{aligned} \tag{C.16}$$

□

BIBLIOGRAPHY

BIBLIOGRAPHY

Airoldi, E. (2007), "Getting Started in Probabilistic Graphical Models," *PLoS Computational Biology*, 3, e252.

Babbie, E. (2010), *The Basics of Social Research*, Wadsworth Publishing, 5th ed.

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008), "Model selection through sparse maximum likelihood estimation," *Journal of Machine Learning Research*, 9, 485–516.

Banfield, J. and Raftery, A. (1993), "Model-based Gaussian and non-Gaussian clustering," *Biometrics*, 49, 803–821.

Barabasi, A.-L. and Albert, R. (1999), "Emergence of scaling in random networks," *Science*, 286, 509–512.

Bazaraa, M., Sherali, H., and Shetty, C. (1993), *Nonlinear programming: theory and algorithms*, John Wiley & Sons, New York.

Bickel, P. and Levina, E. (2008), "Regularized estimation of large covariance matrices," *Annals of Statistics*, 36, 199–227.

Bickel, P. and Levina, L. (2004), "Some theory for Fisher's linear discriminant function, "naive Bayes", and some alternatives when there are many more variables than observations," *Bernoulli*, 10, 989–1010.

Bliss, C. (1935), "The calculation of the dosage-mortality curve," *Annals of Applied Biology*, 22, 134–167.

Bondell, H. and Reich, B. (2009), "Simultaneous factor selction and collapsing levels in ANOVA," *Biometrics*, 65, 169–177.

Borggaard, C. and Thodberg, H. (1992), "Optimal minimal neural interpretation of spectra," *Analytic Chemistry*, 64, 545–551.

Clinton, J., Jackman, S., and Rivers, D. (2004), "The statistical analysis of roll call data," *American Political Science Review*, 98, 355–370.

d'Aspremont, A., Banerjee, O., and El Ghaoui, L. (2008), "First-order methods for sparse covariance selection," *SIAM Journal on matrix Analysis and its Applications*, 30, 56–66.

Daye, Z. and Jeng, X. (2009), “Shrinkage and model selection with correlated variables via weighted fusion,” *Computational Statistics and Data Analysis*, 53, 1284–1298.

de Leeuw, J. (2006), “Principal component analysis of senate voting patterns,” in *Real Data Analysis*, ed. Sawilowski, S., Information Age Publishing, North Carolina, pp. 405–411.

Diaconis, P., Goel, S., and Holmes, S. (2008), “Horseshoes in multidimensional scaling and local kernel methods,” *Annals of Applied Statistics*, 777–807.

Drton, M. and Perlman, M. (2004), “Model selection for Gaussian concentration graphs,” *Biometrika*, 91, 591–602.

Dumais, S. (1991), “Improving the retrieval of information from external source,” *Behavior Research Methods, Instruments and Computers*, 23, 229–236.

Enelow, J. and Hinich, M. (1984), *The spatial theory of voting: an introduction*, Cambridge University Press, Cambridge.

Fan, J., Feng, Y., and Wu, Y. (2009), “Network exploration via the adaptive LASSO and SCAD penalties,” *Annals of Applied Statistics*, 3, 521–541.

Fan, J. and Li, R. (2001), “Variable selection via nonconcave penalized likelihood and its oracle properties,” *Journal of the American Statistical Asscociation*, 96, 1348–1360.

Faraway, J. (2004), *Linear model in R*, CRC Press.

Fraley, C. (1993), “Algorithms for model-based Gaussian hierarchical clustering,” *SIAM Journal on Scientific Computing*, 20, 270–281.

Friedman, J., Hastie, T., Hofling, H., and Tibshirani, R. (2007), “Pathwise coordinate optimization,” *Annals of Applied Statistics*, 1, 302–332.

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

— (2010), “Regularized paths for generalized linear models via coordinate descent,” *Journal of Statistical Software*, 33.

Friedman, J. and Meulman, J. (2004), “Clustering objects on subsets of attributes (with discussion),” *Journal of the Royal Statistical Society, Series B*, 66, 815–849.

Gordon, A. (2008), “A review of hierarchical classification,” *Journal of the Royal Statistical Society, Series A*, 150, 119–137.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2010), “Joint structure estimation for categorical Markov networks,” Tech. rep., Department of Statistics, University of Michigan, Ann Arbor.

— (2011), "Joint estimation of multiple graphical models," *Biometrika*, 98, 1–15.

Hoeffding, W. (1963), "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Asscociation*, 58, 13–30.

Hoefling, H. and Tibshirani, R. (2009), "Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods," *Journal of Machine Learning Research*, 10, 883–906.

Hoff, P. (2006), "Model-based subspace clustering," *Bayesian Analysis*, 1, 321–344.

Hull, J. (1994), "A database for handwritten text recognition research," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16, 550–554.

Hunter, D. and Li, R. (2005), "Variable selection using MM algorithms," *Annals of Statistics*, 33, 1617–1642.

Jeffers, J. (1967), "Two cases studies in the application of principal component," *Applied Statistics*, 16, 225–236.

Johnson, N., Kotz, S., and Balakrishnan, N. (1994), *Continuous Univariate Distributions*, vol. 1, John Wiley & Sons, New Jersey, 2nd ed.

Jollife, I. (1995), "Rotation of principal components: choice of normalization constraints," *Journal of Applied Statistics*, 22, 29–35.

Jollife, I., Trendafilov, N., and Uddin, M. (2003), "A modified principal component technique based on the LASSO," *Journal of Computational and Graphical Statistics*, 12, 531–547.

Jornsten, R. and Keles, S. (2008), "Mixture models with multiple levels, with application to the analysis of multifactor gene expression data," *Biostatistics*, 9, 540–554.

Jung, S.-Y., Park, Y., Choi, K.-S., and Kim, Y. (1996), "Markov random field based English part-of-speech tagging system," in *Proceedings of the 16th Conference on Computational Linguistics*, pp. 236–242.

Kaufman, L. and Rousseeuw, P. (1990), *Finding groups in data: an introduction to cluster analysis*, John Wiley & Sons, New York.

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., and Meltzer, P. S. (2001), "Classification and diagnostic prediction of cancers using gene expression profiling andartificial neural networks," *Nature Medicine*, 7, 673–679.

Kolar, M. and Xing, E. (2008), "Improved estimation of high-dimensional Ising models," in *Eprint arXiv:0811.1239.*

Koren, Y., Bell, R., and Volinsky, C. (2009), "Matrix factorization techniques for recommender systems," *IEEE Computer*, 42, 30–37.

Kotecha, J. and Djuric, P. (1999), "Gibbs sampling approach for generation of truncated multivariate Gaussian random variables," *IEEE Computer Society*, 3, 1757–1760.

Lam, C. and Fan, J. (2009), "Sparsistency and rates of convergence in large covariance matrices estimation," *Annals of Statistics*, 37, 4254–4278.

Land, S. and Friedman, J. (1996), "Variable fusion: a new method of adaptive signal regression," Tech. rep., Department of Statistics, Stanford University, Stanford.

Lee, L.-F. (1979), "On the first and second moments of the truncated multi-normal distribution and a simple estimator," *Economics Letters*, 3, 165–169.

Leppard, P. and Tallis, G. (1989), "Evaluation of the mean and covariance of the truncated multinormal." *Applied Statistics*, 38, 543–553.

Li, C. and Li, H. (2008), "Network-constraint regularization and variable selection for analysis of genomic data," *Bioinformatics*, 24, 1175–1182.

Li, H. and Gui, J. (2006), "Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks," *Biostatistics*, 7, 302–317.

Li, S. (2001), *Markov Random Field Modeling in Image Analysis*, Springer, New York.

MacQueen, J. (1967), "Some methods for classification and analysis of multivariate observations," in *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, vol. 1, pp. 281–297.

Matthews, D. and J.A., S. (1975), *Yeas and nays: normal decision-making in the U.S. House of Representatives*, Wiley Press, New York.

McCabe, G. (1984), "Principal variables," *Technometrics*, 26, 137–144.

McCullagh, P. (1980), "Regression models for ordinal data," *Journal of the Royal Statistical Society, Series B*, 42, 109–142.

McCullagh, P. and Nelder, J. (1989), *Generalized Linear Models*, Chapman and Hall/CRC, London, UK, 2nd ed.

McLachlan, G. and Basford, K. (1988), *Mixture models: inference and applications to clustering*, Marcel Dekker, New York.

McLachlan, G. and Peel, D. (2002), *Finite mixture models*, New York: John Wiley & Sons.

Meinshausen, N. and Buhlmann, P. (2006), "High-dimensional graphs with the lasso," *Annals of Statistics*, 34, 1436–1462.

Meng, X. L. and Rubin, D. (1993), "Maximum likelihood estimation via the ECM algorithm: a general framework," *Biometrika*, 80, 267–278.

Michailidis, G. and de Leeuw, J. (2001), "Multilevel homogeneity analysis with differential weighting," *Computational Statistics and Data Analysis*, 32, 411–442.

Morton, R. (1999), *Methods and models: a guide to the empirical analysis of formal models in political science*, Cambridge University Press, Cambridge.

O'Connell, A. (2005), *Logistic Regression Models for Ordinal Response Variables*, Sage Publications, Inc, 1st ed.

Pan, W. and Shen, X. (2006), "Penalized model-based clustering with application to variable selection," *Journal of Machine Learning Research*, 8, 1145–1164.

Parsons, L., Haque, E., and Liu, H. (2004), "Evaluating subspace clustering algorithms," in *SIAM International Conference on Data Mining*, SIAM, pp. 48–56.

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009), "Partial correlation estimation by joint sparse regression model," *Journal of the American Statistical Asscociation*, 104, 735–746.

Peterson, B. (1990), "Partial proportional odds models for ordinal response variables," *Applied Statistics*, 39, 205–217.

Peterson, C. and Anderson, J. (1987), "A mean field theory learning algorithm for neural networks," *Complex systems*, 1, 995–1019.

Poole, K. and Rosenthal, H. (1997), *Congress: a political-economic history of roll-call voting*, Oxford University Press, Oxford.

Raftery, A. and Dean, N. (2006), "Variable selection for model-based clustering," *Journal of the American Statistical Asscociation*, 101, 168–178.

Ravikumar, P., Wainwright, M., and Lafferty, J. (2010), "High-dimensional Ising model selection using ℓ_1-regularized logistic regression," *Annals of Statistics*, 38, 1287–1319.

Ravikumar, P., Wainwright, M., Raskutti, G., and Yu, B. (2008), "High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence," Tech. rep., Department of Statistics, University of California, Berkeley.

Ravikumar, P., Wainwright, M. J., and Lafferty, J. D. (2009), "High-dimensional Ising model selection using ℓ_1-regularized logistic regression," *Annals of Statistics*, 38, 1287–1319.

Rocha, G., Zhao, P., and Yu, B. (2008), "A path following algorithm for Sparse Pseudo-Likelihood Inverse Covariance Estimation (SPLICE)," Tech. rep., Department of Statistics, University of California, Berkeley.

Rothman, A., Bickel, P., Levina, E., and Zhu, J. (2008), "Sparse permutation invariant covariance estimation," *Electronic Journal of Statistics*, 2, 494–515.

Shojaie, A. and Michailidis, G. (2010), "Penalized likelihood methods for estimation of sparse high dimensional directed acyclic graphs," *Biometrika*, 97, 519–538.

Tadesse, M. G., Sha, N., and Vannucci, M. (2005), "Bayesian variable selection in clustering high-dimensional data," *Journal of the American Statistical Asscociation*, 100, 602–617.

Tallis, G. (1961), "The moment generating function of the truncated multinormal distribution," *Journal of the Royal Statistical Society, Series B*, 23, 223–229.

Thodberg, H. (1996), "A review of Bayesian neural networks with an application to nearinfrared spectroscopy," *IEEE Transactions on Neural Networks*, 7, 56–72.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005), "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society, Series B*, 67, 91–108.

Tutz, G. and Ulbricht, J. (2009), "Penalized regression with correlation-based penalty," *Statistics and Computing*, 19, 239–253.

Vines, S. (2000), "Simple principal components," *Applied Statistics*, 49, 441–451.

von Davier, M. and Carstensen, C. (2010), *Multivariate and Mixture Distribution Rasch Models: Extensions and Applications*, Springer, New York, 1st ed.

Wainwright, M. and Jordan, M. (2008), "Graphical models, exponential families, and variational inference," *Foundations and Trends in Machine Learning*, 1, 1–305.

Walker, S. and Duncan, D. (1967), "Estimation of the probability of an event as a function of several independent variables," *Biometrika*, 54, 167–179.

Wang, P., Chao, D., and Hsu, L. (2009), "Learning networks from high dimensional binary data: an application to genomic instability data," *Biometrics*, To appear.

Wang, S. and Zhu, J. (2007), "Variable selection for model-based high-dimensional clustering and its application to microarray data," *Biometrics*, 64, 440–448.

Xie, B., Pan, W., and Shen, X. (2008), "Penalized model-based clustering with cluster-specific diagonal covariance matrices and grouped variables," *Electronic Journal of Statistics*, 2, 168–212.

Yeoh, E.-J., Ross, M., Shurtleff, S., Williams, W., Patel, D., Mahfouz, R., Behm, F., Raimondi, S., Relling, M., Patel, A., Cheng, C., Campana, D., Wilkins, D., Zhou, X., Li, J., Liu, H., Pui, C.-H., Evans, W., Naeve, C., Wong, L., and Downing, J. (2002), "Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling," *Cancer Cell*, 1, 133–143.

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

Zhou, N. and Zhu, J. (2010), "Group variable selection via a hierarchical lasso and its oracle property," *Statistics and Its Interface*, 3, 557–574.

Zou, H. (2006), "The adaptive LASSO and its oracle properties," *Journal of the American Statistical Asscociation*, 101, 1418–1429.

Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society, Series B*, 67, 301–320.

Zou, H., Hastie, T., and Tibshirani, R. (2006), "Sparse principal component analysis," *Journal of Computational and Graphical Statistics*, 15, 265–286.

— (2007), "On the degrees of freedom of the LASSO," *Annals of Statistics*, 35, 2173–2192.

Zou, H. and Li, R. (2008), "One-step sparse estimates in nonconcave penalized likelihood models," *Annals of Statistics*, 36, 1108–1126.