LEARNING GRAPH STRUCTURE WITH PARAMETRIC AND NON-PARAMETRIC MODELS

by

Shilin Ding

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2012

Date of final oral examination: 05/29/2012

The dissertation is approved by the following members of the Final Oral Committee:

Grace Wahba, Professor, Statistics
Xiaojin Zhu, Associate Professor, Computer Sciences
Zhengjun Zhang, Associate Professor, Statistics
Sunduz Keles, Associate Professor, Statistics
Sijian Wang, Assistant Professor, Statistics

To my wife, Haiyan Xu, and parents, Ximing Ding and Jinghua Ma.

ACKNOWLEDGMENTS

First and foremost, I want to express my utmost gratitude to my advisor, Professor Grace Wahba. Without her, the research would not have been possible. Her deep and broad statistical culture introduced me to the amazing world of statistics. Her dedication in statistical and scientific research motivated me to conduct original research in pursuit of knowledge and excellence. Her inspiration and support has guided me throughout my PhD studies. It is my great honor and privilege to work with her and learn from her. She is not only a mentor, but also a symbol of wisdom, integrity and passion. I will look up to her and follow the example she set for us in my future career.

I want to thank Professor Xiaojin Zhu, Professor Zhengjun Zhang, Professor Sijian Wang, and Professor Sunduz Keles for their service in my thesis committee. I want to dedicate my sincere gratitude to Professor Zhu for his guidance. His kindness, patience and insightful suggestions encouraged me a lot during my work of the thesis. Professor Zhang is a mentor to me in various ways. His broad knowledge and sharp sense of statistics inspired me during my research. Professor Wang shared with me his insightful ideas in statistical analysis which helped my research in many aspects. Professor Sunduz Keles gave me a lot of valuable suggestions in doing research that could make impact both the theory and the practice. I also want to thank Professor Karl Rohe and Professor Stephen Wright for their valuable comments. I have learned many things when discussing with them.

I want to thank Bin Dai, Xiwen Ma, Héctor Corrada Bravo, Zhigeng Geng, Tai Qin, Jing Kong, Kevin Eng, Dongjun Chung, and other graduate students. The former and current Thursday group members helped me in many ways and make my study in Madison enjoyable.

Finally, I would like to thank my wife, Haiyan Xu, who witnessed the happy times and hard ones together with me. It is her love and support that accompanied me during my PhD studies. I also want to thank my parents, Ximing Ding and Jinghua Ma, who have been always supportive since I was born. This work is dedicated to them.

DISCARD THIS PAGE

TABLE OF CONTENTS

Page

DISCARD THIS PAGE

LIST OF TABLES

DISCARD THIS PAGE

LIST OF FIGURES

ABSTRACT

In discrete undirected graphical models, the conditional independence of the node labels Y is specified by the graph structure. We study the case where there is another input random vector X (e.g. observed features) such that the distribution $P(Y \mid X)$ is determined by functions of X that characterize the (higher-order) interactions among the Y's. The main contribution is to learn the graph structure and the functions conditioned on X at the same time.

Parameterizing the graphical models with potential functions might lead to overparameterization. We prove that the discrete undirected graphical models with feature X are equivalent to the multivariate discrete models. The reparameterization of the potential functions in graphical models by conditional log odds ratios of the latter offers advantages in the representation of the conditional independence structure. And the two parameterizations are proved to be equivalent. In addition, the spaces of conditional log odds ratios can be chosen flexibly. They could be linear functional spaces (parametric), or separable Reproducing Kernel Hilbert Spaces determined by kernels (non-parametric).

To obtain a sparse estimation of the graph structure, we impose a Structure Lasso (SLasso) penalty on groups of the conditional log odds ratios to learn the graph structure. These groups with overlaps are designed to enforce hierarchical function selection. An efficient gradient descent algorithm is given to estimate the complete model. The global convergence of the algorithm is guaranteed. And a greedy approach is applied when the graph is large. The BGACV tuning method is derived to select the tuning parameter. It achieves satisfactory numerical results in simulation studies.

The asymptotic analysis shows that the SLasso method is consistent in terms of estimating the graph structure. The consistency properties hold for both the parametric models and the non-parametric models. The experiments show that the SLasso method is able to recover the graph structure with increasing sample size. It also outperforms other methods in the simulation studies.

Chapter 1

Overview

In undirected graphical models (UGMs), a graph is defined as $G = (\Omega, E)$, where $\Omega = \{1, \cdots, K\}$ is the set of nodes and $E \subseteq \Omega \times \Omega$ is the set of edges between the nodes. In fact, Ω is associated to a multivariate response variable $Y = (Y_1, \cdots, Y_K)^T$, and E specifies the conditional independence structure among the components. The UGMs have been widely used in computer vision, natural language processing and other applications. For example, the Conditional Random Fields (CRFs) (Lafferty et al. (2001) [48]) and the extensions, e.g., dynamic CRF (Sutton et al. (2007) [82]), are well known in Natural Language Processing community. The CRFs achieve great success by modeling the effects of features X on the labels (responses) Y of the nodes. There are also numerous applications of the UGMs to computer vision (Szeliski et al. (2007) [83], Schnitzspan et al. (2009) [77]), image processing (Schmidt et al. (2008) [76]), social networks (Banerjee et al. (2008) [8]), and so on.

Graphical Models facilitate the prediction of Y by modeling the relations between its components. Multi-task learning (Caruana (1997) [15]) is related to Graphical Models in this sense. The difference is that the multi-task learning is not focused on higher order interactions on the responses. In the Multi-task learning setting, a set of observations are given for each of the T tasks. In many cases, these tasks will share the same set of features. The general assumption is that there are certain relations between the tasks. Therefore, modeling the T tasks at the same time and considering the relations will be a better choice than treating each task independently. For example, learning speech recognition models for different speakers could be treated as a multi-task learning problem, since the commonality between different speakers could be utilized to improve

the performance. Another example is identifying different but related objects in computer vision (Torralba et al. (2004) [85]).

Evgeniou et al. (2005) [22] considered the embedding of the features into another space and proposed to learn at the same time the T task functions that are in Reproducing Kernel Hilbert Spaces (RKHS). The algorithm works for the linear task functions with linear embedding. Argyriou et al. (2008) [3] proposed a framework to learn sparse representations shared across multiple tasks. The objective function is non-convex because it trys to learn the feature map and the regression parameters at the same time. They proved that the non-convex problem is equivalent to a convex problem and provided the corresponding iterative alternating algorithm. This method is also related to multiple kernel learning (Bach et al. (2004) [7]). Caponnetto et al. (2008) [13] studied the theoretical conditions under which every continuous function in a RKHS can be uniformly approximated in the multi-task settings.

The UGMs are powerful in modeling the joint distribution of Y conditioned on input variables X. The graph structure specifies the conditional independence among the nodes. In many applications, the graph is pre-determined by certain domain knowledge. For example, Duan et al. (2008) [20] proposed a collective model for labeling music signals with fully connected graph, which they called collective conditional random fields. They have 50 labels in 10 semantic categories such as genre (blues, rap, ...), instrument (guitar, piano, ...), production (studio, live), rhythm(strong, weak, middle), and etc. It is possible that some links should not appear, e.g., production and instrument. Estimating the parameters with these interactions included will possibly lead to overfitting. It is important to learn the graph structure and the functions associated with the structure at the same time.

Many prior works have focused on the graphical structure learning without conditioning on X. For instance, Meinshausen and Bühlmann (2006) [63] and Peng et al. (2009) [67] studied the sparse covariance estimation of the Gaussian Markov Random Fields (Speed and Kiiveri (1986) [80]). The covariance matrix fully determines the independence structure in the Gaussian distribution, and thus, specifies the linkage. But it is not the case for non-elliptical distributions, such as the distribution of the multivariate discrete random variables. Ravikumar et al. (2010) [71] and

Xue et al. (2010) [98] discussed consistent structure selection of Ising models based on the l_1-regularized logistic regression, while Höfling and Tibshirani (2009) [33] proposed using pseudo-likelihood with l_1 penalty for estimating sparse Ising models. Ising models are special cases of discrete UGMs with only pairwise interactions, and (usually) without features. We focuse on the discrete UGMs with both higher order interactions and features. It is important to note that the graph structure may change conditioned on different X's, thus our approach may lead to better estimations and interpretation.

In addressing the problem of structure learning with features, Liu et al. (2010) [55] assumed that Y is Gaussian distributed given X, and they partitioned the space of X into bins. We do not assume any special structures of X's in this work but focus on Y which is multivariate discrete when conditioned on X. Schmidt et al. (2008) [76] proposed a framework to jointly learn the pairwise CRFs and the parameters with block-l_1 regularization. Bradley and Guestrin (2010) [11] learned tree CRF that recovers a max spanning tree of a complete graph based on heuristic pairwise link scores. These methods utilize only pairwise information to scale to large graphs. The closest work is Schmidt and Murphy (2010) [75], which examined the higher-order graphical structure learning problem without considering features. They used an active set method to learn higher order interactions in a greedy manner. Their model is over-parameterized, and the hierarchical assumption is sufficient but not necessary for conditional independence in the graph. Buchmann et al. (2012) [12] proposed a structure learning method of binary UGMs without features based on spectral parameterization. This parameterization is equivalent to the multivariate Bernoulli parameterization discussed in Section 2.2. They compared different parameterizations and showed that the spectral parameterization is one of the best performing parameterizations.

To the best of our knowledge, no previous work addressed the issue of graph structure learning of all orders while conditioning on input features. The advantage is the combination of the graph structure learning and the flexible choice of the functional spaces on X. Our contributions include a reparemeterization of the UGMs with bivariate outcomes by the multivariate Bernoulli (MVB) models. It can be easily extended to general discrete UGMs as shown in Section 2.3. The set of conditional log odds ratios in the MVB models are complete to represent the effects of features

on responses and their interactions at all levels. The sparsity in the set of functions are sufficient and necessary for the conditional independence in the graph, i.e., two nodes are conditionally independent if and only if all the interactions that contain these two nodes are constant zero; and the higher order interaction among a subset of nodes means none of the variables is separable from the others in the joint distribution.

To obtain a sparse graph structure, we impose Structure Lasso (SLasso) penalty on groups of the conditional log odds ratios with overlaps. SLasso can be viewed as the group lasso with overlaps. The group lasso that is proposed in Yuan and Lin (2006) [100] leads to the selection of variables in groups. They showed that it is consistent when the groups are exclusive and cover the whole set. Jacob et al. (2009) [35] considered the penalty on groups with arbitrary overlaps. Zhao et al. (2009) [103] set up the general framework for hierarchical variable selection with overlapping groups, which we adopt here for the functions. Our groups are designed to enforce the sparsity on the set of functions and shrink higher order interactions similar to the hierarchical inclusion restriction in Schmidt and Murphy (2010) [75]. We give a proximal linearization algorithm that efficiently learns the complete model, where the normalization factor is calculated by the junction tree algorithm (Koller and Friedman (2009) [44]). The global convergence is guaranteed (Wright (2010) [96]). It can be used in applications where the number of responses is small, such as the Census Bureau data in Section 5.2. It can also be applied to model the relations of multiple clinical responses (hypertension, diabetes, etc.) and how they are affected by the person's genetic and environmental variables (smoking, income, etc). We then propose a greedy search algorithm to scale our method to large graphs as the number of parameters grows exponentially. This algorithm can scale to large graphs (100 nodes or more) by a greedy type search from main effects to higher order interactions.

In addition, we allow the conditional log odds ratios of the joint distribution be functions in any separable Reproducing Kernel Hilbert Spaces. In this way, we extend the linear models to the non-parametric models. For the non-parametric regression in exponential families, Lin and Zhang (2006) [53] proposed the component selection and smoothing operator (COSSO) method for model selection and estimation. They proposed iterative alternating algorithm for learning the

model parameters and the dummy variables that determines the sparsity in the model. Although optimizing over the two sets of parameters is not convex, they showed it is equivalent to a convex optimization problem. They also showed the rate of COSSO estimators converging to the true model in terms of the l_2 norm of the function values on the observations. Other references about the asymptotic results of non-parametric models include but not restricted to Bach (2008) [6], Radchenko and James (2010) [69], Ravikumar et al. (2009) [70], Meier et al. (2009) [62], Huang et al. (2010) [34], and Koltchinskii and Yuan (2010) [45]. Our contribution is to give the sufficient and necessary conditions for the model selection consistency of SLasso with parametric and non-parametric models. Due to the special design of the structure penalty, the SLasso method is consistent in terms of graphs structure estimation. That is, if the true model satisfies the hierarchical structure assumption, the SLasso method is consistent in estimating the set of non-zero conditional log odds ratios. If not, the SLasso method will recover a superset of the non-zero conditional log odds ratios in the true model. The superset will still give the same graph structure, so the result will still preserve the conditional independence structure.

The thesis is organized as follows: Chapter 2 introduces the Graphical Models, multivariate Bernoulli model and its generalizations, multivariate discrete model. We show that Graphical Models are equivalent to the multivariate discrete models. Chapter 3 discussed the SLasso method and the structure penalty. We provided the gradient descent algorithm for learning the model. We derive the GACV and BGACV score to select the tuning parameter. Chapter 4 discusses the asymptotic results for parametric and non-parametric models. The experiments are discussed in Chapter 5. Chapter 6 gives the concluding remarks.

The notations in this paper are summarized in Table 1.1. Without special notice, $\|\cdot\|_n$ denotes the Euclidean l_n norm if $n = 1, 2, \cdots$; $\|\cdot\|_{\mathcal{H}}$ denotes the norm in the space $\mathcal{H}$; $\|\cdot\|_K$ denotes the norm induced by the kernel K; $\|\cdot\|_{\mathcal{J}}$ denotes the conjugate norm with respect to $\mathcal{J}$ if $\mathcal{J}$ is itself a norm (or a penalty) as defined in Definition 4.17.

Table 1.1 Notations

Symbol	Description
$\|\cdot\|$	Euclidean l_2 norm
n	Sample size
p	Number of covariates
K	Number of Response/Output
Y	K dimensional Response/Output
X	Covariates/Feature/Input, $n \times p$ matrix
Ω	Set of $\{1, 2, \ldots, K\}$
$\wp(\Omega)$	Power set of Ω except the empty set
ω, κ, v	Element of $\wp(\Omega)$ used for indexing
$y^\omega(i)$	$y^\omega(i) = \prod_{k\in\omega} y_k(i)$
$\mathcal{Y}(i)$	Augmented responses $(y^1(i), \ldots, y^\Omega(i))$ where $y^\omega(i) = \prod_{k\in\omega} y_k(i)$
c	Model parameters
$\tilde{p}$	Dimension of c^ω. It is $(p+1)$ in linear models
$\tilde{K}$	Number of f^ω's. It is $\|\wp(\Omega)\|$ if there is no restriction on the model
T_v	$T_v = \{\omega \| v \subseteq \omega\}$ is the subgraph rooted at v containing all its descendants
f^{T_v}	$f^{T_v} = (f^\omega), \omega \in T_v$, where f^ω is the conditional log odds ratio
$\mathcal{J}(f)$	Penalty on f
$\mathcal{I}_\lambda$	The objective with tuning parameter λ
p_v	Weight for penalty on structure T_v
s_v, r_v	Subgradient of $\lambda J(f^{T_v})$ of the vth group
$S^\omega(y; x)$	$S^\omega(y; x) = \sum_{\kappa\in T^\omega} y^\kappa f^\kappa$

Chapter 2

Graphical Models and Multivariate Discrete Distribution

In this chapter, we will discuss the distribution of a multivariate discrete random vector which has higher order interactions. We will show that the formulation of the multivariate discrete distribution is equivalent to the discrete Undirected Graphical Models. And the former is more suitable for learning the graph structure.

2.1 Discrete Undirected Graphical Models

In Undirected Graphical Models (UGMs), a graph is defined as $G = (\Omega, E)$, where $\Omega = \{1, \cdots, K\}$ is the set of nodes and $E \subseteq \Omega \times \Omega$ is the set of edges between the nodes. A UGM is also called a Markov Random Field (Kindermann et al. (1980) [40]) because of its Markov properties we will discuss later.

Suppose the multivariate response vector associated with the nodes is $Y = (Y_1, \cdots, Y_K)^T$, and suppose there is a p dimensional predictive variable X which can be viewed as common features shared by the K response variables. We call a UGM with discrete response variables as a discrete UGM.

The Markov property formulates the conditional independence structure of a UGM: given three sets of nodes A, B, C in Ω, A and B are independent given C if all the paths from a node in A to a node in B will go through C. Define a clique to be a fully connect subgraph of G, and define a maximal clique to be a clique which is not properly contained in any other cliques. The Markov property leads to the conclusion that any two nodes not in a clique are conditionally independent

given others. This property gives a reasonable decomposition of the graph G according to the cliques.

One formulation of the joint distribution of discrete $Y = (Y_1, \ldots, Y_K)^T$ conditioned on X is parameterized by a set of potential functions on a set of partitions (Bishop (2006) [10] Chapter 8)

$$P(Y_1 = y_1, \ldots, Y_K = y_K|X) = \frac{1}{Z(X)} \prod_{C \in \mathcal{C}} \Phi_C(y_C; X) \tag{2.1}$$

where $Z(X)$ is a normalization factor that ensures $P(\cdot)$ is a well defined probability measure

$$Z(X) = \sum_{y} \prod_{C \in \mathcal{C}} \Phi_C(y_C; X) \tag{2.2}$$

The distribution in Equation (2.1) is factorized according to $\mathcal{C}$, which is usually the set of cliques in the graph. $\Phi_C(X)$ is a potential function of X on C, indexed by the realization of $Y_C = y_C = (y_i)_{i \in C}$, that is

$$\Phi_C(X) = \sum_{y_C} I(Y_C = y_C) \Phi_C(y_C; X) \tag{2.3}$$

where $I(\cdot)$ is the indicator function. And we only consider $\Phi_C \geq 0$ to make sure the probability will always ≥ 0.

For the purpose of efficient computation, $\mathcal{C}$ is often chosen to be the set of maximal cliques. Different representations by non-maximal cliques can be converted to maximal cliques representation by reformulation of the potential functions (Wainwright and Jordan (2008) [94] Chapter 2). So $\mathcal{C}$ does not have to be the set of cliques implied by the graph structure, as long as it is sufficient to represent the joint distribution. For example, the most general and trivial choice for any given graph is $\mathcal{C} = \{\Omega\}$. In this case, we cannot infer the conditional independence from the formulation in Equation (2.1). There are 2^K potential functions for any given graph, even a sparse one. This number is much more than that in a maximal cliques representations for a sparse graph. This is because choosing $\mathcal{C}$ as $\{\Omega\}$ is over-parameterized. And a lot of those potential functions are trivial in the sense of being constant functions. In this case, the conditional independence between the response variables is implicitly formulated by the form of the potential functions.

Example 2.1. *In Figure 2.1(a), we have a triangle clique* $\{Y_1, Y_2, Y_3\}$ *indicate a third order interaction.* Y_4 *is independent with other nodes. Additionally, there is a pairwise interaction between* Y_3 *and* Y_4*.* Y_4 *is conditionally independent with* Y_1 *or* Y_2 *given* Y_3*.* $\mathcal{C}$ *in Figure 2.1(a) is* $\{\{1,2,3\},\{3,4\}\}$.

In Figure 2.1(b), Y_5, Y_6 *are another set of interacted random variables which are independent of other 4 nodes. In this case,* $\{Y_2, Y_4\}$ *are conditionally independent given* Y_3*, so are* $\{Y_1, Y_4\}$*.* $\mathcal{C} = \{\{1,2,3\},\{3,4\},\{5,6\}\}$.

In Figure 2.1(c), Y_5, Y_6, Y_7, Y_8 *form a 4-node clique that are independent to* Y_1, Y_2, Y_3 *and* Y_4*.* $\mathcal{C} = \{\{1,2,3\},\{3,4\},\{5,6,7,8\}\}$.

In Figure 2.1(d), $Y_5, \cdots, Y_8$ *form a 4-node clique.* Y_9, Y_{10} *are connected to the clique through* Y_7*.* $\mathcal{C} = \{\{1,2,3\},\{3,4\},\{4,5\},\{5,6,7,8\},\{7,9\},\{9,10\}\}$.

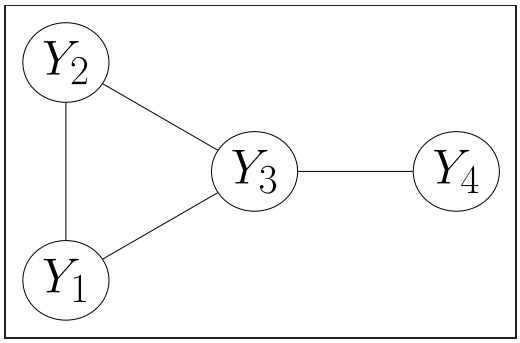
2.2 Multivariate Bernoulli Distribution

The Graphical Model representation in Equation (2.1) is powerful in formulating the joint distribution of the multivariate discrete random variables if the graph structure is known in advance. It greatly reduces the number of parameters. But if the graph is unknown in advance, estimating the potential functions on all possible cliques tends to be over-parameterized (Schmidt and Murphy (2010) [75]). Furthermore, forcing $\log \Phi_C(y_C; X) = 0$ is sufficient for the conditional independence among the nodes but not necessary (see Section 2.2.3). Therefore, we introduce another parameterization to learn the joint distribution when the conditional independence (graph structure) is not known.

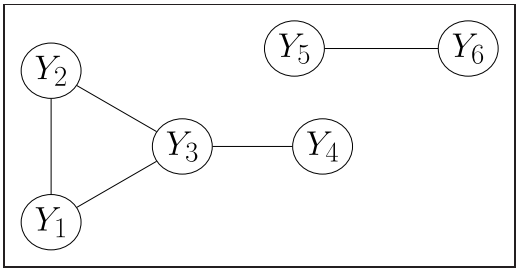
In this section, we consider the multivariate Bernoulli (MVB) random variables, i.e. $Y_k = 0$ or 1. The general results of multivariate discrete random variables are provided in Section 2.3.

2.2.1 Multivariate Bernoulli Distribution formulation

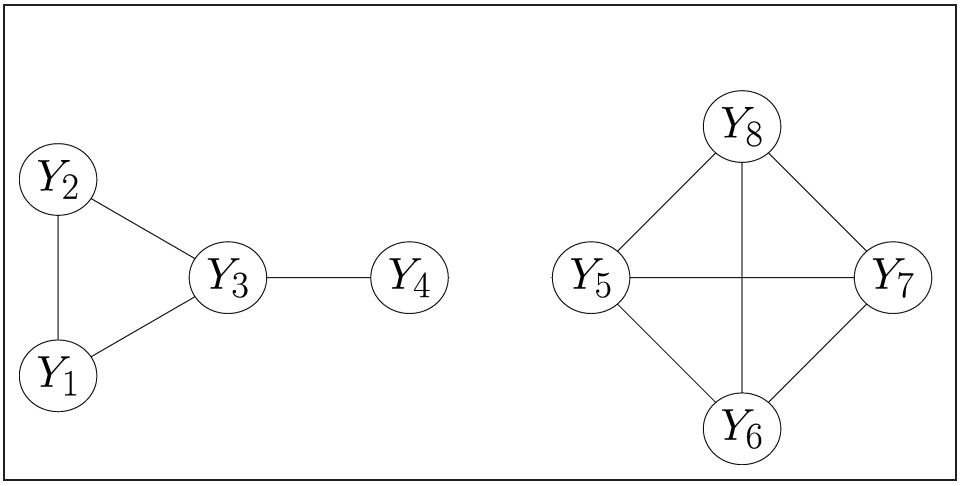
The multivariate Bernoulli (MVB) model of K random variables is equivalent to Equation (2.1) with binary nodes (see Theorem 2.3). It has $2^K - 1$ natural parameters (Whittaker (1990) [95]) if



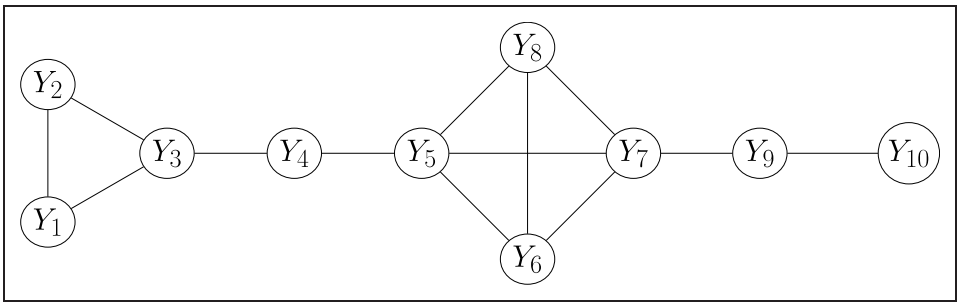
(a) Graph 1



(b) Graph 2



(c) Graph 3



(d) Graph 4

Figure 2.1 Graphical model examples.

the graph is fully connected. The distribution of the MVB model is

$$\begin{aligned} P(Y_1 = y_1, \ldots, Y_K = y_k|X) &= \exp\left\{\mathcal{Y}^T f(X) - b(f(X))\right\} \\ &= \exp\left\{y_1 f^1(X) + \cdots + y_K f^K(X)\right. \\ &\quad + y_1 y_2 f^{1,2}(X) + \cdots + y_{K-1} y_K f^{K-1,K} \\ &\quad + \cdots \\ &\quad \left. + y_1 \ldots y_K f^{1,\ldots,K}(X) - b(f(X))\right\} \end{aligned} \tag{2.4}$$

Here, we use the following notations. Let $\Omega = \{1, \ldots, K\}$ be the set of the nodes in the graph. Denote $\wp(\Omega)$ the power set of Ω leaving out the empty set $\{\emptyset\}$ to index the components from main effects to higher order interactions in the model. There are $|\wp(\Omega)| = 2^K - 1$ components (f^ω's) in (2.4) as free parameters. Let ω denotes a set in $\wp(\Omega)$, define $\mathcal{Y} = (y^1, \cdots, y^\omega, \cdots, y^\Omega)$ be the augmented response with

$$y^\omega = \prod_{i \in \omega} y_i \tag{2.5}$$

Given the predictive variable X, $f = (f^1, \ldots, f^\omega, \ldots, f^\Omega)$ is a vector of functions of X, called conditional log odds ratios (Gao et al. (2001) [25]). It is also referred to as natural parameters in the exponential family (McCullagh and Nelder (1989) [60]). We will call $f^1, \cdots, f^K$ main effects, and $f^{1,2}, \cdots, f^{1,\cdots,K}$ the interactions between the response variables.

From the distribution of a MVB random variable, $f^\omega(x)$ is equivalent to

$$f^\omega(x) = \log OR(Y_i, i \in \omega | Y_j = 0, j \notin \omega; X = x) \tag{2.6}$$

Here, the odds ratios are calculated recursively as

$$OR(Y_i|X = x) = \frac{P(Y_i = 1|X = x)}{1 - P(Y_i = 1|X = x)}, \tag{2.7}$$

$$OR(Y_i, i \in \omega \cup \{k\}|X = x) = \frac{OR(Y_i, i \in \omega|Y_k = 1, X = x)}{OR(Y_i, i \in \omega|Y_k = 0, X = x)}, \text{ with } k \notin \omega \tag{2.8}$$

The following two notations are useful in optimization and parameter tuning

$$S^\omega(y; x) = \sum_{\kappa \subseteq \omega} y^\kappa f^\kappa(x); \quad S^\omega(x) = \sum_{\kappa \subseteq \omega} f^\kappa(x); \tag{2.9}$$

Then the normalization factor is

$$\exp(b(f(x))) = 1 + \sum_{\omega \in \wp(\Omega)} \exp(S^{\omega}(x)) \tag{2.10}$$

In practice, the $\exp(b(f(x)))$ is calculated by the junction tree algorithm (Koller and Friedman (2009) [44] Chapter 10) to avoid enumerating 2^K possible values of Y, which is intractable in large graphs.

We assume f^{ω} is in a separable Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_{\omega}$ with kernel K_{ω} (Wahba (1990) [90]). The details are discussed in Section 2.4. We focus on estimating the set of $f^{\omega}(x)$ with feature x where the sparsity in the set specifies the graph structure.

2.2.2 Relations to Binary Undirected Graphical Models

We present the following lemma and theorem which show the equivalence between the binary UGM in Equation (2.1) and the MVB model in Equation (2.4):

Lemma 2.2. *In a MVB model, define the odd-even partition of the power set of ω as: $\wp_o(\omega) = \{\kappa \subseteq \omega \mid |\kappa| = |\omega| - k, \text{where } k \text{ is odd}\}$, and $\wp_e(\omega) = \{\kappa \subseteq \omega \mid |\kappa| = |\omega| - k, \text{ where } k \text{ is even}\}$. Note $|\wp_o(\omega)| = |\wp_e(\omega)| = 2^{|\omega|-1}$. The following properties hold:*

$$\exp(S^{\omega}(X)) = \frac{P(Y_i = 1, i \in \omega, \text{and } Y_j = 0, j \in \Omega \backslash \omega | X)}{P(Y_i = 0, i \in \Omega | X)} \tag{2.11}$$

$$f^{\omega}(X) = \log \frac{\prod_{\kappa \in \wp_e(\omega)} P(Y_i = 1, i \in \kappa; Y_j = 0, j \in \Omega \backslash \kappa | X)}{\prod_{\kappa \in \wp_o(\omega)} P(Y_i = 1, i \in \kappa; Y_j = 0, j \in \Omega \backslash \kappa | X)} \tag{2.12}$$

$$b(f(X)) = \log \frac{Z(X)}{\prod_{C \in \mathcal{C}} \Phi_C(0; X)} \tag{2.13}$$

Proof. This lemma follows from the formulation of MVB in Equation (2.4) and the definition of odds ratios in Equation (2.6). □

Theorem 2.3. *A UGM of the general form (2.1) with binary nodes is equivalent to a MVB model of (2.4). In addition, the followings are equivalent:*

1. *There is no $|C|$-order interaction in $\{Y_i, i \in C\}$;*

2. *There is no clique $C \subseteq \Omega$ in the graph;*

3. $f^\omega = 0$ *for all* ω *such that* $C \subseteq \omega$.

Proof. The proof is given in Appendix A.1. □

The $|C|$-order interaction in $\{Y_i, i \in C\}$ is defined as: $\{Y_i, i \in C\}$ are not separable in the joint distribution. Theorem 2.3 states that there is a clique C in the graph, if and only if there is $\omega \supseteq C, f^\omega \neq 0$ in the MVB model. The advantage of modeling by MVB is that the sparsity in f^ω's is sufficient and necessary for the conditional independence in the graph, thus fully specifying the graph structure. Specially, Y_i, Y_j are conditionally independent if and only if $f^\omega = 0$ for any ω such that $\{i, j\} \subseteq \omega$. This showed the interaction is non-zero if all the nodes involved are not pairwisely conditionally independent.

2.2.3 Examples in Bivariate Case

For a graph with K nodes, suppose we choose $\mathcal{C} = \{\Omega\}$, the parameters in binary UGM are $\{\Phi_\omega \mid \omega \subseteq \Omega\}$, where $\Phi_\omega = \Phi_\Omega(Y_i = 1, i \in \omega, \text{and } Y_j = 0, j \in \Omega - \omega)$ is the potential function. We usually restrict $\Phi_\emptyset = 1$ to make the model identifiable. So there are $2^K - 1$ free parameters. Similarly, there are also $2^K - 1$ free parameters in MVB model $(f^1, \ldots, f^\Omega)$

When $K = 2$, $\Omega = \{1, 2\}, \mathcal{C} = \{\Omega\}$, write $p_{y_1 y_2} = P(Y_1 = y_1, Y_2 = y_2|X)$ for simlicity, the distribution of Y given X is:

$$\begin{aligned} P(Y_1 = y_2, Y_2 = y_2|X) &= p_{11}^{y_1 y_2} p_{10}^{y_1(1-y_2)} p_{01}^{(1-y_1)y_2} p_{00}^{(1-y_1)(1-y_2)} \\ &= \exp\left\{ y_1 \log\frac{p_{10}}{p_{00}} + y_2 \log\frac{p_{01}}{p_{00}} + y_1 y_2 \log\frac{p_{11}p_{00}}{p_{10}p_{01}} + \log(p_{00}) \right\} \end{aligned} \tag{2.14}$$

The MVB formulation of the distribution is (f^ω denotes $f^\omega(X)$ for simplicity):

$$\begin{aligned} P(Y_1 = y_2, Y_2 = y_2|X) = \exp\big\{ & y_1 f^1 + y_2 f^2 + y_1 y_2 f^{1,2} \\ & - \log\left[\exp(f^1) + \exp(f^2) + \exp(f^1 + f^2 + f^{1,2})\right] \big\} \end{aligned} \tag{2.15}$$

Write $\Phi_\Omega(Y_1 = y_2, Y_2 = y_2; X)$ as $\Phi_{y_1 y_2}$ for simplicity, then the distribution with UGM parameterization is

$$P(Y_1 = y_2, Y_2 = y_2|X) = \frac{1}{Z(X)}\Phi_{y_1 y_2}(X) \tag{2.16}$$

Comparing Equation (2.14), (2.15) and (2.16), and applying the results in Lemma 2.2, we know

$$p_{00} = \frac{1}{Z}\Phi_{00}, \quad p_{01} = \frac{1}{Z}\Phi_{01}, \quad p_{10} = \frac{1}{Z}\Phi_{10}, \quad p_{11} = \frac{1}{Z}\Phi_{11}$$
$$f^1 = \log(p_{10}), \quad f^2 = \log(p_{01}), \quad f^{1,2} = \log\frac{p_{11}p_{00}}{p_{10}p_{01}}$$

And the relations between UGM and MVB are

$$f^1 = \log\frac{\Phi_{10}}{\Phi_{00}},$$
$$f^2 = \log\frac{\Phi_{01}}{\Phi_{00}},$$
$$f^{1,2} = \log\frac{\Phi_{11}\cdot\Phi_{00}}{\Phi_{01}\cdot\Phi_{10}}$$

Note, the independence between Y_1 and Y_2 implies:

$$f^{1,2} = 0 \quad \text{or} \quad \log\frac{\Phi_{11}\cdot\Phi_{00}}{\Phi_{01}\cdot\Phi_{10}} = 0$$

Therefore, $f^{1,2}$ being zero in the bivariate MVB model is sufficient and necessary for the conditional independence in the model. On the other hand, $\log\Phi_C \equiv 0$ is a sufficient condition but not necessary.

2.3 Multivariate Discrete Distribution

The distribution of a general multivariate discrete random vector where $Y_k \in \{0, \cdots, m-1\}$ can be extended from Equation (2.4).

Proposition 2.4. *Let $V = \{1, \ldots, m-1\}, y_\omega = (y_i)_{i\in\omega}$, then*

$$P(Y_1 = y_1, \cdots, Y_K = y_K|X) = \exp\Big\{\sum_{\omega=1}^{\Omega}\sum_{v\in V^{|\omega|}} I(y_\omega = v)f_v^\omega - b(f)\Big\} \tag{2.17}$$

where I is an indicator function and $V^n = V\times\cdots\times V$ is the Cartesian product of n V's. Each f^ω is a $(m-1)^{|\omega|}$ dimensional vector.

Note

$$(m-1)\binom{K}{1}+(m-1)^2\binom{K}{2}+\cdots+(m-1)^K\binom{K}{K}=m^K-1 \tag{2.18}$$

Thus, the number of free parameters in Equation (2.17) is equal to the number in Equation (2.1). And similarly, the multivariate discrete distribution formulation is equivalent to UGM whose response variables taking value in $V=\{0,1,\cdots,m-1\}$.

2.4 Multivariate functions in Reproducing Kernel Hilbert Spaces

The Reproducing Kernel Hilbert Space (Aronszajn (1950) [4]) $\mathcal{H}$ is a Hilbert space of functions on $\mathbb{X}$ for which all the evaluation functionals are bounded and linear. It is associated with a unique Kernel function K which is positive definite in the sense that for any $n=1,2,\cdots$, $x(1),\cdots,x(n)\in\mathbb{X}$ and $a_1,\cdots,a_n\in\mathbb{R}$, $\sum_{i,j=1}^n a_i a_j K(x(i),x(j))\geq 0$. $K(x,\cdot)$ is the Reisz representer of the evaluation functional such that $\langle K(x,\cdot),f\rangle_{\mathcal{H}}=f(x)$, for any $f\in\mathcal{H}$. More details about the related theorems and choices of K can be referred to Wahba (1990) [90].

The extension to the general Reproducing Kernel Hilbert Spaces of multivariate functions is discussed in Wahba (1992) [91]. Micchelli and Pontil (2005) [65] gave another general extension to Hilbert space valued functions. They showed the representer theorem holds and provided practical discussions about the regularization problems, as well as the form of Kernels. Another good reference to vector valued Reproducing Kernel Hilbert Space can be found in Carmeli et al. (2006) [14].

Let f be a M dimensional vector valued function on $\mathbb{X}$, that is $f(x)=(f^1(x),\cdots,f^M(x))^T\in\mathbb{R}^M$. Let u,v index the u-th and v-th components of $f(x)$; $\mathcal{M}=\{1,\cdots,M\}$. Let K be an positive definite function on $\{\mathcal{M}\times\mathbb{X}\}\times\{\mathcal{M}\times\mathbb{X}\}$ in the sense that for any $n=1,2,\cdots$, $x(1),\cdots,x(n)\in\mathbb{X}$, and $a_{ui}\in\mathbb{R}$ for $u=1,\cdots,M, i=1,\cdots,n$

$$\sum_{u,v=1}^{M}\sum_{i,j=1}^{n}a_{ui}a_{vj}K(u,x(i);v,x(j))\geq 0 \tag{2.19}$$

For any fixed v, x, we define the M dimensional vector function as

$$K_{v,x}(\cdot) = \begin{pmatrix} K(1, \cdot; v, x) \\ K(2, \cdot; v, x) \\ \cdots \\ K(M, \cdot; v, x) \end{pmatrix} \tag{2.20}$$

Then, the RKHS $\mathcal{H}_K$ associated with kernel K is defined as the closure of all the countable linear combinations of Equation (2.20) in the form of $f(x) = \sum_{i=1}^{n} \sum_{v=1}^{M} c_{vi} K_{v,x(i)}(x)$. $K_{v,x}$ is the Reisz representer of the evaluation functional such that $\langle f, K_{v,x} \rangle_{\mathcal{H}_K} = f^v(x)$.

The famous Kimerdolf and Wahba representer theorem (Kimeldorf and Wahba (1971) [39]) can be extended to the multivariate case in Wahba (1992) [91]. Let $y^v(i)$ denote the v-th component of the i-th response. Suppose the observation is Gaussian data such that $y^v(i) = f^v(x(i)) + \epsilon^v(i)$, where $\epsilon(i), i = 1, \cdots, n$ are iid multivariate Gaussian random variable $N(0, \sigma^2 I_M)$. The minimizer of the following objective function

$$\min_{f \in \mathcal{H}_K} \sum_{i=1}^{n} \sum_{v=1}^{M} (y^v(i) - f^v(x(i))) + \lambda \|f\|^2_{\mathcal{H}_K} \tag{2.21}$$

has the form of

$$\hat{f}(x) = \sum_{i=1}^{n} \sum_{v=1}^{M} \hat{c}_{vi} K_{v,x_i}(x) \tag{2.22}$$

We assume in the MVB distribution, f is in a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_K$ with kernel K. Since we do not assume any special connection between any pair of conditional log odds ratios, we will suppose $f^v \in \mathcal{H}_v$ which is only associated with a reproducing kernel K_v. It is equivalent to assume $K(u, \cdot; v, \cdot) = 0$ for any $u \neq v$ in the general representation. In this case, $\|f\|^2_{\mathcal{H}_K} = \sum_{v=1}^{M} \|f^v\|^2_{\mathcal{H}_k}$.

Chapter 3

Structure Lasso Model for Graph Learning

3.1 Structure Lasso

3.1.1 Structure Penalty

In many applications, the assumption is that the graph has very few large cliques. It is natual to build up higher order interactions if (at least some) of the lower order interactions exist. One example is the forward search strategy in multivariate adaptive regression splines (MARS, Friedman (1991) [24]). In terms of graph structure learning, we are mainly interested in the set of maximal cliques $\mathcal{C}$ which determines the conditional independence structure of the graph. Any $C \in \mathcal{C}$ contains the cliques of its subset with smaller size. It leads us to include a higher order interaction only when all its lower order interactions are included. Although, with careful choice of the potential function Φ_C on the maximal clique C, we might obtain a MVB distribution where some lower order interactions are zero given f^C is non-zero in the model. For example, $f^{1,2,3} \neq 0$ but $f^{1,2} = 0$ in the true mode. This situation is highly related to the parameterization of the joint distribution, but it does not affect the conditional independence structure of the graph. Later on, the theoretical studies show that this hierarchical restriction will lead to the estimation consistency in graph structure learning. Schmidt and Murphy (2010) [75] applied the same hierarchical inclusion restriction in structure learning with the graphical model parameterization (Equation (2.1)). Radchenko and James (2010) [69] also suggested to include the main effects ahead of the pairwise interaction terms in high dimensional settings for linear regression.

Our model is to fit the graphical model by its multivariate Bernoulli parameterization. We consider the conditional distribution of the nodes (Y) given the predictive variables (X). The

sparsity in the set of conditional log odds ratios is sufficient and necessary for the conditional independence in the graph. Our model are very flexible that $f^{\omega}(x)$ can be in an arbitrary separable RKHS.

Let $Y(i) = (Y_1(i), \ldots, Y_K(i)), X(i) = (X_1(i), \ldots, X_p(i))$ be the ith observation. The augmented representation of the multivariate response is:

$$\mathcal{Y}(i) = (y^1(i), \ldots, y^{\omega}(i), \ldots, y^{\Omega}(i))^T \tag{3.1}$$

The joint distribution in Equation (2.4) contains $\tilde{K} = |\wp(\Omega)| = 2^K - 1$ conditional log odds ratios in the complete model. Suppose the input variable $X \in \mathbb{X} \subset \mathbb{R}^p$, the model has $\tilde{p} = p \cdot \tilde{K}$ free parameters. In cases where there are no predictive variables, e.g., Höfling and Tibshirani (2009) [33] and Ravikumar et al. (2010) [71], the complete MVB distribution has $\tilde{p} = 2^K - 1$ free parameters. We first consider learning the full model when K is small, and later propose a greedy search algorithm to scale to large graphs.

To obtain a sparse estimation of the conditional log odds ratios, we follow the framework of penalized likelihood method (Good and Gaskins (1971) [29])

$$\min_{f \in \mathcal{H}_K} \mathcal{I}_{\lambda}(f) = \mathcal{L}(f) + \lambda \mathcal{J}(f) = \frac{1}{n} \sum_{i=1}^{n} \Big(- \mathcal{Y}(i)^T f(x(i)) + b(f) \Big) + \lambda \mathcal{J}(f) \tag{3.2}$$

Here, the loss function is the negative log likelihood of the observation $Z = (X, Y)$

$$L(Y; f(X)) = -\mathcal{Y}^T f(X) + b\left(f(X)\right) \tag{3.3}$$

and denote $L_Z(f) = L(Y; f(X))$. Then, $\mathcal{L}(f) = \frac{1}{n} \sum_{i=1}^{n} L_{Z_i}(f)$ is the negative log likelihood that evaluate the goodness-of-fit. $\mathcal{J}(\cdot)$ is the penalty that enforce the smoothness and sparsity of the vector valued function f. And λ is the tuning parameter, which controls the trade-off between $\mathcal{L}$ and $\mathcal{J}$.

Our objective is to obtain a sparse estimation of the cliques in the graph through the sparsity of the components in the vector-valued function f. Take the pairwise links for example. No link between Y_s, Y_t in the graphical model means they are conditionally independent given other nodes, or equivalently, $f^{\omega} = 0$ for all $\omega \supseteq \{s, t\}$ (Theorem 2.3). For example, in Figure (2.1(b)), Y_1, Y_4 are

conditionally independent means $f^{1,4}$, $f^{1,2,4,}$, $f^{1,3,4}$, $f^{1,2,3,4}$ are all zero. This objective is similar to the sparse covariance matrix estimation in Gaussian Markov Random Fields for neighborhood selection with lasso (Meinshausen and Bühlmann (2006) [63]). The sparse penalty $\mathcal{J}(\cdot)$ is designed to construct such a graph with sparse cliques. However, our model will deal with higher order covariance structures that do not exist in Gaussian data. In addition, we not only consider the graph structure of responses Y alone, but also the effects of predictive variables X on Y.

To satisfy this intuition, the penalty is designed to shrink higher order interactions in a hierarchical manner. The hierarchical assumption is that if there is no interaction on clique C, then f^{ω} should be zero, for all $\omega \supseteq C$. We consider the Structure Lasso (SLasso) penalty to shrink such f^{ω} toward zero. It is guided by a lattice like Figure (3.1). The lattice has $\tilde{K}$ nodes: $1, \ldots, \omega, \ldots, \Omega$. There is an edge from ω_1 to ω_2 if and only if $\omega_1 \subset \omega_2$ and $|\omega_1| + 1 = |\omega_2|$. Jenatton et al. (2011) [36] discussed how to define the groups to achieve different non-zero patterns in a structured way.

Let $T_v = \{\omega \in \wp(\Omega) | v \subseteq \omega\}$ be the subgraph rooted at v in the lattice, including all the descendants of v. $\mathcal{T} = \{T_1, \ldots, T_\Omega\}$ categorize all the functions into groups with overlaps. Denote f^{T_v} be the vector of functions that concatenates all the components of f in T_v such that $f^{T_v} = (f^{\omega})_{\omega \in T_v}$. Based on the discussion of the extension of RKHS theorems to vector valued functions in Section 2.4, we know $f^{T_v} \in \mathcal{H}_{T_v}$ which is a RKHS associated with kernel K_{T_v}. And K_{T_v} is the original kernel K restricted on the index set T_v. The Structure Lasso (SLasso) penalty on group T_v is:

$$J_v(f) = p_v \|f^{T_v}\|_{\mathcal{H}_{T_v}} = p_v \sqrt{\sum_{\omega \in T_v} \|f^{\omega}\|^2_{\mathcal{H}_{\omega}}}$$

where p_v is the weight for the penalty on T_v. p_v is empirically chosen as $\frac{1}{|T_v|}$, since we do not hope to penalize too much on the components that appear in many groups. And the complete penalty function is

$$\mathcal{J}(f) = \sum_{v \in \wp(\Omega)} J_v(f) = \sum_{v \in \wp(\Omega)} p_v \sqrt{\sum_{\omega \in T_v} \|f^{\omega}\|^2_{\mathcal{H}_{\omega}}} \tag{3.4}$$

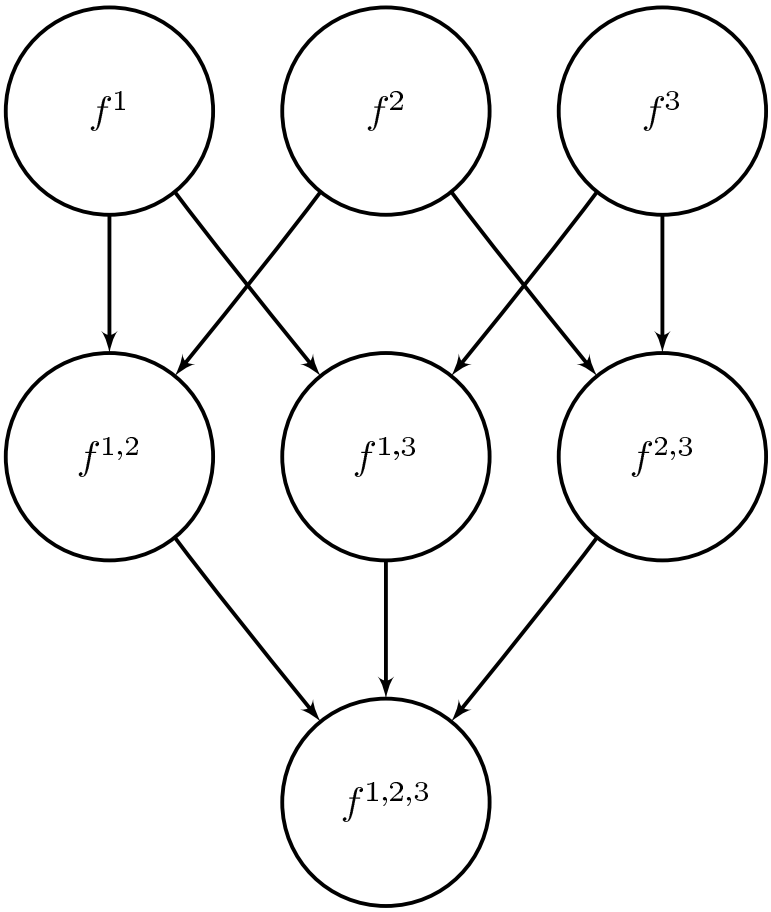


Figure 3.1 Hierarchical lattice for penalty

Then, the objective is:

$$\min_{f} \quad \mathcal{I}_\lambda(f) = \mathcal{L}(f) + \lambda \sum_{v \in \wp(\Omega)} p_v \sqrt{\sum_{\omega \in T_v} \|f^\omega\|^2_{\mathcal{H}_\omega}} \tag{3.5}$$

The following example helps illustrate f, T_v and the objective function.

Example 3.1. *If $K = 3$, the vector-valued function is: $f = (f^1, f^2, f^3, f^{1,2}, f^{1,3}, f^{2,3}, f^{1,2,3})$. The group at node 1 in the lattice (Figure (3.1)) is $f^{T_1} = (f^1, f^{1,2}, f^{1,3}, f^{1,2,3})$ and the objective is*

$$\begin{aligned}
\min_{f \in \mathcal{H}_K} \mathcal{L}(f) + \lambda \Big(& p_1 \sqrt{\|f^1\|^2 + \|f^{1,2}\|^2 + \|f^{1,3}\|^2 + \|f^{1,2,3}\|^2} \\
& + p_2 \sqrt{\|f^2\|^2 + \|f^{1,2}\|^2 + \|f^{2,3}\|^2 + \|f^{1,2,3}\|^2} \\
& + p_3 \sqrt{\|f^3\|^2 + \|f^{1,3}\|^2 + \|f^{2,3}\|^2 + \|f^{1,2,3}\|^2} \\
& + p_{1,2} \sqrt{\|f^{1,2}\|^2 + \|f^{1,2,3}\|^2} + p_{1,3} \sqrt{\|f^{1,3}\|^2 + \|f^{1,2,3}\|^2} \\
& + p_{2,3} \sqrt{\|f^{2,3}\|^2 + \|f^{1,2,3}\|^2} + p_{1,2,3} \sqrt{\|f^{1,2,3}\|^2} \Big)
\end{aligned} \tag{3.6}$$

In non-parametric smoothing regression problems, Lin and Zhang (2006) [53] first proposed the penalty on the sum of RKHS norms instead of the squared norms to select the functional components in Smoothing Spline ANOVA model (Wahba et al. (1995) [92]). The RKHS norm $\|f^{T_v}\|_{\mathcal{H}_{T_v}}$ is nonsmooth at $f^{T_v} = 0$, which leads to the sparse estimation of the components. When $f^{T_v} \neq 0$, the norm penalty will enforce the smoothness of the function. The penalty on the norm of a function in a RKHS can be viewed as a penalty on a group of model parameters, if the RKHS is finite dimensional. Yuan and Lin (2006) [100] proposed Group Lasso for the parametric regression with similar philosophy. The structure penalty has the same effect, except we are dealing with vector-valued functions and we group the components of the functions with overlaps.

The negative log likelihood $L(Y; f(X))$ of the MVB distribution ensures the loss functional $L_Z(\cdot) : \mathcal{H}_K \to \mathbb{R}$ is strictly convex and continuously twice differentiable. Since it does not cause problems for understanding the next theorem, we postpone the discussion of the differentials and other functional operations on $\mathcal{H}_K$ in Section 4.3.1, where we are dealing with the asymptotic results. The following theorem is the extension of the Kimerdolf and Wahba representer theorem to vector valued functions and structure penalty.

Theorem 3.2. *If the loss function $\mathcal{L}(f)$ is convex and continuously twice differentiable, and the penalty function $\mathcal{J}(f)$ is a norm on $\mathcal{H}_K$, then the objective in Equation (3.5) is convex, and there exists a minimizer of Equation (3.5). Let $\hat{f}$ be such minimizer, and assume the kernel K is diagonal in the sense that $K(u,\cdot;v,\cdot) = 0$ for any $u, v \in \wp(\Omega)$ if $u \neq v$, then the ω-th component of $\hat{f}$ is $\hat{f}^v \in span\{K_{v,X(i)}(\cdot), i = 1, \cdots n\}$. That is, $\hat{f}^v(\cdot) = \sum_{i=1}^n c_i^v K_{v,X(i)}(\cdot)$, for some real valued c_i^v.*

Proof. See Appendix A.2. □

The representer theorem ensures that the solution of the non-parametric functional optimization in Equation (3.5) is in a finite dimensional space. This is a crucial property for the feasibility of solving the objective function.

In addition, the following theorem shows that SLasso method achieves the hierarchical inclusion restriction we impose on the graphical model. That is, by minimizing the objective (3.5), $\hat{f}^{\omega_1}$ will enter the model before $\hat{f}^{\omega_2}$ if $\omega_1 \subset \omega_2$. Or equivalently, if $\hat{f}^{\omega_1}$ is zero, there will be no higher order interactions on $\omega_2 \supset \omega_1$. It is an extension of Theorem 1 in Zhao et al. (2009) [103].

The reason can be easily perceived in the following example. $\hat{f}^{1,2} = 0$ only occurs when $\hat{f}^{T_{\{1,2\}}} = 0$. Otherwise, $\|f^{T_{\{1,2\}}}\|_{\mathcal{H}_{T_{\{1,2\}}}}$ is not at the singular point, and thus the probability of $\hat{f}^{1,2} = 0$ is almost zero. However, if $\hat{f}^{1,2,3} = 0$, we will still have the penalty on $f^{1,2}$ which may or may not shrink it to zero.

Theorem 3.3. *Let $\omega_1, \omega_2 \in \wp(\Omega)$ and $\omega_1 \subset \omega_2$. If $\hat{f}$ is the minimizer of (3.5) given the observations, then $0 \in \partial I_\lambda(\hat{f})$ which is the subgradient of I_λ at $\hat{f}$. In addition, $\hat{f}^{\omega_2} = 0$ almost surely if $\hat{f}^{\omega_1} = 0$.*

Proof. The proof is given in Appendix A.3. □

3.1.2 Pattern Selection by SLasso/COSSO Penalty

The structure penalty will satisfy the hierarchical inclusion assumption in the estimated model. In some real applications, it might be preferred to allow higher order interactions exist even some or none of its lower order ones are in the model. But SLasso cannot yield sparsity within the groups. Friedman et al. (2010) [23] considered the sparse group lasso criterion with the combination of l_1

and l_2 norm as the penalty for parametric linear regression model $Y(i) = X(i)^T\beta + \epsilon(i)$

$$J(\beta) = \lambda_1\|\beta\|_1 + \lambda_2\sum_{g=1}^{G}\|\beta^g\|_2 \tag{3.7}$$

where β is model parameter, G is a partition of the components in β without overlaps, and $\beta^g = (\beta_j)_{j\in g}$. They proposed to solve the optimization problem by coordinate descent procedure.

Yuan et al. (2011) [99] studied the problem of overlapping group lasso problem with the penalty of the same formulation as in Equation (3.7), except that they allow the groups overlap with each other. They proposed a fast algorithm based on gradient descent methods which solve the convex dual problem to obtain the proximal operator of the original optimization problem.

To extend the idea of sparse group lasso to the vector valued functional space, we consider the following SLasso /COSSO penalty

$$\begin{aligned}\mathcal{J}(f) &= \mathcal{J}^S(f) + \bar{\lambda}\mathcal{J}^C(f) \\ &= \sum_{v\in\wp(\Omega)} p_v\sqrt{\sum_{\omega\in T_v}\|f^\omega\|^2_{\mathcal{H}_\omega}} + \bar{\lambda}\sum_{v\in\wp(\Omega)}\|f^v\|_{\mathcal{H}_v}\end{aligned} \tag{3.8}$$

where $\mathcal{J}^S(f)$ is the structure penalty defined in Equation (3.4) and $\mathcal{J}^C(f)$ is the COSSO type penalty function presented in Lin and Zhang (2006) [53]. $\bar{\lambda}$ is another tuning parameter that controls the trade-off between the two sparse penalties. It is easy to verify that $\mathcal{J}(f)$ is also a norm on $\mathcal{H}_K$, then the representer theorem in Theorem 3.2 holds.

In linear models, to select the features within each conditional log odds ratio, we propose the following feature selection objective

$$\begin{aligned}\mathcal{J}(f) &= \mathcal{J}^S(f) + \mathcal{J}^L(f) \\ &= \sum_{v\in\wp(\Omega)} p_v\sqrt{\sum_{\omega\in T_v}\|f^\omega\|^2_{\mathcal{H}_\omega}} + \bar{\lambda}\sum_{v\in\wp(\Omega)}\|c^v\|_1\end{aligned} \tag{3.9}$$

where c^v is the vector of paramters in f^v. This objective can be used to select features in multivariate Bernoulli data where not all of the predictive variables are related to the response.

3.2 Estimating the Complete Model for Small Graphs

In this section, we discuss the parameter estimation where the ωth RKHS $\mathcal{H}_\omega$ is composed of a constant functions space and a non-constant RKHS: $\mathcal{H}_\omega = \mathcal{H}_{\omega,0} \oplus \mathcal{H}_{\omega,1}$. We take $\mathcal{H}_{\omega 0} = \{1\}$, which refers to the constant function space. And $\mathcal{H}_{\omega,1}$ could be a linear function space (Example 3.4), a B-Spline function space (Example 3.5), or a general RKHS.

Example 3.4. *Suppose $\mathbb{X} = [-1,1]^p$, $\mathcal{H}_{\omega,1} = \{x_1\} \oplus \cdots \oplus \{x_p\}$ is a RKHS of linear functions. We denote $\{x_j\}$ as a space of linear functions on j-th component of x, and assume the L_2 inner product on $\{x_j\}$: $\langle f, g\rangle_{\{x_j\}} = \int_{[-1,1]} fg$. For example, the functions in $\{x_j\}$ has the form of cx_j. It is easy to obtain the following results: the associated kernel is $K_{\omega,\{x_j\}}(s,t) = \frac{3}{2}st$; the function $f_j^\omega \in \{x_j\}$ must be in the span of the basis functions obtained from $K_{\omega,\{x_j\}}$: $f_1^\omega \in span\{K_{\omega,\{x_j\}}(\cdot, x(i)), i = 1, \cdots, n\}$ for some n and $x(1), \cdots, x(n) \in \mathbb{R}$. So it has the form of*

$$f_j^\omega(x) = \frac{3}{2}\sum_{i=1}^{n} b_i x_i \cdot x = c_j^\omega x \tag{3.10}$$

for some $b_1, \cdots, b_n \in \mathbb{R}$, and $c_1 = \sum_{i=1}^n b_i x(i)$. Thus, $\|f_1^\omega\|^2_{\mathcal{H}_{\omega,1}} = \frac{2}{3}(c_1^\omega)^2$.

The function in $\{1\}$ is a constant: $f_0^\omega = c_0^\omega$; the associated kernel with $\{1\}$ is $K_{\omega,\{1\}}(s,t) = \frac{1}{2}$, $\|f\|^2_{\mathcal{H}_{\omega 0}} = 2(c_0^\omega)^2$.

Theorefore, by specificly choosing K_ω, i.e. $K_\omega = \sum_{j=0}^p \theta_{\omega,p} K_{\omega p} = \frac{1}{2}K_{\omega 0} + \frac{3}{2}\sum_{j=1}^p K_{\omega,\{x_j\}}$, the function $f^\omega \in \mathcal{H}_\omega$ has the form of

$$f^\omega(x) = c_0^\omega + \sum_{j=1}^{p} c_j^\omega x_j \tag{3.11}$$

Its norm is $\|f^\omega\|_{\mathcal{H}^\omega} = \|c^\omega\|$, where $\|\cdot\|$ stands for Euclidean l_2 norm. Here, we denote $c^\omega = (c_0^\omega, \ldots, c_p^\omega)^T \in \mathbb{R}^{p+1}$ as a vector of length $p+1$ and $c = (c^\omega)_{\omega \in \wp(\Omega)} \in \mathbb{R}^{\tilde{K}\cdot\tilde{p}}$ is the concatenated vector of all parameters, where $\tilde{p} = (p+1)$. Let $c^{T_v} = (c^\omega)_{\omega \in T_v}$ be a $(p+1)\cdot|T^v|$ vector, then the objective (3.5) is now

$$\begin{aligned} \min_{c \in \mathbb{R}^{\tilde{K}\cdot\tilde{p}}} \quad \mathcal{I}_\lambda(c) &= \mathcal{L}(c) + \lambda \sum_v p_v \sqrt{\sum_{\omega \in T_v} \|f^\omega\|^2_{\mathcal{H}_\omega}} \\ &= \mathcal{L}(c) + \lambda \sum_v p_v \|c^{T_v}\| \end{aligned} \tag{3.12}$$

and T_v is the subgraph rooted at v defined in the previous section.

Example 3.5. *B-spline basis functions are used in many applictions, e.g. high dimensional additive modeling (Meier et al. (2009) [62] and Huang et al. (2010) [34]). The details about B-spline basis functions can be found in Appendix A.8 and in De Boor (1986) [18].*

Suppose $\mathbb{X} = [-1,1]^p$, $\mathcal{H}_{\omega,1}$ is a B-Spline function space, with reasonable amount of basis functions. And we also assume the L_2 norm on $\mathcal{H}_\omega$. Each $f^\omega(x) \in \mathcal{H}_\omega$ has the form of

$$f^\omega(x) = c_0^\omega + \sum_{j=1}^{D} g_j^\omega(x_j) \tag{3.13}$$

where $g_j^\omega(x_j) = \sum_{k=1}^{D} c_{jk}^\omega B_k(x_j)$ is spanned by the B-spline basis functions $\{B_k(\cdot)\}_{k=1,\cdots,D}$; D is the number of basis functions, and it is determined by the number of knots. See Section A.8 for more details.

Let B_j^ω be a $D \times D$ matrix whose k,l-th element is $(B_j^\omega)_{k,l} = \int_{[-1,1]} g_k(x) g_l(x) dx$; and $B^\omega = diag(1, B_1^\omega, \cdots, B_p^\omega)$ be the blockwise diagonal matrix. Then, the norm on $\mathcal{H}_\omega$ is

$$\|f^\omega\|_{\mathcal{H}_\omega}^2 = \|c_0^\omega\|^2 + \sum_{j=1}^{p} (c_j^\omega)^T B_j^\omega c_j^\omega = (c^\omega)^T B^\omega c^\omega := \|c^\omega\|_{K_\omega} \tag{3.14}$$

Here, we denote $c^\omega = (c_0^\omega, c_{11}^\omega, \cdots, c_{1D}^\omega, \cdots, c_{pD}^\omega)$ as the finite dimensional parameter in $\mathbb{R}^{pD+1}$ for ω-th component of f; $c = (c^\omega)_{\omega \in \wp(\Omega)} \in \mathbb{R}^{\tilde{K} \cdot \tilde{p}}$ is the concatenated vector of all parameters, where $\tilde{p} = pD + 1$. Denote $c^{T_v} = (c^\omega)_{\omega \in T_v}$ be a $(pD+1) \cdot |T^v|$ vector. We will obtain a similar objective function as in Equation (3.12)

$$\begin{aligned} \min_{c \in \mathbb{R}^{\tilde{K} \cdot \tilde{p}}} \quad \mathcal{I}_\lambda(c) &= \mathcal{L}(c) + \lambda \sum_v p_v \sqrt{\sum_{\omega \in T_v} \|f^\omega\|_{\mathcal{H}_\omega}^2} \\ &= \mathcal{L}(c) + \lambda \sum_v p_v \|c^{T_v}\|_{K_{T_v}} \end{aligned} \tag{3.15}$$

where $\|c^{T_v}\|_{K_{T_v}} = \|f^{T_v}\|_{\mathcal{H}_{T_v}} = \sum_{\omega \in T_v} \|c^\omega\|_{K_\omega}$.

3.2.1 Gradient Method by Proximal Linearization

Many applications do not involve a large amount of responses, e.g., the Census Bureau data in our experiment. In these applications, the deep understandings of the higher order interaction

structure are preferrable. So it is desirable to learn the complete model when the graph is small. In this section, we propose a method to optimize Equation (3.5) in the general form of a complete model.

In solving the optimizaition problem, each conditonal log odds ratio f^ω is in a finite dimensional function space spanned by the kernel function evaluated on the n oberservations: $x(1), \cdots, x(n) \in \mathbb{R}^p$. This is enusred by the representer theorem. It is natural to view f^ω as a vector of parameters $c^\omega \in \mathbb{R}^{\tilde{p}}$. $\tilde{p} = p + 1$ for linear case; $\tilde{p} = pD + 1$ for the B-spline case; and $p = n$ for a general RKHS case. Denote the basis functions are $\left\{\phi_j^\omega \mid j = 1, \cdots, \tilde{p}\right\}$, and ϕ^ω be the $\tilde{p}$ dimensional vector of the basis functions, then the form of f^ω can be written as

$$f^\omega = (\phi^\omega)^T c = \sum_{j=1}^{\tilde{p}} c_j^\omega \phi_j^\omega(\cdot) \tag{3.16}$$

Here, we use Σ_ω to denote the $\tilde{p} \times \tilde{p}$ kernel matrix, which is determined by the observations in an infinite dimensional RKHS, i.e., the j, k-th element of Σ_ω is $K_\omega(x(j), x(k))$; or which is determined by the basis functions, i.e., $\left\langle \phi_j^\omega, \phi_k^\omega \right\rangle_{\mathcal{H}_\omega}$, for any $j, k = 1, \cdots, \tilde{p}$. Without special notice, we use c^ω instead of f^ω, and the norm of c^ω is $\|c^\omega\|_{K_\omega} = \|f^\omega\|_{\mathcal{H}_\omega} = (c^\omega)^T \Sigma_\omega c^\omega$, and the inner product on $\mathbb{R}^{\tilde{p}}$ with respect to the kernel matrix K_ω is $\left\langle c^\omega, d^\omega \right\rangle_{K_\omega} = (c^\omega)^T \Sigma_\omega d^\omega$ for any $c^\omega, d^\omega \in \mathbb{R}^{\tilde{p}}$.

The definition of Σ_ω can be extended from ω-th component to $\tilde{K} \cdot \tilde{p} \times \tilde{K} \cdot \tilde{p}$ kernel matrix Σ of the vector valued function space, which is blockwise diagonal. The (u, v)-th block of Σ is 0 if $u \neq v$; the (v, v)-th block is Σ_v. Similarly, $\langle \cdot, \cdot \rangle_K$, and $\| \cdot \|_K$ can be defined.

Lin and Zhang (2006) [53] proposed an equivalent formulation of the COSSO objective to solve the functional optimization problem with RKHS norms which are nonsmooth at the singular point. The equivalent formulation of Equation (3.5) is

$$\min_{f \in \mathcal{H}_K, \gamma_v \geq 0} \quad \mathcal{L}(f) + \lambda_1 \sum_{v \in \wp(\Omega)} \frac{1}{\gamma_v} \Big(\sum_{\omega \in T_v} \|f^\omega\|_{\mathcal{H}_\omega}^2 \Big) + \lambda_2 \sum_{v \in \wp(\Omega)} \gamma_v \tag{3.17}$$

In the equivalent formulation, γ_v is the dummy variable. The procedure is to iteratively fix γ to get an optimal solution of f, and then fix f to obtain a solution of γ. It is efficient for the quadratic loss function on Gaussian data, but the alternating optimization might not scale well in

our case. In stead, we estimate the complete model with all interaction levels by iteratively solving the following proximal linearization problem similar to Wright (2010) [96]. Other references use the proximal method include Mairal et al. (2010) [59]. We will develop the formulation for the general RKHS cases.

$$\min_{c\in\mathbb{R}^{\tilde{K}\cdot\tilde{p}}} \mathcal{L}_k + \langle \nabla\mathcal{L}_k, c - c_k\rangle_{\mathcal{H}_K} + \frac{\alpha_k}{2}\|c - c_k\|^2_{\mathcal{H}_K} + \lambda\mathcal{J}(c) \tag{3.18}$$

In Equation (3.18), let ϕ be the concatenated vector of all basis functions; $\mathcal{L}_k = \mathcal{L}(\phi^T c_k)$; $\nabla\mathcal{L}_k = \nabla\mathcal{L}(\phi^T c_k)$; and α_k is a positive scalar chosen adaptively at kth step. Without causing ambiguity, we denote c_k as the value of c at kth step. Algorithm 1 summarized the framework of solving (3.5).

Following the analysis in Wright (2010) [96], we can show that the proximal linearization algorithm will converge for the negative log-likelihood loss function with the SLasso penalty.

Proposition 3.6. *Let the objective function be defined in Equation (*3.5*), with $\mathcal{L}$ be the negative log-likelihood of the MVB distribution, and $\mathcal{J}$ be the SLasso penalty. Suppose the scaling factor α_k is chosen as described. Then, the sequence $\{c_k\}$ generated by Algorithm 1 will converge to the global minimum of $\mathcal{I}$, and the convergence rate is Q-Quadratic.*

See Nocedal and Wright (1999) [66] page 29 for the definition of the convergence rate.

3.2.2 Dual of the Proximal Linearization Problem

Since the framework of gradient descent method works for solving the SLasso problem, it remains to solve the proximal linearization subproblem in Equation (3.18). Although we can view it as solving a local problem of group lasso with overlaps, it is by no means trivial due to the non-smoothness at the singular point, which is complicated by the overlaps.

In recent years, several papers have addressed the problem of solving group lasso with overlaps. Jacob et al. (2009) [35] duplicated the design matrix columns that appear in group overlaps, then solved the problem as group lasso without overlaps. Kim and Xing (2010) [38] reparameterized the group norm with additional dummy variables as did in Lin and Zhang (2006) [53]. They alternatively optimized the model parameters and the dummy ones at each step. As stated before,

Algorithm 1 Proximal Linearization Algorithm

Input: $c_0, \alpha_0, \alpha_{min}, \alpha_{max}, \zeta > 1, tol > 0$, observations $(y(1), x(1)), \cdots, (y(n), x(n))$

Output: c_k

repeat

Choose $\alpha_k \in [\alpha_{min}, \alpha_{max}]$

Solve Eq (3.18) for $d_k = c - c_k$

while $\delta_k = \mathcal{I}_\lambda(\phi^T c_k) - \mathcal{I}_\lambda(\phi^T(c_k + d_k)) < \|d_k\|^3_{\mathcal{H}_K}$ **do**

// Insufficient decrease

Set $\alpha_k = \max(\alpha_{min}, \zeta\alpha_k)$

Solve Eq (3.18) for d_k

end while

Set $\alpha_{k+1} = \alpha_k/\zeta$

Set $c_{k+1} = c_k + d_k$

until $\delta_k < tol$

this method might not scale well in multivariate Bernoulli data with SLasso penalty. Instead, we will solve (3.18) by its smooth and convex dual problem in Yuan et al. (2011) [99] and Mairal et al. (2010) [59].

To solve the following objective of the proximal linearization problem in Equation (3.18), we solve its dual problem as suggested in Yuan et al. (2011) [99]. Let $\mathcal{A}^{\mathsf{c}} = \{v \in \wp(\Omega) | \|c^{T_v}\| = 0\}$, and $\mathcal{A} = \wp(\Omega) \setminus \mathcal{A}^{\mathsf{c}}$ be the complement. Define s_v for every $v \in \wp(\Omega)$ as

$$s_v \in \mathbb{S}_v = \{s = (s^\omega)_{\omega \in \wp(\Omega)} \mid s \in \mathbb{R}^{\tilde{K} \cdot \tilde{p}}, \|s\|_K \leq \lambda p_v, s^\omega = 0 \text{ if } \omega \notin T_v\} \tag{3.19}$$

Then the subgradient of (3.18) is:

$$K \cdot \nabla L + \alpha_k K(c - c_k) + \sum_{v \in \mathcal{A}^{\mathsf{c}}} s_v + \sum_{u \in \mathcal{A}} r_u \tag{3.20}$$

where s_v is the subgradient of $\lambda p_v \|c^{T_v}\|_{K_{T_v}}$ for $v \in \mathcal{A}^{\mathsf{c}}$ and r_u is the subgradient for $u \in \mathcal{A}$:

$$r_u = \arg\max_{s_u \in \mathbb{S}} \langle s_u, c \rangle_K, \text{ for } u \in \mathcal{A} \tag{3.21}$$

The subgradient s_v is in a unit ball of certain subspace of $\mathbb{R}^{\tilde{K} \cdot \tilde{p}}$ for the linear case. These subspaces are not orthogonal to each other. Thus, s_v's are not separable, and closed form solution of (3.18) cannot be obtained. We solve the proximal subproblem (3.18) by its smoothing and convex dual problem as suggested by Yuan et al. (2011) [99]. Note (3.18) is equivalent to

$$\min_{c \in \mathbb{R}^{\tilde{K} \cdot \tilde{p}}} \max_{S \in \mathbb{S}} \psi(c, S) = \langle \nabla L_k, c - c_k \rangle_K + \frac{\alpha_k}{2} \|c - c_k\|_K^2 + \sum_{v \in \wp(\Omega)} \langle s_v, c \rangle_K \tag{3.22}$$

where S is a $\tilde{K} \cdot \tilde{p} \times |\wp(\Omega)|$ matrix whose columns are s_v. $\mathbb{S} = \{S | S = (s_1, \ldots, s_v, \ldots, s_\Omega), s_v \in \mathbb{S}_v \text{ for } v \in \wp(\Omega)\}$ is the feasible region of S. Since $\psi(\cdot, S)$ is lower semicontinuous and $\psi(c, \cdot)$ is upper semicontinuous, there exists a saddle point and the $\max$ and $\min$ are exchangeable (Barbu and Precupanu (2012) [9]). The solution of minimizing $\psi(c, S)$ is:

$$\tilde{c} = \arg\min_c \psi(c, S) = c_k - \frac{1}{\alpha_k} \nabla L_k - \frac{1}{\alpha_k} \sum_v s_v \tag{3.23}$$

Substitute $\tilde{c}$ back into (3.22), we have the dual problem of (3.18) as:

$$\max_{S \in \mathbb{S}} \eta(S) = -\frac{1}{2} \| \sum_v s_v \|_K^2 + \left\langle \alpha_k c_k - \nabla L_k, \sum_v s_v \right\rangle_K \tag{3.24}$$

Following the proof in Yuan et al. (2011) [99], we can show that $\eta(S)$ is convex and Lipschitz continuous. The differential is $\alpha_k \tilde{c} e^T$ where $e \in \mathbb{R}^{\tilde{K} \cdot \tilde{p}}$ is a vector of ones. Hence, (3.24) can be solved by existing gradient methods. We use the accelerated gradient descent method in Liu et al. (2009) [56].

3.3 Estimating Large Graphs by Greedy Search Algorithm

The above algorithm is efficient on small graphs ($K < 20$). It usually terminates within 20 iterations in our experiments. However, the issue of estimating a complete model is the exponential number of f^ω's and the same amount of groups involved in objective (3.12). It is intractable when the graph becomes large. The hierarchical assumption and the SLasso penalty lend themselves naturally to a greedy search algorithm:

1. Start from the set of main effects as $A_0 = \{f^1, \cdots, f^K\}$. Suppose all higher order interactions are zero.

2. In step i, remove the nodes that are not in A_i from the lattice in Figure 3.1. Obtain a sparse estimation of the functions in A_i by algorithm (1). Denote the resulting sparse set A_i'.

3. Let $A_{i+1} = A_i'$. Keep adding the higher order interactions into A_{i+1} if all its subsets of interactions are included in A_i'. And also add the nodes into the lattice in Figure 3.1.

Iterate step 2 and 3 until convergence. The algorithm is similar to the active set method in Schmidt and Murphy (2010) [75]. It has multiple runs of Algorithm 1 to enforce the hierarchical assumption. It is not guaranteed to converge to the global optimum. Nonetheless, our empirical experiments show its ability to scale to large graphs.

3.4 Parameter Tuning

In the regularization problems, choosing a good tuning parameter λ is a crucial part in fitting the model. Some model selection criteria could be used to choose λ, such as Akaike information criterion (AIC)(Akaike (1973) [1]) and Bayesian information criterion (BIC) (Schwarz (1978)

[78]). These criteria requires the estimation of the degree of freedom, which is not trivial for non-Gaussian data penalized with structure penalty. In Efron (2004) [21], the generalized degree of freedom is defined as

$$\hat{df} = \sum_{i=1}^{n} cov\Big(\hat{f}_i, \mathcal{Y}(i)\Big) \tag{3.25}$$

where $\hat{f}_i = \hat{f}(X_i)$ is the estimated conditional log odds ratios evaluated on X_i. The alternative is the cross validation procedure based on the predictive mean square error. In Gaussian data, if σ is known, the Stein's unbiased risk estimator (SURE) (Stein (1981) [81]) can be used. When σ is unknown, generalized cross validation (GCV) was proposed in Golub et al. (1979) [28] and Craven and Wahba (1979) [16]. The minimizer of the GCV score is a good estimator of the minimizer of the predictive mean square error. Other references about the asymptotic properties of GCV are Li (1985) [50], Li (1986) [51], and Li (1987) [52].

In the non-Gaussian exponential family, Xiang and Wahba (1996) [97] proposed generalized approximate cross validation (GACV) to obtain the λ as a minimizer of the comparative Kullback-Leibler (CKL) distance, which serves as a proxy of the KL distance between the true regression function f^* and the estimated function $\hat{f}$. The goal of GACV is to minimize the KL distance, instead of selecting the "true" model. The consequence is that GACV tends to be conservative in the screening and therefore includes noisy patterns. Shi et al. (2008) [79] proposed B-type GACV (BGACV), which is aimed to balance the KL divergence and the penalty of selecting a noise pattern.

In this section, we will derive the GACV and BGACV tuning criteria for learning graph structure with SLasso penalty in general non-parametric settings. In the end, we will give the approximation of the degrees of freedom of SLasso for AIC and BIC tuning criteria.

Suppose we have n observations, $(Y(i), X(i))$, for $i = 1, \cdots, n$. Denote the grand design matrix as

$$D = \Big(D(1)^T \quad \cdots \quad D(n)^T\Big)^T \tag{3.26}$$

where $D(i)$ is a $\tilde{K} \times \tilde{K} \cdot \tilde{p}$ matrix

$$D(i) = \begin{pmatrix} \phi^1(X(i))^T & 0 & \cdots & 0 \\ 0 & \phi^2(X(i))^T & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \phi^\Omega(X(i))^T \end{pmatrix} \tag{3.27}$$

where $\phi^\omega(X(i))$ is a $\tilde{p}$ dimensional vector of basis functions evaluated on $X(i)$ for $\omega \in \wp(\Omega)$, i.e.

$$\phi^\omega(X(i)) = (\phi_1^\omega(X(i))^T, \cdots, \phi_j^\omega(X(i))^T, \cdots, \phi_{\tilde{p}}^\omega(X(i))^T)^T \tag{3.28}$$

Let $\vec{f}$ be the vector of evaluations of f on the n observations, Then, we have $\vec{f} = Dc$ where c is the $\tilde{p}$ dimensional model parameter, and $\tilde{K} \cdot \tilde{p}$ is determined by the number of the basis functions.

Denote $S_i^\omega = S^\omega(X(i))$, where S^ω is defined in Equation (2.9). Then the normalization factor of the i-th data is denoted as $b_i = b(f(X(i))) = \log(1 + \sum_\omega \exp S_i^\omega)$, and write $\vec{b} = (b_1, \cdots, b_n)^T$. See Section 2.2.1 for more details.

The mean of the augmented response $\mathcal{Y}(i)$ in the MVB model is a $\tilde{K}$ dimensional vector

$$\begin{aligned} \mu(i) &= (\mu^1(i), \cdots, \mu^\omega(i), \cdots, \mu^\Omega(i))^T \\ &= \mathbb{E}[\mathcal{Y}(i) | X(i), f] \end{aligned} \tag{3.29}$$

where

$$\mu^\omega(i) = \mathbb{E}[y^\omega(i) \mid X(i), f] = \frac{\partial b_i}{\partial f^\omega} = \frac{\sum_{\kappa \in T_\omega} \exp S_i^\kappa}{\exp b_i} \tag{3.30}$$

Denote f_λ the minimizer of Equation (3.5) with tuning parameter λ; denote $f_{\lambda,\epsilon}$ the minimizer of Equation (3.5) with tuning parameter λ and small perturbation ϵ on $\mathcal{Y}$; and denote $f_\lambda^{[-i]}$ the minimizer of Equation (3.5) with i-th data point omitted. Let $\vec{f}_\lambda$, $\vec{f}_{\lambda,\epsilon}$, $\vec{f}_\lambda^{[-\cdot]}$ be the corresponding evaluation of f_λ, $f_{\lambda,\epsilon}$, $f_\lambda^{[-i]}$ on the observations respectively, with model parameter c_λ, $c_{\lambda,\epsilon}$, $c_\lambda^{[-i]}$.

The CKL distance between the true model and the estimated model is

$$\mathrm{CKL}(\lambda) = \frac{1}{n} \sum_{i=1}^n \left[-\mu(i)^T f_\lambda(x(i)) + b(f_\lambda(x(i))) \right] \tag{3.31}$$

As a good estimator of $\mathrm{CKL}(\lambda)$, the leaving-out-one cross validation function is

$$\begin{aligned} CV(\lambda) &= \frac{1}{n}\sum_{i=1}^{n}\Big[-\mathcal{Y}(i)^T f_\lambda^{[-i]}(x(i)) + b(f_\lambda(x(i)))\Big] \\ &= \mathrm{OBS}(\lambda) + \frac{1}{n}\sum_{i=1}^{n}\mathcal{Y}(i)^T\Big[f_\lambda(x(i)) - f_\lambda^{[-i]}(x(i))\Big] \\ &= \mathrm{OBS}(\lambda) + \frac{1}{n}\mathcal{Y}^T\Big(\vec{f}_\lambda - \vec{f}_\lambda^{[-\cdot]}\Big) \end{aligned} \tag{3.32}$$

where

$$\mathrm{OBS}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left[-\mathcal{Y}(i)^T f_\lambda(x(i)) + b(f_\lambda(x(i)))\right] = \frac{1}{n}\Big(-\mathcal{Y}^T\vec{f}_\lambda + 1^T\vec{b}\Big) \tag{3.33}$$

The GACV method provides a good approximation of $\left(\vec{f}_\lambda - \vec{f}_\lambda^{[-\cdot]}\right)$ for fast computation. The key idea is to identify the $n\cdot\tilde{K}\times n\cdot\tilde{K}$ influence matrix H (Xiang and Wahba (1996) [97] and Ma (2010) [58] Chapter 3) which implies

$$\vec{f}_{\lambda,\epsilon} - \vec{f}_\lambda = H\epsilon \tag{3.34}$$

where ϵ is the perturbation on $\mathcal{Y}$. We suppose the perturbation is very small such that the non-zero patterns in the estimated model will not change. We will derive the formulation of the influence matrix H for the GACV score.

We first state the Leaving-out-one Lemma which is first discussed in Craven and Wahba (1979) [16] and extended to multivariate case in Ma (2010) [58].

Lemma 3.7. ***Leaving-out-one Lemma***

Replace the i-th observed response $\mathcal{Y}(i)$ by a new response $\tilde{\mathcal{Y}}$. Suppose $h_\lambda[i,\tilde{\mathcal{Y}}]$ be the minimizer of

$$\sum_{k\neq i} L_{Z_i}(f) + \Big(-\tilde{\mathcal{Y}}^T f(X(i)) + b(f(X(i)))\Big) + \lambda\mathcal{J}(f) \tag{3.35}$$

Then $h_\lambda[i,\mu_\lambda^{[-i]}(i)] = f_\lambda^{[-i]}$, where $\mu_\lambda^{[-i]}(i)] = \mathbb{E}\Big[\mathcal{Y}|X, f_\lambda^{[-i]}\Big]$

The $\tilde{K} \times \tilde{K}$ covariance matrix of the i-th augmented response under the estimated distribution is

$$W(i) = var(\mathcal{Y}(i) \mid X(i), f_\lambda) \tag{3.36}$$

where the (α, β)-th element of $W(i)$ is:

$$\begin{aligned} W(i)_{\alpha,\beta} &= \frac{\partial^2 b_i}{\partial f^\alpha (\partial f^\beta)^T} \\ &= \frac{1}{(\exp b_i)^2} \left(\exp b_i \cdot \sum_{\omega \in T_\alpha \cap T_\beta} \exp S_i^\omega - \sum_{\omega \in T_\alpha} \exp S_i^\omega \cdot \sum_{\omega \in T_\beta} \exp S_i^\omega \right) \\ &= \frac{1}{\exp b_i} \left(\sum_{\omega \in T_\alpha \cap T_\beta} \exp S_i^\omega \right) - \mu^\alpha(i) \cdot \mu^\beta(i) \end{aligned} \tag{3.37}$$

Remember, $\mathcal{A} = \{v \in \wp(\Omega) | \|f^{T_v}\| = 0\}^{\mathsf{c}}$ is the cover of the non-zero patterns in c. Let $c^{\mathcal{A}}$ be a sub-vector of c with all the components in $\mathcal{A}$, i.e. $c^{\mathcal{A}} = (c^\omega)_{\omega \in \mathcal{A}}$. For any $v \in \mathcal{A}$, let $I_{T_v}^{\mathcal{A}}$ be a $|\mathcal{A}| \cdot \tilde{p} \times |\mathcal{A}| \cdot \tilde{p}$ diagonal matrix whose ω-th diagonal $\tilde{p} \times \tilde{p}$ block is a identity matrix if $\omega \in T_v$. Then, the v-th group penalty $J_v(f)$ can be written as:

$$J_v(f) = p_v \sqrt{\sum_{\omega \in T_v} \|f^\omega\|_{\mathcal{H}_\omega}^2} = p_v \|I_{T_v}^{\mathcal{A}} c^{\mathcal{A}}\|_{K_{\mathcal{A}}} \tag{3.38}$$

Note $I_{T_v}^{\mathcal{A}}$ is symmetric and $I_{T_v}^{\mathcal{A}} \cdot I_{T_v}^{\mathcal{A}} = I_{T_v}^{\mathcal{A}}$, direct calculation yields the derivative and Hessian of the penalty term:

$$\frac{\partial J}{\partial c^{\mathcal{A}}} = \sum_{v \in \mathcal{A}} p_v \frac{I_{T_v}^{\mathcal{A}} \Sigma_{\mathcal{A}} c^{\mathcal{A}}}{\|I_{T_v}^{\mathcal{A}} c^{\mathcal{A}}\|_{K_{\mathcal{A}}}} \tag{3.39}$$

$$\frac{\partial^2 J}{\partial c^{\mathcal{A}} \partial c^{\mathcal{A}^T}} = \sum_{v \in \mathcal{A}} \ddot{J}_v = \sum_{v \in \mathcal{A}} p_v \frac{\left(I_{T_v}^{\mathcal{A}} \Sigma_{\mathcal{A}} \|I_{T_v}^{\mathcal{A}} c^{\mathcal{A}}\|_{K_{\mathcal{A}}}^2 - (I_{T_v}^{\mathcal{A}} \Sigma_{\mathcal{A}} c) \cdot (I_{T_v}^{\mathcal{A}} \Sigma_{\mathcal{A}} c)^T\right)}{\|I_{T_v}^{\mathcal{A}} c^{\mathcal{A}}\|_{K_{\mathcal{A}}}^3} \tag{3.40}$$

where $\ddot{J}_v$ is the second order derivative of J_v.

Let $\tilde{D}$ be the matrix composed by the columns of D whose index is in $\mathcal{A}$.

The analysis of the first order Taylor expansion of $\frac{\partial I_\lambda}{\partial c^{\mathcal{A}}}(c^{\mathcal{A}}_{\lambda,\epsilon}, \mathcal{Y}+\epsilon)$ leads to the formulation of H. The Taylor approximation is

$$\begin{aligned} 0 &= \frac{\partial I_\lambda}{\partial c^{\mathcal{A}}}(c^{\mathcal{A}}_{\lambda,\epsilon}, \mathcal{Y}+\epsilon) \\ &\approx \frac{\partial \mathcal{I}_\lambda}{\partial c^{\mathcal{A}}}(c^{\mathcal{A}}_{\lambda}, \mathcal{Y}) + \frac{\partial^2 \mathcal{I}_\lambda}{\partial c^{\mathcal{A}} \partial c^{\mathcal{A}^T}}(c^{\mathcal{A}}_{\lambda}, \mathcal{Y})(c^{\mathcal{A}}_{\lambda,\epsilon} - c^{\mathcal{A}}_{\lambda}) + \frac{\partial^2 \mathcal{I}_\lambda}{\partial c^{\mathcal{A}} (\partial \mathcal{Y})^T}(c^{\mathcal{A}}_{\lambda}, \mathcal{Y})\epsilon \\ &= \frac{\partial^2 \mathcal{I}_\lambda}{\partial c^{\mathcal{A}} \partial c^{\mathcal{A}^T}}(c^{\mathcal{A}}_{\lambda}, \mathcal{Y})(c^{\mathcal{A}}_{\lambda,\epsilon} - c^{\mathcal{A}}_{\lambda}) + \frac{\partial^2 \mathcal{I}_\lambda}{\partial c^{\mathcal{A}} (\partial \mathcal{Y})^T}(c^{\mathcal{A}}_{\lambda}, \mathcal{Y})\epsilon \end{aligned} \tag{3.41}$$

Note

$$\begin{aligned} \frac{\partial^2 \mathcal{I}_\lambda}{\partial c^{\mathcal{A}} \partial c^{\mathcal{A}^T}}(c^{\mathcal{A}}_{\lambda}, \mathcal{Y}) &= \frac{\partial^2 \mathcal{L}}{\partial c^{\mathcal{A}} \partial c^{\mathcal{A}^T}} + \lambda \frac{\partial^2 \mathcal{J}}{\partial c^{\mathcal{A}} \partial c^{\mathcal{A}^T}} \\ &= \frac{1}{n} \tilde{D}^T W \tilde{D} + \lambda \sum_{v \in \mathcal{A}} p_v \ddot{J}_v \end{aligned} \tag{3.42}$$

and,

$$\frac{\partial^2 \mathcal{I}_\lambda}{\partial c^{\mathcal{A}} (\partial \mathcal{Y})^T}(c^{\mathcal{A}}_{\lambda}, \mathcal{Y}) = -\frac{1}{n} \tilde{D} \tag{3.43}$$

Therefore

$$\begin{aligned} c^{\mathcal{A}}_{\lambda,\epsilon} - c^{\mathcal{A}}_{\lambda} &= -\left(\frac{\partial^2 \mathcal{I}_\lambda}{\partial c^{\mathcal{A}} \partial c^{\mathcal{A}^T}}(c^{\mathcal{A}}_{\lambda}, \mathcal{Y}) \right)^{-1} \frac{\partial^2 \mathcal{I}_\lambda}{\partial c^{\mathcal{A}} (\partial \mathcal{Y})^T}(c^{\mathcal{A}}_{\lambda}, \mathcal{Y})\epsilon \\ &= \left(\tilde{D}^T W \tilde{D} + \lambda n \sum_{v \in \mathcal{A}} p_v \ddot{J}_v \right)^{-1} \tilde{D}\epsilon \end{aligned} \tag{3.44}$$

Remember, ϵ is a small perturbation on $\mathcal{Y}$; $\vec{f_\lambda} = \tilde{D} c^{\mathcal{A}}_{\lambda}$ is the estimated function value with tuning parameter λ; and $\vec{f}_{\lambda,\epsilon} = \tilde{D} c^{\mathcal{A}}_{\lambda,\epsilon}$ is the estimated function value with the perturbation. Therefore, the influence matrix H is

$$H = \tilde{D} \Big(\tilde{D}^T W \tilde{D} + \lambda n \sum_{v \in \mathcal{A}} p_v \ddot{J}_v \Big)^{-1} \tilde{D}^T \tag{3.45}$$

The (i,j)-th $\tilde{K} \times \tilde{K}$ submatrix of H is

$$H(i,j) = \tilde{D}(i)^T \Big(\tilde{D}^T W \tilde{D} + \lambda n \sum_{v \in \mathcal{A}} p_v \ddot{J}_v \Big)^{-1} \tilde{D}(j) \tag{3.46}$$

Note the following two approximations

$$f_\lambda^{[-i]}(X(i)) - f_\lambda(X(i)) \approx H(i,i)\Big(\mu_\lambda^{[-i]}(i) - \mathcal{Y}(i)\Big) \tag{3.47}$$
$$\mu_\lambda^{[-i]}(i) - \mu_\lambda(i) \approx W(i)\Big(f_\lambda^{[-i]}(X(i)) - f_\lambda(X(i))\Big)$$

The approximation of the CV score in Equation (3.32) is

$$ACV(\lambda) = \mathrm{OBS}(\lambda) + \frac{1}{n}\sum_{i=1}^{n} \mathcal{Y}(i)^T (I - H(i,i)W(i))^{-1} H(i,i)(\mathcal{Y}(i) - \mu_\lambda(i)) \tag{3.48}$$

Let $Q(i) = I - H(i,i)W(i)$ for $i = 1, \ldots, n$, define the generalized average matrix (Gao et al. (2001) [25]), denoted as $\bar{Q}$, of $\{Q(i), i = 1, \ldots, n\}$ as follows

$$\bar{Q} = (\delta - \gamma) I_{q\times q} + \gamma \cdot ee^T = \begin{pmatrix} \delta & \gamma & \cdots & \gamma \\ \gamma & \delta & \cdots & \gamma \\ \vdots & \vdots & \ddots & \vdots \\ \gamma & \gamma & \cdots & \delta \end{pmatrix} \tag{3.49}$$

where e is the unit vector of length $\tilde{K}$ and

$$\delta = \frac{1}{nq\sum_{i=1}^{n} tr(Q(i))}, \quad \gamma = \frac{1}{nq(q-1)}\big[e^T Q(i) e - tr(Q(i))\big] \tag{3.50}$$

Let $\bar{H}$ be the generalized average of $\{H(i,i), i = 1, \cdots, n\}$, the GACV score is

$$GACV(\lambda) = OBS(\lambda) + \frac{1}{n}\sum_{i=1}^{n} \mathcal{Y}(i)^T \bar{Q}^{-1} \bar{H}\big(\mathcal{Y}(i) - \mu_\lambda(i)\big) \tag{3.51}$$

The degrees of freedom of multivariate Bernoulli data is generally difficult to obtain. But we can have a good approximation from GACV (Shi et al. (2008) [79]) as

$$\hat{df}(\lambda) = \sum_{i=1}^{n} \mathcal{Y}(i)^T \bar{Q}^{-1} \bar{H}\big(\mathcal{Y}(i) - \mu(i)\big) \tag{3.52}$$

So the BGACV score can be defined as

$$BGACV(\lambda) = OBS(\lambda) + \frac{1}{n}\frac{\log n}{2}\sum_{i=1}^{n} \mathcal{Y}(i)^T \bar{Q}^{-1} \bar{H}\big(\mathcal{Y}(i) - \mu(i)\big) \tag{3.53}$$

For the model selection criteria AIC and BIC, Ma (2010) [58] (page 53) showed that the degree of freedom can be approximated by

$$\hat{df} = tr(WH) \tag{3.54}$$

Therefore, the AIC and BIC criteria are provided as follows

$$AIC(\lambda) = OBS(\lambda) + \frac{1}{n} tr(WH) \tag{3.55}$$

$$BIC(\lambda) = OBS(\lambda) + \frac{1}{n} \frac{\log n}{2} tr(WH) \tag{3.56}$$

Chapter 4

Asymptotic Results

In this chapter, we consider the model selection consistency of graph structure learning. The problem is described in Equation (3.2). Suppose the set of non-zero conditional log odds ratios is $\mathcal{P}$ in the true model, we will prove that SLasso can identify the cover of $\mathcal{P}$. The cover of $\mathcal{P}$ follows the hierarchical inclusion assumption. Thus, SLasso will eventually recover the same graph structure as in the true model. We derive the necessary and sufficient conditions for the consistency of SLasso in terms of graph structure learning when $n \to \infty$.

Sparse penalties have been widely used in the model selection problems and in high dimensional data. Here, we by no means intend to give a comprehensive review of the asymptotic results of the model selection methods, but only discuss the most relevant literatures.

The asymptotic properties of Lasso (Tibshirani (1996) [84]) have been studied in many references. Knight and Fu (2000) [42] showed that the Lasso type estimator of a linear regression problem is $\sqrt{n}$-consistent for a Gaussian random variable with the mean being the true model parameters, and the variance controlled by the noise and the design matrix. But this estimation consistency does not lead to the sparsistency, which means

$$P(\hat{\mathcal{P}} = \mathcal{P}) \to 1 \tag{4.1}$$

where $\hat{\mathcal{P}}$ is the set of estimated non-zero patterns. In this chapter, we will use non-zero patterns and non-zero conditional log odds ratios interchangeably.

Zhao and Yu (2006) [102] gave the Irrepresentable Condition for sign consistency, which is a stronger version of the sparsistency, of the Lasso type estimators. Roughly speaking, if the covariances between the predictor variables in the true model and those not in the true model are

small, the Lasso can select the true model when $n \to \infty$. We will follow this idea to show the SLasso method is sparsistent under some regularity conditions.

The structure penalty in SLasso can be viewed as an extension of the overlapping Group Lasso on functions in certain Reproducing Kernel Hilbert Spaces. The Group Lasso has been proposed in Yuan and Lin (2006) [100]. The advantage is that the Group Lasso will select the variables in groups, which predetermined by certain domain knowledge. Liu and Zhang (2009) [54] extended the L_2 consistent results from Lasso (Meinshausen and Yu (2009) [64]) to Group Lasso. Bach (2008) [6] derived the necessary and sufficient conditions for the model selection consistency of Group Lasso. The results apply to both the linear regression and the non-parametric regression of Gaussian data where the functions are in separable RKHS's. Radchenko and James (2010) [69] studied the variable selection in the nonlinear Gaussian regression models of up to second order interactions. Their results showed the sparsistency of the model with overlapped group lasso penalty. Jenatton et al. (2011) [36] gave the general guideline for constructing the sparsity-inducing norms for specific requirements based on the overlapping among the group penalties. For the Gaussian data, they derived the necessary and sufficient conditions for the model selection consistency. Since the groups have overlaps, the method is also consistent in terms of the cover of the non-zero patterns. Percival (2012) [68] derived the asymptotic distribution of linear regression with overlapping group lasso penalty. They also presented the finite sample bounds on prediction and estimation.

The asymptotic results about non-Gaussian exponential families are not trivial to obtain because of the complexity of the loss function. Meier et al. (2008) [61] extended the Group Lasso to logistic regression models. Their consistency results showed that the squared distance between the conditional log odds ratio of the fitted model and that of the true model goes to 0 in probability. Van De Geer (2008) [86] proved that under certain regularity conditions, the excess risk of the estimator is bounded above with probability that goes to 1 exponentially fast and the estimator will converge to the true parameters. Rocha et al. (2009) [72] provided the asymptotic distribution for the Lasso type estimators with the logistic regression loss and the hinge loss for SVM.

And they derived the necessary and sufficient conditions for sparsistency based on the asymptotic distribution of the estimator.

The asymptotic properties with nonlinear models have also been studied in many references. As mentioned before, Bach (2008) [6] and Radchenko and James (2010) [69] studied the consistency results for nonlinear regression problems. Additive models are popular in non-parametric regression problems (Hastie and Tibshirani (1990) [31]). Ravikumar et al. (2009) [70] proposed sparse additive models for high dimensional non-parametric regression. The penalty can be viewed as the summation of the functional norms. They showed the estimator is sparsistent with increasing number of orthogonal basis functions. Meier et al. (2009) [62] proposed the sparsity-smoothness penalty for non-parametric additive models. The penalty on each function contains both the l_2 norm of the function values and the quadratic smoothness penalty. They showed the asymptotic optimality of the estimator with increasing number of basis functions. Huang et al. (2010) [34] applied adaptive Group Lasso to select non-zero components in the non-parametric additive models. They showed the estimation consistency in terms of l_2 norm, and sparsistency for adaptive Group Lasso penalty. Koltchinskii and Yuan (2010) [45] discussed the asymptotic properties in the general multiple kernel learning setting. The target is to minimize the empirical risk penalized on the function norms. They established the oracle inequalities for the excess risk of the estimators in Reproducing Kernel Hilbert Spaces. The inequalities that hold with large probability gave the diminishing bound for the excess risk as the number of observations goes to infinity.

In this chapter, we will focus on the consistency of the SLasso method in terms of graph structure learning in parametric and non-parametric settings. We will provide the necessary and sufficient conditions for the consistency.

4.1 Consistency of Graph Structure Learning of Linear Models

In this section, we show that the model with the linear conditional log odds ratios will consistently estimate the graph structure under certain conditions. We assume the random design where $Z_i = (Y_i, X_i), i = 1, \cdots, n$ are random variables. We suppose $Y_i \in \mathbb{Y}$, $X_i \in \mathbb{X}$, and P is the probability measure on $\mathbb{Y} \times \mathbb{X}$. In real applications, we have $\mathbb{Y} = \mathbb{R}^K$ and $\mathbb{X} = \mathbb{R}^p$. Let the conditional

log odds ratios be the same as defined in Equation (3.11)(Section 3.2)

$$f^{\omega}(x) = c_0^{\omega} + \sum_{j=1}^{p} c_j^{\omega} x_j$$

where $c^{\omega} = (c_0^{\omega}, \ldots, c_p^{\omega})^T \in \mathbb{R}^{\tilde{p}}$ is a vector of length $\tilde{p} = p + 1$ and $c = (c^{\omega})_{\omega \in \wp(\Omega)} \in \mathbb{R}^{\tilde{K} \cdot \tilde{p}}$ is the concatenated vector of all parameters of length $\tilde{K} \cdot \tilde{p}$.

To make the notations consistent throughout this chapter, we will use f to denote the model parameters c if it does not lead to ambiguity.

Definition 4.1. *Let $\mathcal{T} = \{T_v | v \in \wp(\Omega)\}$ be a partition (with overlaps) of Ω, as a collection of groups T_v. Let $\mathcal{P} = \{v \in \wp(\Omega) \mid \|f^v\| \neq 0|\}$ be the set of indices of non-zero patterns. The cover of $\mathcal{P}$ with respect to the partition $\mathcal{T}$ is:*

$$\begin{aligned} \mathcal{A} = \mathrm{cover}(\mathcal{P}) = \left(\bigcup_{v: T_v \cap \mathcal{P} = \emptyset} T_v \right)^{\mathsf{c}} &= \left(\bigcup_{v: \|f^{T_v}\| = 0} T_v \right)^{\mathsf{c}} \\ &= \bigcup_{v: \|f^v\| \neq 0} \{\omega \in \wp(\Omega) \mid \omega \subseteq v\} \\ &= \{v \in \wp(\Omega) | \|f^{T_v}\| = 0\}^{\mathsf{c}} \end{aligned} \tag{4.2}$$

Note the last equality holds due to the specialty of the hierarchical structure of $\mathcal{T}$. Write the complement of $\mathcal{A}$ as $\mathcal{A}^{\mathsf{c}} = \Omega \setminus \mathcal{A}$. The following notations are useful in the later derivations

$$\mathcal{T}_{\mathcal{A}} = \{T_v | T_v \cap \mathcal{A} \neq \emptyset\} = \{T_v | \|f^{T_v}\| \neq 0\} \tag{4.3}$$

$$\gamma^{\mathcal{A}} = (\gamma^{\omega})_{\omega \in \mathcal{A}}, \quad \gamma^{\omega} = f^{\omega} \sum_{v \subseteq \omega} \frac{p_v}{\|f^{T_v}\|} \tag{4.4}$$

$$\mathcal{L}_{\mathcal{A}}(f^{\mathcal{A}}) = \mathcal{L}(I_{\mathcal{A}} f); \quad \mathcal{J}_{\mathcal{A}}(f^{\mathcal{A}}) = \sum_{v \in \wp(\Omega)} J_v(I_{\mathcal{A}} f) \tag{4.5}$$

where $I_{\mathcal{A}}$ is a diagonal matrix whose i-th diagonal block is a $\tilde{p} \times \tilde{p}$ identity matrix if $i \in \mathcal{A}$.

Here we will give an example of the cover. In Figure 4.1, $\Omega = A \cup B \cup C$; A, B, C are the groups; the true positive patterns are in P. Then, the cover of P with respect to the groups is $A \cup (B \setminus C)$ (all the red region).

Note that from the above definition, $\forall \omega \in \mathcal{A}, v \subseteq \omega \Rightarrow \omega \in T_v, T_v \in \mathcal{T}^{\mathcal{A}}$. Also note that the graph induced by $\mathcal{P}$ and $\mathcal{A}$ are the same. This means that if the estimated non-zero patterns are consistent to $\mathcal{A}$, the estimation is consistent in terms of graph structure. The conclusion of the following corollary follows Theorem 3.3 and the definition of the cover in Equation (4.2).

Corollary 4.2. *Let $\hat{\mathcal{P}}$ be the non-zero patterns in the SLasso estimation of Equation (*3.5*). Then* $\mathrm{cover}(\hat{\mathcal{P}}) = \hat{\mathcal{P}}$ *almost surely.*

Proof. The corollary follows directly from Theorem 3.3 and the definition of the cover. □

We are interested in developing the theory that shows the estimated non-zero patterns (or equivalently, their cover $\hat{\mathcal{A}}$) converge to the true cover $\mathcal{A}$. Let $L(Y; f(X))$ be the loss function as defined in Equation (3.3), and denote $L_Z(f) = L(Y; f(X))$. Suppose f^* is the true model parameter, such that $(f^*)^v = 0$ if $v \notin \mathcal{P}$. In the exponential family, $f^* = \arg\min_f \mathbb{E}_{f^*}[L_Z(f)]$, and $\nabla \mathbb{E}\left[L_{Z_i}(f^*)\right] = \mathbb{E}\left[\nabla L_{Z_i}(f^*)\right] = 0$, since f^* is optimal.

Assume the loss function has the following properties:

1. $\mathbb{E}_P |L(Y; f(X))| < \infty$ for any $f \in \mathbb{R}^{\tilde{K} \cdot \tilde{p}}$

2. L is convex and twice-continuously differentiable in the second component, and

$$\mathbb{E}_P \left[\nabla L(Y; f^*(X)) \nabla L(Y; f^*(X))^T\right] < \infty \tag{4.6}$$

3. The risk function $R(f) = \mathbb{E}_P[L(Y; f(X))]$ is twice differentiable at f^* and its Hessian matrix

$$H(f) = \nabla^2 \mathbb{E}_P[L(Y; f(X))] \tag{4.7}$$

   is strictly positive definite at f^*.

It is obvious that the loss function takes the form of the negative log-likelihood of the exponential family satisfies the above properties.

To show the model consistency of the SLasso method, we will first derive the asymptotic distribution of the estimated parameters $\hat{f}_n$, and then lead to the necessary and sufficient conditions

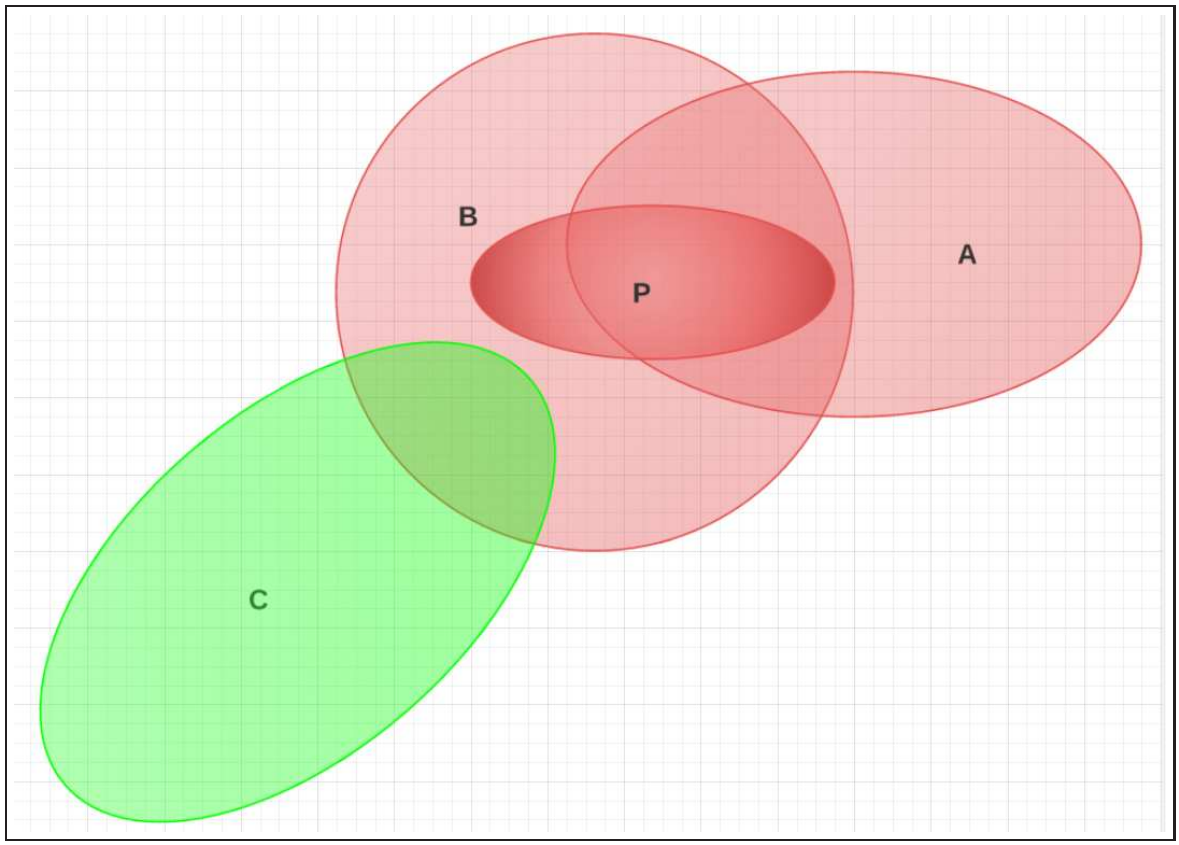


Figure 4.1 Cover of the positive patterns

similarly to the irrepresentible condition in Zhao and Yu (2006) [102] and Rocha et al. (2009) [72]. The following lemma presents the asymptotic distribution which is the key to the later proofs.

Lemma 4.3. *Suppose λ_n is a sequence of positive values which satisfies $\lambda_n \to 0$ and $\lambda_n\sqrt{n} \to \infty$ as $n \to \infty$. Let f^* denote the true model parameters, $\hat{f}_n = \arg\min_f \mathcal{I}_{\lambda_n}(f)$, $H(f^*) = \nabla^2\mathbb{E}\left[L_Z(f^*)\right]$. Then,*

$$\frac{1}{\lambda_n}\left(\hat{f}_n - f^*\right) \xrightarrow{d} \hat{\delta} = \arg\min_{\delta} W(\delta) = \frac{1}{2}\delta^T H(f^*)\delta + \left[(\gamma^{\mathcal{A}})^T\delta^{\mathcal{A}} + \mathcal{J}(\delta^{\mathcal{A}^c})\right] \tag{4.8}$$

where $\gamma^{\mathcal{A}}$ is defined in Equation (4.4) when $f = f^*$.

Proof. See Appendix A.4. □

Before getting to the necessary and sufficient conditions of the sparsistency of SLasso model, we define the conjugate norm with respect to penalty $\mathcal{J}$. Let $f \in \mathbb{R}^{\tilde{K}\cdot\tilde{p}}$, $\mathcal{J}$ as defined before, define the conjugate $\mathcal{J}$-norm as

$$\|f\|_{\mathcal{J}} = \max_{g\in\mathbb{R}^{\tilde{K}\cdot\tilde{p}}, \mathcal{J}(g)\leq 1} \langle f, g\rangle \tag{4.9}$$

We can think f as a linear operator that maps from $\mathbb{R}^{\tilde{K}\cdot\tilde{p}}$ onto $\mathbb{R}$ as $f(g) = \langle f, g\rangle$. Then, the norm $\|f\|_{\mathcal{J}}$ is the conjugate norm defined on the linear operator with respect to the penalty function $\mathcal{J}$.

Theorem 4.4. ***Necessary condition**:*

Let λ_n, f^* and $H(f^*)$ as defined in Lemma 4.3, $\mathcal{A}$ defined in Equation (4.2). Let $H_{\mathcal{A}\mathcal{A}}$ be the sub-matrix of $H(f^*)$ where the rows in $\mathcal{A}$ and the columns in $\mathcal{A}$ of H are selected. Let $H_{\mathcal{A}^c\mathcal{A}}$ be defined similarly. If $\mathcal{A}$ is estimated consistently, that is, $\mathbb{P}(\hat{\mathcal{A}}_n = \mathcal{A}) \to 1$ as $n \to \infty$, where $\hat{\mathcal{A}}_n = \hat{\mathcal{P}}_n = \{\omega | \hat{f}_n^{\omega} \neq 0\}$, then $\left\|H_{\mathcal{A}^c\mathcal{A}}{H_{\mathcal{A}\mathcal{A}}}^{-1}\gamma^{\mathcal{A}}\right\|_{\mathcal{J}_{\mathcal{A}^c}} \leq 1$.

Proof. Let $\hat{\delta}_n = \hat{f}_n - f^*$, then $\hat{\delta}_n \xrightarrow{d} \hat{\delta}$. From Lemma 4.3, the assumption $P(\hat{\mathcal{A}}_n^c = \mathcal{A}^c) \to 1$ leads to $\hat{\delta}^{\mathcal{A}^c} = 0$. The KKT condition of Equation (4.8) is

$$H_{\mathcal{A}\mathcal{A}}\hat{\delta}^{\mathcal{A}} + \gamma^{\mathcal{A}} = 0 \tag{4.10}$$

$$H_{\mathcal{A}^c\mathcal{A}}\hat{\delta}^{\mathcal{A}} + \sum_{v\in\mathcal{A}^c} s_v = 0 \tag{4.11}$$

where s_v is defined slightly different from Equation (3.19) as

$$s_v \in \mathbb{S}_v = \{s = (s^\omega)_{\omega \in \mathcal{A}^c} \mid s \in \mathbb{R}^{|\mathcal{A}^c|}, \|s\| \leq p_v, s^\omega = 0 \text{ if } \omega \notin T_v\} \tag{4.12}$$

From equation 4.10, we get $\hat{\delta}^{\mathcal{A}} = -{H_{\mathcal{A}\mathcal{A}}}^{-1}\gamma^{\mathcal{A}}$, therefore:

$$H_{\mathcal{A}^c\mathcal{A}}{H_{\mathcal{A}\mathcal{A}}}^{-1}\gamma^{\mathcal{A}} = \sum_{v \in \mathcal{A}^c} s_v \tag{4.13}$$

For any $f^{\mathcal{A}^c} \in \mathbb{R}^{|\mathcal{A}^c|}$

$$-\mathcal{J}_{\mathcal{A}^c}(f^{\mathcal{A}^c}) = -\sum_{v \in \mathcal{A}^c} J_v(f^{\mathcal{A}^c}) \leq \langle s_v, f^{\mathcal{A}^c}\rangle \leq \sum_{v \in \mathcal{A}^c} J_v(f^{\mathcal{A}^c}) = \mathcal{J}_{\mathcal{A}^c}(f^{\mathcal{A}^c}) \tag{4.14}$$

The inequality holds when for each $v \in \mathcal{A}^c$

$$s_v = p_v \frac{I_{T_v} f^{\mathcal{A}^c}}{\|I_{T_v} f^{\mathcal{A}^c}\|} \tag{4.15}$$

where I_{T_v} is a diagonal matrix whose i-th diagonal element is 1 if $i \in T_v$. So,

$$|\langle H_{\mathcal{A}^c\mathcal{A}}{H_{\mathcal{A}\mathcal{A}}}^{-1}\gamma^{\mathcal{A}}, f^{\mathcal{A}^c}\rangle| \leq \mathcal{J}_{\mathcal{A}^c}(f^{\mathcal{A}^c}) \tag{4.16}$$

This leads to the conclusion that

$$\left\|H_{\mathcal{A}^c\mathcal{A}}{H_{\mathcal{A}\mathcal{A}}}^{-1}\gamma^{\mathcal{A}}\right\|_{\mathcal{J}_{\mathcal{A}^c}} \leq 1 \tag{4.17}$$

□

Before moving to the theorem of sufficient condition, we present the lemma about the SLasso estimation on the restricted problem on $\mathcal{A}$.

Lemma 4.5. *Let $\hat{f}_n^{\mathcal{A}}$ be the solution of the following problem restricted on $\mathcal{A}$:*

$$\hat{f}_n^{\mathcal{A}} = \underset{f^{\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|}}{\arg\min}\, \mathcal{L}_{\mathcal{A}}(f^{\mathcal{A}}) + \lambda_n \mathcal{J}_{\mathcal{A}}(f^{\mathcal{A}}) \tag{4.18}$$

where $\mathcal{L}_{\mathcal{A}}$ and $\mathcal{J}_{\mathcal{A}}$ are defined in Equation (4.5). Let $\hat{\mathcal{P}}_{n\mathcal{A}} = \{v | \|\hat{f}_n^v\| \neq 0\}$ and $\hat{\mathcal{A}}_{n\mathcal{A}} = \text{cover}(\hat{\mathcal{P}}_{n\mathcal{A}})$, then

$$\hat{f}_n^{\mathcal{A}} \xrightarrow{p} f^{*\mathcal{A}} \quad \textit{and} \quad \mathbb{P}\left(\hat{\mathcal{A}}_{n\mathcal{A}} = \mathcal{A}\right) \to 1 \tag{4.19}$$

Proof. See Appendix A.5. □

Theorem 4.6. ***Sufficient condition:***

Let λ_n, f^* and $H(f^*)$ as defined in Lemma 4.3. If $\left\|H_{\mathcal{A}^c\mathcal{A}}H_{\mathcal{A}\mathcal{A}}{}^{-1}\gamma^{\mathcal{A}}\right\|_{\mathcal{J}_{\mathcal{A}^c}} < 1$, then $\mathcal{A}$ is consistently estimated in the sense that $\mathbb{P}(\hat{\mathcal{A}}_n = \mathcal{A}) \to 1$ as $n \to \infty$, where $\hat{\mathcal{A}}_n = \hat{\mathcal{P}}_n = \{v | \|\hat{f}_n^v\| \neq 0\}$.

Proof. We prove the result based on the primal dual witness technique which were used in Ravikumar et al. (2009) [70] and Wainwright (2009) [93].

Let $\hat{f}_n^{\mathcal{A}}$ be the solution of the restricted problem as defined in Equation (4.18), and pad $\hat{f}_n^{\mathcal{A}}$ with zeros on $\mathcal{A}^c$ to obtain $\hat{f}_n$. From Lemma 4.5, $\hat{f}_n^{\mathcal{A}} \xrightarrow{p} f^{*\mathcal{A}}$ and $\mathbb{P}\left(\hat{\mathcal{A}}_{n\mathcal{A}} = \mathcal{A}\right) \to 1$. Thus, to prove the conclusion, we need to show that $\hat{f}_n$ satisfies the optimality condition of objective in Equation (3.5).

For large n, $\hat{\gamma}^{\mathcal{A}}$ is well defined as

$$\hat{\gamma}^{\mathcal{A}} = (\hat{\gamma}^{\omega})_{\omega \in \mathcal{A}} \quad \text{where } \hat{\gamma}^{\omega} = \hat{f}_n^{\omega} \sum_{v \subseteq \omega} \frac{p_v}{\|\hat{f}_n^{T_v}\|} \tag{4.20}$$

and $\hat{\gamma}^{\mathcal{A}} \xrightarrow{p} \gamma^{\mathcal{A}}$.

The optimality condition on $\mathcal{A}$ is already satisfied due to the definition of $\hat{f}_n$

$$\left(\nabla \mathcal{L}(\hat{f}_n)\right)_{\mathcal{A}} + \lambda_n \hat{\gamma}^{\mathcal{A}} = 0 \tag{4.21}$$

It remains to show that there exist s_v as defined in Equation (4.12) such that

$$\left(\nabla \mathcal{L}(\hat{f}_n)\right)_{\mathcal{A}^c} + \lambda_n \sum_{v \in \mathcal{A}^c} s_v = 0 \tag{4.22}$$

that is, $\|\left(\nabla \mathcal{L}(\hat{f}_n)\right)_{\mathcal{A}^c}\|_{\mathcal{J}_{\mathcal{A}^c}} < 1$.

Let $\hat{\delta}_n = \hat{f}_n - f^* \xrightarrow{p} = 0$. Note

$$\begin{aligned}
\nabla \mathcal{L}(\hat{f}_n) &= \frac{1}{n}\sum_{i=1}^{n} \nabla L_{Z_i}(\hat{f}_n) \\
&= \left[\frac{1}{n}\sum_{i=1}^{n} \nabla L_{Z_i}(f^*)\right] + \left[\frac{1}{n}\sum_{i=1}^{n} \nabla^2 L_{Z_i}(f^*)\hat{\delta}_n\right] + o_p(\|\hat{\delta}_n\|) \\
&= D_n + H_n\hat{\delta}_n + o_p(\|\hat{\delta}_n\|)
\end{aligned} \tag{4.23}$$

As we have shown in Lemma 4.3, $D_n \xrightarrow{p} 0$ and $H_n \xrightarrow{p} H(f^*)$.

Since $\hat{f}_n^{\mathcal{A}^c} = f^{*\mathcal{A}^c} = 0$, we have from the above equation that

$$\left(\nabla \mathcal{L}(\hat{f}_n)\right)_{\mathcal{A}} = H_{\mathcal{A}\mathcal{A}}\hat{\delta}_n^{\mathcal{A}} + o_p(1) \tag{4.24}$$

$$\left(\nabla \mathcal{L}(\hat{f}_n)\right)_{\mathcal{A}^c} = H_{\mathcal{A}^c\mathcal{A}}\hat{\delta}_n^{\mathcal{A}} + o_p(1) \tag{4.25}$$

From Equation (4.24) we have

$$\hat{\delta}_n = -\lambda_n {H_{\mathcal{A}\mathcal{A}}}^{-1}\hat{\gamma}^{\mathcal{A}} + o_p(1) \tag{4.26}$$

Then, the following equality holds because $\hat{\gamma}^{\mathcal{A}} \xrightarrow{p} \gamma^{\mathcal{A}}$

$$\left(\nabla \mathcal{L}(\hat{f}_n)\right)_{\mathcal{A}^c} = -\lambda_n H_{\mathcal{A}^c\mathcal{A}} {H_{\mathcal{A}\mathcal{A}}}^{-1}\gamma^{\mathcal{A}} + o_p(1) \tag{4.27}$$

Therefore, for any $f^{\mathcal{A}^c} \in \mathbb{R}^{|\mathcal{A}^c|}$, the following inequality holds by the sufficient condition

$$\begin{aligned}\left|\left\langle f^{\mathcal{A}^c}, \left(\nabla \mathcal{L}(\hat{f}_n)\right)_{\mathcal{A}^c}\right\rangle\right| &= \lambda_n |\langle f^{\mathcal{A}^c}, H_{\mathcal{A}^c\mathcal{A}} {H_{\mathcal{A}\mathcal{A}}}^{-1}\gamma^{\mathcal{A}}\rangle| + o_p(1) \\ &< \lambda_n \mathcal{J}_{\mathcal{A}^c}(f^{\mathcal{A}^c}) \quad \text{for large } n\end{aligned} \tag{4.28}$$

This completes the proof. □

4.2 Sparsistency of SLasso on Pattern Selection

Friedman et al. (2010) [23] proposed the sparse group lasso criterion with the combination of the l_1 and l_2 norm as the penalty for the parametric linear regression model. We extended the idea to multivariate Bernoulli data as presented in Section 3.1.2. The objective is

$$\mathcal{I}(f) = \mathcal{L}(f) + \lambda_n \left(\mathcal{J}(f) + \bar{\lambda} \sum_{\omega \in \wp(\Omega)} \|f^\omega\|_{\mathcal{H}_\omega} \right) \tag{4.29}$$

Zou and Hastie (2005) [104] proposed the Elastic Net that combines two different types of penalties. Yuan and Lin (2007) [101] showed the regularity conditions for Elastic Net to consistently estimate the non-zero patterns in linear models. Jia and Yu (2010) [37] studied the model selection property of Elastic Net in general settings where the number of non-zero parameters and that of the sample size all go to infinity. Here, we study the sparsistency property of SLasso on pattern selection.

Theorem 4.7. ***Necessary condition**:*

Let λ_n, f^* and $H(f^*)$ defined similarly, $\mathcal{P}$ and $\mathcal{A}$ defined in Equation (4.2) when $f = f^*$. Let $H_{\mathcal{P}\mathcal{P}}$ be the sub-matrix of $H(f^*)$ where the rows $\mathcal{P}$ and columns $\mathcal{P}$ of H are selected. Let $H_{\mathcal{A}\setminus\mathcal{P},\mathcal{P}}, H_{\mathcal{A}^c\mathcal{P}}$ be defined similarly. If $\mathcal{P}$ is estimated consistently, that is, $\mathbb{P}(\hat{\mathcal{P}}_n = \mathcal{P}) \to 1$ as $n \to \infty$, where $\hat{\mathcal{P}}_n = \{\omega | \hat{f}_n^\omega \neq 0\}$, then

$$\left\| H_{\mathcal{A}\setminus\mathcal{P},\mathcal{P}} {H_{\mathcal{P}\mathcal{P}}}^{-1} \left(\gamma^{\mathcal{P}} + \bar{\lambda}\,\mathrm{sign}(f^{*\mathcal{P}})\right) \right\|_\infty \leq \bar{\lambda}$$
$$\left\| H_{\mathcal{A}^c\mathcal{P}} {H_{\mathcal{P}\mathcal{P}}}^{-1} \left(\gamma^{\mathcal{P}} + \bar{\lambda}\,\mathrm{sign}(f^{*\mathcal{P}})\right) \right\|_{\mathcal{J}_{\mathcal{A}^c} + \bar{\lambda} l_1} \leq 1$$

Theorem 4.8. ***Sufficient condition**:*

Let λ_n, f^* and $H(f^*)$ defined similarly. If

$$\left\| H_{\mathcal{A}\setminus\mathcal{P},\mathcal{P}} {H_{\mathcal{P}\mathcal{P}}}^{-1} \left(\gamma^{\mathcal{P}} + \bar{\lambda}\,\mathrm{sign}(f^{*\mathcal{P}})\right) \right\|_\infty < \bar{\lambda}$$
$$\left\| H_{\mathcal{A}^c\mathcal{P}} {H_{\mathcal{P}\mathcal{P}}}^{-1} \left(\gamma^{\mathcal{P}} + \bar{\lambda}\,\mathrm{sign}(f^{*\mathcal{P}})\right) \right\|_{\mathcal{J}_{\mathcal{A}^c} + \bar{\lambda} l_1} < 1$$

then $\mathcal{P}$ is consistently estimated in the sense that $\mathbb{P}(\hat{\mathcal{P}}_n = \mathcal{P}) \to 1$ as $n \to 1$, where $\hat{\mathcal{P}}_n = \{v | \|\hat{f}_n^v\| \neq 0\}$.

Proof. The proof of the above two theorems are similar to those presented in the previous section. We only need to note the following KKT conditions of the objective.

$$H_{\mathcal{P}\mathcal{P}}\hat{\delta}^{\mathcal{P}} + \gamma^{\mathcal{P}} + \bar{\lambda}\,\mathrm{sign}(\hat{\delta}^{\mathcal{P}}) = 0 \tag{4.30}$$
$$H_{\mathcal{A}\setminus\mathcal{P},\mathcal{P}}\hat{\delta}^{\mathcal{P}} + \gamma^{\mathcal{A}\setminus\mathcal{P}} + \bar{\lambda}\,\mathrm{sign}\, t_{\mathcal{A}\setminus\mathcal{P}} = 0 \tag{4.31}$$
$$H_{\mathcal{A}^c\mathcal{P}}\hat{\delta}^{\mathcal{P}} + \sum_{v\in\mathcal{A}^c} s_v + \bar{\lambda} t_{\mathcal{A}^c} = 0 \tag{4.32}$$

where $t_{\mathcal{A}}$ is the subgradient of $\|f^{\mathcal{A}}\|_1$.

Note $\|t_{\mathcal{A}}\|_1 \leq 1$; $\gamma^{\mathcal{A}\setminus\mathcal{P}} = 0$. Applying the techinques to the above KKT conditions will complete the proof.

□

4.3 Consistency of Graph Structure Learning with Non-parametric Model

4.3.1 Fréchet derivative

We will review the Fréchet derivative in functional space in this section. More details can be found in Akerkar (1999) [2].

Definition 4.9. ***Fréchet derivative***

Let $\mathbb{X}, \mathbb{Y}$ be Banach spaces; $U \subseteq \mathbb{X}$ be an open subset of $\mathbb{X}$, and $F : U \to \mathbb{Y}$ a map. The Fréchet derivative of F at x_0, $\nabla F(x_0)$, is a linear map from $\mathbb{X}$ to $\mathbb{Y}$ if and only if

$$\lim_{\delta \in \mathbb{X}, \|\delta\| \to 0} \frac{\|F(x_0 + \delta) - F(x_0) - \nabla F(x_0)(\delta)\|}{\|\delta\|} = 0 \tag{4.33}$$

If F is a continuous map, then ∇F is a continuous linear map. Let $\mathbb{L}(\mathbb{X}, \mathbb{Y})$ be the space of linear operators that map elements of $\mathbb{X}$ to $\mathbb{Y}$. Then, $\nabla F(x_0) \in \mathbb{L}(\mathbb{X}, \mathbb{Y})$, and $\nabla F : U \to \mathbb{L}(\mathbb{X}, \mathbb{Y})$. We now define the second order Fréchet derivative.

Definition 4.10. ***Higher order Fréchet derivative***

Let $\mathbb{X}, \mathbb{Y}$ be Banach spaces; $U \subseteq \mathbb{X}$ be an open subset of $\mathbb{X}$; and $F : U \to \mathbb{Y}$ a map. If the Fréchet derivative ∇F is continuous and differentiable at x_0, we write the second order derivative of F at x_0 as $\nabla^2 F(x_0)$, which is the Fréchet derivative of ∇F.

m-th order Fréchet derivative can be defined similarly for $m = 3, 4, \cdots$.

If $\nabla^2 F(x)$ exists on for any $x \in U$, we denote the second order derivative as $\nabla^2 F : U :\to \mathbb{L}(\mathbb{X}, \mathbb{L}(\mathbb{X}, \mathbb{Y}))$. Since $\mathbb{L}(\mathbb{X}, \mathbb{L}(\mathbb{X}, \mathbb{Y}))$ and $\mathbb{L}(\mathbb{X} \times \mathbb{X}, \mathbb{Y})$ are isomorphic, $\nabla^2 F(x_0)$ can be treated as a bilinear operator: $\mathbb{X} \times \mathbb{X} \to \mathbb{Y}$.

We summarize the Taylor's theorem extended to Banach space based on Theorem 2.5 in Akerkar (1999) [2].

Proposition 4.11. ***Taylor's theorem on Banach space***

Let $\mathbb{X}, \mathbb{Y}$ are Banach spaces; $U \subseteq \mathbb{X}$ open; $\{x + \tau\delta | 0 \leq \tau \leq 1\} \subseteq U$; $F : U \to \mathbb{Y}$ be m-times differentiable and the derivatives up to the order of m continuous. Then,

$$\begin{aligned} F(x+\delta) =& F(x) + \nabla F(x)(\delta) + \frac{1}{2}\nabla^2 F(x)(\delta,\delta) \\ &+ \cdots + \frac{1}{p!}\nabla^m F(x)(\delta, \cdots, \delta) + o(\|\delta\|^p) \end{aligned} \tag{4.34}$$

Definition 4.12. ***Partial Fréchet derivative***

Let $\mathbb{X}_1, \mathbb{X}_2, \mathbb{Y}$ be Banach spaces; $U_i \subseteq \mathbb{X}_i$, be an open subset of $\mathbb{X}_i$ for $i = 1, 2$. Let $\mathbb{X} = \mathbb{X}_1 \times \mathbb{X}_2$, $U = U_1 \times U_2$; and $F : U \to \mathbb{Y}$ be differentiable. For any $(x_1, x_2) \in U$, define the partial map of F on the first part (x_1) as

$$\begin{aligned} F_{x_2} &: U_1 \to Y \\ F_{x_2}(x_1) &= F(x_1, x_2) \end{aligned} \tag{4.35}$$

The partial Fréchet derivative of F on the first part at (x_1, x_2) is defined as

$$\begin{aligned} \nabla_1 F(x_1, x_2) &: \mathbb{X}_1 \to Y \\ \nabla_1 F(x_1, x_2)(\delta_1) &= \nabla F_{x_2}(x_1)(\delta_1) \end{aligned} \tag{4.36}$$

The partial Fréchet derivative on the second part, $\nabla_2 F$, can be defined similarly. Also, the second order partial Fréchet derivative, e.g., $\nabla^2_{11} F$, can be defined according to the definition of higher order derivative and partial derivative.

If F is a continuous map, it is continuously differentiable at (x_1, x_2), if and only if F is partially differentiable and the partial derivatives are continuous maps. And we have the following relation:

$$\nabla F(x_1, x_2)(\delta_1, \delta_2) = \nabla_1 F(x_1, x_2)(\delta_1) + \nabla_2 F(x_1, x_2)(\delta_2) \tag{4.37}$$

Now, we will present the final result that is useful in the later derivations. It is the chain rule of Fréchet derivative from the Theorem 2.1 in Akerkar (1999) [2].

Proposition 4.13. ***Chain rule***

Let $\mathbb{X}, \mathbb{Y}, \mathbb{Z}$ be Banach spaces, $U \subseteq \mathbb{X}$, $V \subseteq \mathbb{Y}$ open subsets. Let $F : U \to \mathbb{Y}$, $G : V \to \mathbb{Z}$ be continuous maps, such that $F(U) \subseteq V$. Let $y_0 = F(x_0)$, suppose the Fréchet derivatives $\nabla F(x_0)$ and $\nabla G(y_0)$ exist. Then $G \circ F$ is Fréchet differentiable at x_0 and

$$\nabla(G \circ F)(x_0) = \nabla G(y_0) \circ \nabla F(x_0) \tag{4.38}$$

4.3.2 Differential Calculus of the Loss Functions

We first define the loss function

$$\begin{aligned} L : \mathbb{Z} \times \mathbb{F} &\to \mathbb{R} \\ (Z, f) &\to L(Z; f) \end{aligned} \tag{4.39}$$

where $\mathbb{F}$ is the domain of the second element of L. In the linear case, $\mathbb{F} = \mathbb{R}^{\tilde{K} \cdot \tilde{p}}$.

Assume L satisfies the following properties (similar to the assumption in Section 4.1 but on the $\mathbb{F}$ domain instead of the domain of model parameters in the linear case).

1. $\mathbb{E}_P \|L(Z; f)\| < \infty$ for each $f \in \mathbb{F}$.

2. L is convex and twice-continuously differentiable in the second component.

$$\mathbb{E}_P \|\nabla_f L(Z; f^*)\|^2 < \infty \tag{4.40}$$

3. The risk function, $R(f) = \mathbb{E}_P[L(Z; f)]$ is twice differentiable at f^* and its Hessian is strictly positive definite at f^*:

$$H(f) = \nabla^2_{ff} \mathbb{E}_P[L(Z; f)] \tag{4.41}$$

Suppose we have the random design and each observation is $Z_i = (Y_i, X_i)$, for $i = 1, \cdots, n$. $Y_i \in \mathbb{Y}, X_i \in \mathbb{X}$ are random vectors. Let $\mathbb{Z} = \mathbb{Y} \times \mathbb{X}$, and P be the probability measure on $\mathbb{Z}$. Let $\mathbb{C}(\mathbb{Z}, \mathcal{R})$ be the space of continuous and bounded functions $F \in \mathbb{C}(\mathbb{Z}, \mathcal{R}) : \mathbb{Z} \to \mathcal{R}$, where $\mathcal{R}$ is a Banach space.

Definition 4.14. *Let $\mathbb{Z}, \mathcal{R}$, $\mathbb{C}(\mathbb{Z}, \mathcal{R})$, $\mathcal{H}_K$ as defined before. The expectation operator E_P, and the loss operator L_Z are defined as follows.*

$$E_P : \mathbb{C}(\mathbb{Z}, \mathcal{R}) \to \mathcal{R} \tag{4.42}$$
$$E_P(g) = \mathbb{E}_P(g(Z))$$
$$L_Z : \mathcal{H}_K \to \mathcal{R} \tag{4.43}$$
$$L_Z(f) = L(Z; f)$$

In addition, define the Hessian operator

$$H : \mathcal{H}_K \to \mathbb{L}(\mathcal{H}_K \times \mathcal{H}_K, \mathcal{R} \times \mathcal{R}) \tag{4.44}$$
$$H(f) = \nabla^2(E_P \circ L_{(\cdot)})(f)$$

And the risk operator

$$R : \mathcal{H}_K \to \mathcal{R} \tag{4.45}$$
$$R(f) = E_P \circ L_{(\cdot)}(f)$$

It is obvious that L_Z is convex and twice-continuously Fréchet-differentiable. Since $L_{(\cdot)}(f) \in \mathbb{C}(\mathbb{Z}, \mathcal{R})$, $E_P \circ L_{(\cdot)}(f) = \mathbb{E}_P[L(Z; f)]$ based on the definition of E_P. Note for any linear functional F, we have $\nabla F(f)(g) = F(g)$. And since E_P is a linear operator, we have

$$\begin{aligned} \nabla \mathbb{E}_P[L(Z; f)] &= \nabla(E_p \circ L_{(\cdot)})(f) \\ &= \nabla E_P(L_{(\cdot)}(f)) \circ \nabla L_{(\cdot)}(f) \\ &= E_P \circ \nabla L_{(\cdot)}(f) \end{aligned} \tag{4.46}$$

Similarly, for the second order derivative,

$$\nabla^2(E_P \circ L_{(\cdot)})(f) = E_P \circ \nabla^2 L_{(\cdot)}(f) \tag{4.47}$$

Let f^* be the true model. From the properties of the loss functional L, it is easy to show that $M(f^*) = E_P\|\nabla L_Z(f^*)\|^2 < \infty$ and $H(f^*)$ is strictly positive definite in the sense that for any

$g_1, g_2 \in \mathcal{H}_K$, $H(f^*)(g_1, g_2)$ is positive definite in $\mathcal{R} \times \mathcal{R}$. Note $H(f^*) \in \mathbb{L}(\mathcal{H}_K, \mathbb{L}(\mathcal{H}_K, \mathbb{R}))$, its range is dense and $\inf_{\|f\|=1} \|H(f^*)(f)\| \geq \|H(f^*)(g)\| > 0$ for any $g \in \mathcal{H}_K$ that $g \neq 0$ and $\|g\| \leq 1$. Since a linear operator F is invertible if and only if its range is dense and bounded from below (Halmos (1998) [30]), we know $H(f^*)$ is invertible.

To understand the inverse of the bounded bilinear operator $H(f^*)$, we denote A as the Reisz representer of $H(f^*)$ which is a bounded linear operator from $\mathcal{H}_K$ onto itself. That is, for any $f, g \in \mathcal{H}_K$, $H(f^*)(f, g) = \langle Af, g \rangle_{\mathcal{H}_K}$. Let $G = H(f^*)(f, \cdot)$ which is a bounded linear operator from $\mathcal{H}_K$ onto $\mathbb{R}$, and η_G the Reize representer of G such that $G = \langle \eta_G, \cdot \rangle_{\mathcal{H}_K}$. Then $f = A^{-1}\eta_G := H(f^*)^{-1}(G)$. In a spectial case when $G = \langle \gamma, \cdot \rangle_{\mathcal{H}_K}$, $H(f^*)^{-1}(\langle \gamma, \cdot \rangle_{\mathcal{H}_K}) = A^{-1}\gamma$.

The proofs in the following section needs the strong law of large numbers in Banach space. See Ledoux and Talagrand (1991) [49] for more details. The theorem requires $\mathcal{H}_K$ be a separable Hilbert space, or equivalently, requires K be square integrable, by Mercer-Hilbert-Schmidt theorem.

4.3.3 Consistency Results for Reproducing Kernel Hilbert Space

To show the sparsistency results in Reproducing Kernel Hilbert Space, we need the general version of Lemma 4.3. Lemma 4.16 will give the asymptotic distribution of the estimated function $\hat{f}_n$. Before that, we will present the following lemma about the convergence of a sequence of Hilbert space valued random variables as the estimates of a sequence of essentially strictly convex objective functionals.

Let $(\mathcal{E}, \mathscr{E}, P)$ be a probability space, $\mathcal{H}_K$ be a Reproducing Kernel Hilbert Space with kernel K. $W_n, n = 1, 2, \cdots$ and W are random functionals defined on $H \times \mathcal{E}$ to $\mathbb{R}$. We usually denote $W(\delta, \cdot)$ or $W(\delta)$ for the random functional, $W(\cdot, e)$ for $e \in \mathcal{E}$ as a realized functional on $\mathcal{H}_K$. We are interested in the convergence of the approximation of $\hat{\delta} = \arg\min_{\delta \in \mathcal{H}_K} W(\delta)$ by a sequence $\hat{\delta}_n = \arg\min_{\delta \in \mathcal{H}_K} W_n(\delta)$. For more about the random functions and the convergence results, refer to Korf and Wets (2001) [46] and Vogel and Lachout (2003) [88], and the references therein.

Lemma 4.15. *Let W_n $n = 1, 2, \cdots$ and W be random functions defined on $\mathcal{H}_K \times \mathcal{E}$ to $\mathbb{R}$. Suppose W_n and W are continuous and essentially strictly convex on $\mathcal{H}_K$. If W_n point-wisely converges in*

Table 4.1 Summary of the functional operators when $\mathcal{R} = \mathbb{R}$

Operator	Mapping
$L_Z(\cdot) \in \mathbb{C}(\mathcal{H}_K, \mathbb{R})$	$\mathcal{H}_K \to \mathbb{R}$
$L_{(\cdot)}(f) \in \mathbb{C}(\mathbb{Z}, \mathbb{R})$	$\mathbb{Z} \to \mathbb{R}$
$\nabla L_{(\cdot)}(f) \in \mathbb{C}(\mathbb{Z}, \mathbb{L}(\mathcal{H}_K, \mathbb{R}))$	$\mathbb{Z} \to \mathbb{L}(\mathcal{H}_K, \mathbb{R})$
$E_P \circ \nabla L_{(\cdot)}(f) \in \mathbb{L}(\mathcal{H}_K, \mathbb{R})$	$\mathcal{H}_K \to \mathbb{R}$
$V_P \circ \nabla L_{(\cdot)}(f) \in \mathbb{L}(\mathcal{H}_K \times \mathcal{H}_K, \mathbb{R})$	$\mathcal{H}_K \times \mathcal{H}_K \to \mathbb{R}$
$E_P \circ \nabla^2 L_{(\cdot)}(f) \in \mathbb{L}(\mathcal{H}_K \times \mathcal{H}_K, \mathbb{R})$	$\mathcal{H}_K \times \mathcal{H}_K \to \mathbb{R}$
$H(\cdot) \in \mathbb{L}(\mathcal{H}_K \times \mathcal{H}_K, \mathbb{R})$	$\mathcal{H}_K \times \mathcal{H}_K \to \mathbb{R}$
$H(f)(\cdot) \in \mathbb{L}(\mathcal{H}_K, \mathbb{R})$	$\mathcal{H}_K \to \mathbb{R}$
$H^{-1}(\cdot) \in \mathbb{L}(\mathbb{L}(\mathcal{H}_K, \mathbb{R}), \mathcal{H}_K)$	$\mathbb{L}(\mathcal{H}_K, \mathbb{R}) \to \mathcal{H}_K$
$M(\cdot) \in \mathbb{L}(\mathcal{H}_K \times \mathcal{H}_K, \mathbb{R})$	$\mathcal{H}_K \times \mathcal{H}_K \to \mathbb{R}$
$R(\cdot) \in \mathbb{C}(\mathcal{H}_K, \mathbb{R})$	$\mathcal{H}_K \to \mathbb{R}$

probability to W, i.e., for any $\delta \in \mathcal{H}_K$, $W_n(\delta) \xrightarrow{p} W(\delta)$, then,

$$\arg\min_{\delta \in \mathcal{H}_K} W_n(\delta) \xrightarrow{d} \arg\min_{\delta \in \mathcal{H}_K} W(\delta) \tag{4.48}$$

Proof. See Appendix A.6. □

Lemma 4.16. *Suppose λ_n is a sequence of positive values which satisfies $\lambda_n \to 0$ and $\lambda_n \sqrt{n} \to \infty$ as $n \to \infty$. Let f^* denote the true model in $\mathcal{H}_K$, $\hat{f}_n = \arg\min_{f \in \mathcal{H}_K} \mathcal{I}_{\lambda_n}(f)$, $H(f^*)$ is the Hessian as defined before. Then,*

$$\frac{1}{\lambda_n}(\hat{f}_n - f^*) \xrightarrow{d} \hat{\delta} = \arg\min_{\delta \in \mathcal{H}_K} W(\delta) = \frac{1}{2} H(f^*)(\delta, \delta) + \left\langle \gamma^{\mathcal{A}}, \delta^{\mathcal{A}} \right\rangle_{\mathcal{H}_K} + \mathcal{J}_{\mathcal{A}^c}(\delta^{\mathcal{A}^c}) \tag{4.49}$$

with $\gamma^{\mathcal{A}}$ defined in the following equation when $f = f^*$.

$$\gamma^{\mathcal{A}} = (\gamma^{\omega})_{\omega \in \mathcal{A}}, \quad \gamma^{\omega} = f^{\omega} \sum_{v \subseteq \omega} \frac{p_v}{\|f^{T_v}\|_{\mathcal{H}_{T_v}}} \tag{4.50}$$

Proof. See Appendix A.7. □

Before presenting the consistency results, we define the conjugate $\mathcal{J}$-norm on the linear operator from $\mathcal{H}_K$ onto $\mathbb{R}$.

Definition 4.17. *Let $F : \mathcal{H}_K \to \mathcal{R}$ be a linear operator, $\mathcal{J}$ is a norm on $\mathcal{H}_K$, the conjugate $\mathcal{J}$-norm is defined as*

$$\|F\|_{\mathcal{J}} = \max_{f \in \mathcal{H}_K, \mathcal{J}(f) \leq 1} \|F(f)\|_{\mathcal{R}} \tag{4.51}$$

The conjugate $\mathcal{J}_{\mathcal{A}^c}$-norm can be defined naturally.

Theorem 4.18. ***Necessary condition for RKHS***

Let λ_n, f^* and $H(f^*)$ as defined in Lemma 4.16, $\mathcal{A}, \hat{\mathcal{A}}_n$ defined in Equation (4.2) for $f = f^*$ and $f = \hat{f}_n$ respectively. Let $H_{\mathcal{A}\mathcal{A}}$ and $H_{\mathcal{A}^c\mathcal{A}}$ be the second order partial derivative of L_Z at f^*. If $\mathcal{A}$ is estimated consistently, that is, $\mathbb{P}(\hat{\mathcal{A}}_n = \mathcal{A}) \to 1$ as $n \to \infty$, where $\hat{\mathcal{A}}_n = \hat{\mathcal{P}}_n = \{\omega | \hat{f}_n^{\omega} \neq 0\}$, then $\left\| H_{\mathcal{A}^c\mathcal{A}}(f^*) \circ H_{\mathcal{A}\mathcal{A}}(f^*)^{-1} \left(\left\langle \gamma^{\mathcal{A}}, \cdot \right\rangle_{\mathcal{H}_{\mathcal{A}}} \right) \right\|_{\mathcal{J}_{\mathcal{A}^c}} \leq 1$.

Proof. Since $H(f^*)$ is a symmetric bilinear operator, we have $\nabla H(f^*)(\delta,\delta) : \mathcal{H}_K \times \mathcal{H}_K \to \mathbb{R}$. For any $g \in \mathcal{H}_K$,

$$\begin{aligned}\nabla H(f^*)(\delta,\delta)(g,g) &= \nabla_1 H(f^*)(\delta,\delta)(g) + \nabla_2 H(f^*)(\delta,\delta)(g) \\ &= H(f^*)(g,\delta) + H(f^*)(\delta,g) = 2H(f^*)(g,\delta)\end{aligned} \tag{4.52}$$

We can view $H(f^*)$ as an operator from $\mathcal{H}_K \to \mathbb{R}$ such that it maps any $\delta \in \mathcal{H}_K$ to $H(f^*)(\delta,\delta)$. Then $\nabla H(f^*)(\delta,\delta) : \mathcal{H}_K \to \mathbb{R}$ such that for any $g \in \mathcal{H}_K$, $\nabla H(f^*)(\delta,\delta)(g) = 2H(f^*)(g,\delta)$. In addition, $H(f^*)(\delta) = H(f^*)(\delta,\cdot)$ can be viewed as an linear operator from $\mathcal{H}_K$ onto $\mathbb{R}$. And all these definitions and results can be natually applied to the second order partial derivatives $H_{\mathcal{A}\mathcal{A}}(f^*)$ and $H_{\mathcal{A}^c\mathcal{A}}(f^*)$.

Denote $F_{\gamma^{\mathcal{A}}} = \left\langle \gamma^{\mathcal{A}}, \cdot \right\rangle_{\mathcal{H}_{\mathcal{A}}}$ as a linear operator from $\mathcal{H}_{\mathcal{A}}$ onto $\mathbb{R}$. The generalized KKT condition (Luenberger (1997) [57]) of Equation (4.49) is

$$H_{\mathcal{A}\mathcal{A}}(f^*)(\delta^{\mathcal{A}}) + F_{\gamma^{\mathcal{A}}} = 0 \tag{4.53}$$

$$H_{\mathcal{A}^c\mathcal{A}}(f^*)(\delta^{\mathcal{A}}) + \sum_{v\in\mathcal{A}^c} s_v = 0 \tag{4.54}$$

In the above equation, s_v is an operator from $\mathcal{H}_{\mathcal{A}^c}$ to $\mathbb{R}$ defined below:

$$\begin{aligned}s_v \in \mathbb{S}_v = \{&s = (s^\omega)_{\omega\in\mathcal{A}^c} \mid s \in \mathbb{L}(\mathcal{H}_{\mathcal{A}^c}, \mathbb{R}), \\ &|s(\delta^{\mathcal{A}^c})| \le p_v \|I_{T_v}(\delta^{\mathcal{A}^c})\|_{\mathcal{H}_{T_v}} \text{ for any } \delta^{\mathcal{A}^c} \in \mathcal{H}_{\mathcal{A}^c}, \text{ and } s^\omega = 0 \text{ if } \omega \notin T_v\}\end{aligned} \tag{4.55}$$

where $I_{T_v} : \mathcal{H}_{\mathcal{A}^c} \to \mathcal{H}_{\mathcal{A}^c}$ is a linear operator such that for any $f \in \mathcal{H}_{\mathcal{A}^c}$, $(I_{T_v}(f))^\omega = 0$ if $\omega \notin T_v$; and $(I_{T_v}(f))^\omega = f^\omega$ if $\omega \in T_v$.

Since $H_{\mathcal{A}\mathcal{A}}(f^*) \in \mathbb{L}(\mathcal{H}, \mathbb{L}(\mathcal{H}, \mathbb{R}))$ is invertible, we have $\delta^{\mathcal{A}} = H_{\mathcal{A}\mathcal{A}}(f^*)^{-1}(F_{\gamma^{\mathcal{A}}})$. Then,

$$H_{\mathcal{A}^c\mathcal{A}}(f^*) \circ H_{\mathcal{A}\mathcal{A}}(f^*)^{-1}(F_{\gamma^{\mathcal{A}}}) + \sum_{v\in\mathcal{A}^c} s_v = 0 \tag{4.56}$$

The following inequality completes the proof:

$$
\begin{aligned}
|H_{\mathcal{A}^c\mathcal{A}}(f^*) \circ H_{\mathcal{A}\mathcal{A}}(f^*)^{-1}(F_{\gamma^{\mathcal{A}}})(f^{\mathcal{A}^c})| &= |\sum_{v\in\mathcal{A}^c} s_v(f^{\mathcal{A}^c})| \\
&\leq \sum_{v\in\mathcal{A}^c} |s_v(f^{\mathcal{A}^c})| \\
&\leq \sum_{v\in\mathcal{A}^c} p_v \|I_{T_v}(f^{\mathcal{A}^c})\|_{\mathcal{H}_{T_v}} = \mathcal{J}_{\mathcal{A}^c}(f^{\mathcal{A}^c})
\end{aligned} \tag{4.57}
$$

holds for any $f^{\mathcal{A}^c} \in \mathcal{H}_{\mathcal{A}^c}$. □

Theorem 4.19. ***Sufficient condition for RKHS***

Let λ_n, f^* and $H(f^*)$ as defined in Lemma 4.16. If $\left\| H_{\mathcal{A}^c\mathcal{A}}(f^*) \circ H_{\mathcal{A}\mathcal{A}}(f^*)^{-1} \left(\left\langle \gamma^{\mathcal{A}}, \cdot \right\rangle_{\mathcal{H}_{\mathcal{A}}} \right) \right\|_{\mathcal{J}_{\mathcal{A}^c}} < 1$, then $\mathcal{A}$ is consistently estimated in the sense that $\mathbb{P}(\hat{\mathcal{A}}_n = \mathcal{A}) \to 1$ as $n \to \infty$, where $\hat{\mathcal{A}}_n = \hat{\mathcal{P}}_n = \{v | \|\hat{f}_n^v\| \neq 0\}$.

Proof. We prove the result similarly as we did in the linear case.

Let $\hat{f}_n^{\mathcal{A}}$ be the solution of the restricted problem as defined in Equation (4.18), and pad $\hat{f}_n^{\mathcal{A}}$ with constant zero functions on $\mathcal{A}^c$ to obtain $\hat{f}_n$. From Lemma 4.5, $\hat{\mathcal{A}}_n \xrightarrow{p} \mathcal{A}$. And applying similar techinques, we have $\hat{f}_n^{\mathcal{A}} \xrightarrow{p} f^{\mathcal{A}}$. Thus, to prove the conclusion, we need to show that $\hat{f}_n$ satisfies the optimality condition of objective in Equation (3.5).

For large n, $\hat{\gamma}^{\mathcal{A}}$ is well defined similarly to Equation (4.50)

$$
\hat{\gamma}^{\mathcal{A}} = (\hat{\gamma}^{\omega})_{\omega\in\mathcal{A}} \quad \text{where } \hat{\gamma}^{\omega} = \hat{f}_n^{\omega} \sum_{v\subseteq\omega} \frac{p_v}{\|\hat{f}_n^{T_v}\|_{\mathcal{H}_{T_v}}} \tag{4.58}
$$

and $\hat{\gamma}^{\mathcal{A}} \xrightarrow{p} \gamma^{\mathcal{A}}$.

The optimality condition on $\mathcal{A}$ is already satisfied due to the definition of $\hat{f}_n$, which implies

$$
\left(\nabla \mathcal{L}(\hat{f}_n) \right)_{\mathcal{A}} + \lambda_n \left\langle \hat{\gamma}^{\mathcal{A}}, \cdot \right\rangle_{\mathcal{H}_{\mathcal{A}}} = 0 \tag{4.59}
$$

It remains to show that there exist s_v's as defined in Equation (4.55) such that

$$
\left(\nabla \mathcal{L}(\hat{f}_n) \right)_{\mathcal{A}^c} + \lambda_n \sum_{v\in\mathcal{A}^c} s_v = 0 \tag{4.60}
$$

That is, we need to show

$$\left\| \left(\nabla \mathcal{L}(\hat{f}_n) \right)_{\mathcal{A}^c} \right\|_{\mathcal{J}_{\mathcal{A}^c}} < \lambda_n \tag{4.61}$$

Let $\hat{\delta}_n = \hat{f}_n - f^* \xrightarrow{p} 0$. Note

$$\begin{aligned} \nabla \mathcal{L}(\hat{f}_n) &= \frac{1}{n} \sum_{i=1}^{n} \nabla L_{Z_i}(\hat{f}_n) \\ &= \left[\frac{1}{n} \sum_{i=1}^{n} \nabla L_{Z_i}(f^*) \right] + \left[\frac{1}{n} \sum_{i=1}^{n} \nabla^2 L_{Z_i}(f^*)(\hat{\delta}_n) \right] + o_p(\|\hat{\delta}_n\|) \\ &= D_n + H_n(\hat{\delta}_n) + o_p(\|\hat{\delta}_n\|) \end{aligned} \tag{4.62}$$

As we have shown in Lemma 4.16, $D_n \xrightarrow{p} 0$ and $H_n \xrightarrow{p} H(f^*)$.

Since $\hat{f}_n^{\mathcal{A}^c} = f^{*\mathcal{A}^c} = 0$, we have from the above equation that

$$\left(\nabla \mathcal{L}(\hat{f}_n) \right)_{\mathcal{A}} = H_{\mathcal{A}\mathcal{A}}(f^*)(\hat{\delta}_n^{\mathcal{A}}) + o_p(1) \tag{4.63}$$

$$\left(\nabla \mathcal{L}(\hat{f}_n) \right)_{\mathcal{A}^c} = H_{\mathcal{A}^c\mathcal{A}}(f^*)(\hat{\delta}_n^{\mathcal{A}}) + o_p(1) \tag{4.64}$$

From Equation (4.63) and Equation (4.59) we have

$$\hat{\delta}_n = -\lambda_n H_{\mathcal{A}\mathcal{A}}(f^*)^{-1}(\hat{F}_{\gamma^{\mathcal{A}}}) + o_p(1) \tag{4.65}$$

where

$$\hat{F}_{\gamma^{\mathcal{A}}} = \left\langle \hat{\gamma}^{\mathcal{A}}, \cdot \right\rangle_{\mathcal{H}_{\mathcal{A}}} \quad \text{and} \quad \hat{F}_{\gamma^{\mathcal{A}}} \xrightarrow{p} F_{\gamma^{\mathcal{A}}} = \left\langle \gamma^{\mathcal{A}}, \cdot \right\rangle_{\mathcal{H}_{\mathcal{A}}} \tag{4.66}$$

Then, the following equality holds because $\hat{\gamma}^{\mathcal{A}} \xrightarrow{p} \gamma^{\mathcal{A}}$

$$\left(\nabla \mathcal{L}(\hat{f}_n) \right)_{\mathcal{A}^c} = -\lambda_n H_{\mathcal{A}^c\mathcal{A}}(f^*) \circ H_{\mathcal{A}\mathcal{A}}(f^*)^{-1}(F_{\gamma^{\mathcal{A}}}) + o_p(1) \tag{4.67}$$

Therefore, for any $f^{\mathcal{A}^c} \in \mathcal{H}_{\mathcal{A}^c}$, the following inequality holds because of the sufficient condition

$$\begin{aligned} \left| \left\langle f^{\mathcal{A}^c}, \left(\nabla \mathcal{L}(\hat{f}_n) \right)_{\mathcal{A}^c} \right\rangle_{\mathcal{H}_{\mathcal{A}^c}} \right| &= \lambda_n \left| \left\langle f^{\mathcal{A}^c}, H_{\mathcal{A}^c\mathcal{A}}(f^*) \circ H_{\mathcal{A}\mathcal{A}}(f^*)^{-1} \left(\left\langle \gamma^{\mathcal{A}}, \cdot \right\rangle_{\mathcal{H}_{\mathcal{A}}} \right) \right\rangle_{\mathcal{H}_{\mathcal{A}^c}} \right| + o_p(1) \\ &< \lambda_n \mathcal{J}_{\mathcal{A}^c}(f^{\mathcal{A}^c}) \quad \text{for large } n \end{aligned} \tag{4.68}$$

This completes the proof. □

Chapter 5

Numerical Studies

5.1 Simulations

5.1.1 Simulation Settings

In the simulation, we create 6 graphs. The first four graphs are depicted in Figure 2.1. Graph 5 has 100 nodes where the first 8 nodes have the same structure as in Figure 2.1(c) and the others are independent. Graph 6 also has 100 nodes where the first 10 nodes have the same connection as in Figure 2.1(d) and the others are independent.

We generate 100 independent datasets for each experiment with the same setting, and evaluate the performance based on the averaged results on the 100 independent runs. Here is how the first data set is generated:

The length of the feature vector, p, is set to 0 or 5 in our experiment. When $p = 0$, we are considering the graphical models without input features. For $p = 5$, $X = (X_1, \ldots, X_5)$, each $f^{\omega}(x) = c_0^{\omega} + \sum_{j=1}^{5} c_j^{\omega} x_j$, for $\omega \in \wp(\Omega)$. The true sets of the model parameters, c_{jk}^{ω}, are provided in Appendix B. The features, X_j, are i.i.d uniform on [-1, 1]. Y is sampled according to the probability in Equation (2.4). Gibbs sampling is applied for Graph 5 and 6.

We use BGACV (B-type generalized approximate cross validation GACV) (Xiang and Wahba (1996) [97], Shi et al. (2008) [79]) to choose the regularization parameter λ for the complete model (graphs 1-4). The performance of choosing the tuning parameter by GACV is not presented here because it is comparable to BGACV in terms of recovering the true non-zero patterns, but with more false detections. We use BIC for greedy search algorithm in Graph 5 and 6 due to the

computational consideration. The range of λ is chosen according to Koh et al. (2007) [43]. The details of the tuning methods are discussed in Section 3.4.

5.1.2 Estimation Consistency of SLasso

We evaluate the graph structure estimation accuracy of our SLasso method, and compared it to two closely related graph structure learning methods. Höfling and Tibshirani (2009) [33] proposed using pseudo-likelihood with l_1 penalty for estimating sparse pairwise binary Markov models. They only consider pairwise interactions, and there are no input features involved. The method is published as an R package, BMN[1]. Schmidt and Murphy (2010) [75] considered the problem of learning higher-order graphical structure without features. They used the log-linear models and overlapping penalties. Their code, LLM, is published online[2]. We choose the tuning parameters for BMN and LLM by cross validation.

5.1.2.1 When $p = 0$

To make a fair comparison, we first let $p = 0$, which corresponds to the graphical models without input features.

In Table 5.1, we count, for each conditional log odds ratio, f^ω, the number of runs out of 100 where f^ω is recovered ($\|c^\omega\| \neq 0$). If a recovered f^ω is in the true model, it is considered as true positive, otherwise false positive. The sample size is 1000. We list in the table the average discovery rate on a selected subset of the non-zero conditional log odds ratios in the true model. The last column is the average discovery rate of all the zero patterns in the true model. The main effects are always detected correctly, thus, are not listed in the table. LLM takes too long to converge on Graph 5 and 6. So the corresponding results are not provided.

According to Table 5.1, BMN, LLM, and SLasso achieve very similar results on the simplest graph (Graph 1). On Graph 3-6, SLasso is more effective compared to BMN and LLM. BMN cannot detect higher order interactions because it only considers the pairwise interactions.

[1] `http://cran.r-project.org/web/packages/BMN/index.html`
[2] `http://www.di.ens.fr/~mschmidt/Software/thesis.html`

Table 5.1 The average discovery rate of a selected set of the non-zero patterns when $p = 0$, $n = 1000$. The last column, FP (False Positive), is the average discovery rate of all the zero patterns in the true model. Note, the numbers of the zero patterns in the true model for the 6 graphs are 6, 51, 231, 994, 10^{30}, and 10^{30} respectively.

Graph	Method	$f^{1,2}$	$f^{1,3}$	$f^{2,3}$	$f^{3,4}$	$f^{1,2,3}$	$f^{5,7,8}$	$f^{5,6,7,8}$	FP
1	BMN	1.00	1.00	1.00	1.00	0	-	-	0.25
	LLM	1.00	1.00	1.00	1.00	0.93	-	-	0.96
	SLasso	1.00	1.00	1.00	1.00	1.00	-	-	0.48
2	BMN	0.98	0.53	0.16	0.21	0	-	-	0
	LLM	1.00	1.00	0.89	1.00	0.89	-	-	7.61
	SLasso	0.94	0.92	0.90	0.96	0.90	-	-	0.94
3	BMN	0.72	0.75	0.24	0.34	0	0	0	0.01
	LLM	1.00	1.00	1.00	1.00	0.50	0.70	0.13	11.96
	SLasso	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.68
4	BMN	0.81	0.96	0.99	0.01	0	0	0	0.02
	LLM	1.00	0.99	1.00	0.96	0.10	0.22	0	2.09
	SLasso	1.00	1.00	1.00	0.94	1.00	1.00	0.20	0.98
5	BMN	0.41	0.24	0.05	0.07	0	0	0	0.91
	SLasso	0.99	0.99	0.98	0.97	0.95	0.92	0	3.97
6	BMN	0.29	0.78	0.71	0	0	0	0	0.01
	SLasso	1.00	0.95	1.00	0.99	0.94	0.82	0	3.58

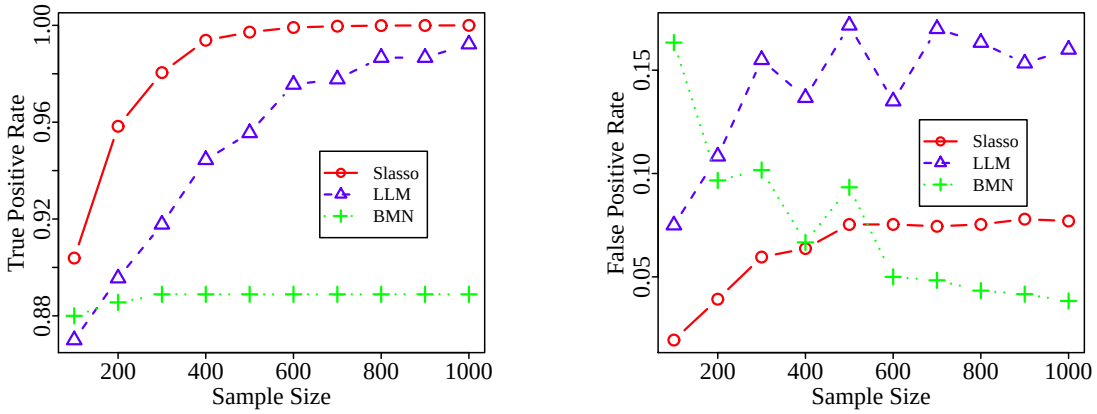
SLasso achieves relative better performance on true positive rate as well as false positive rate compared to LLM.

Now, we evaluate the true positive rate and false positive rate of recovering the conditional log odds ratios in the model, with increasing sample size. In Figure (5.1) - (5.4), we show the learning results in terms of true positive rate (TPR) and false positive rate (FPR) as sample size increases from 100 to 1000. Subfigure (a) and (b) are measured on the unit of the conditional log odds ratios (some times are called patterns). Subfigure (c) and (d) are measured on the unit of the cliques. There are 2, 3, 3, 6, 3, 6 cliques in the true models, respectively. We consider all the possible cliques of any size in the graph. The total number in a graph of K nodes is $2^K - 1$. We calculate the TPR on cliques by dividing the average overall number of correctly discovered cliques by the number of cliques in the graph. The FPR on cliques is calculated by dividing the average overall number of false discovered cliques by the number of nonexistent cliques in the graph. As we discussed in Chapter 2, the graph structure and the conditional independence is determined by the cliques. And according to the asymptotic analysis in Chapter 4, the estimation of SLasso is consistent in terms of cliques. So, the results on cliques are important criteria for evaluating the estimation consistency of a graph structure learning method. Since BMN does not consider interactions higher than second order, we will not include its TPR/FPR on the unit of cliques.

The experimental settings are the same as before. The true model parameters are listed in Appendix B.1. In these figures, we can see that SLasso achieves satisfying performance in TPR with FPR well controlled. With increasing sample size, the estimate of SLasso is getting close to the true graph structure.

In Graph 2, LLM outperforms SLasso in terms of pattern TPR, at the cost of high pattern FPR. As a result, more noisy cliques (possibly larger cliques) are recovered by LLM, which causes the worse performance of clique TPR and clique FPR compared to SLasso . SLasso outperforms LLM in other scenarios.

BMN has good performance on simpler graphs (Graph 1 and 2). However, it misses many pairwise interactions when graphs are large and contain higher order interactions.



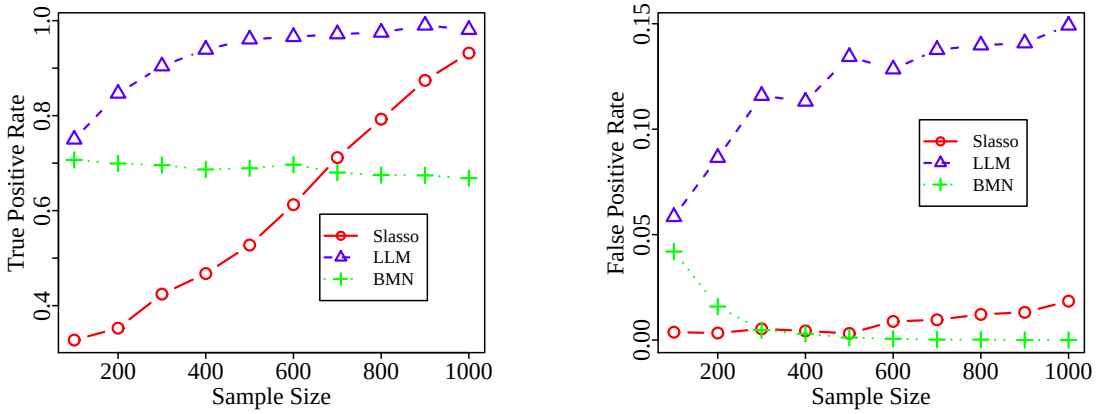
(a) True Positive Rate - Pattern

(b) False Positive Rate - Pattern

(c) True Positive Rate - Clique

(d) False Positive Rate - Clique

Figure 5.1 Graph 1, $p = 0$



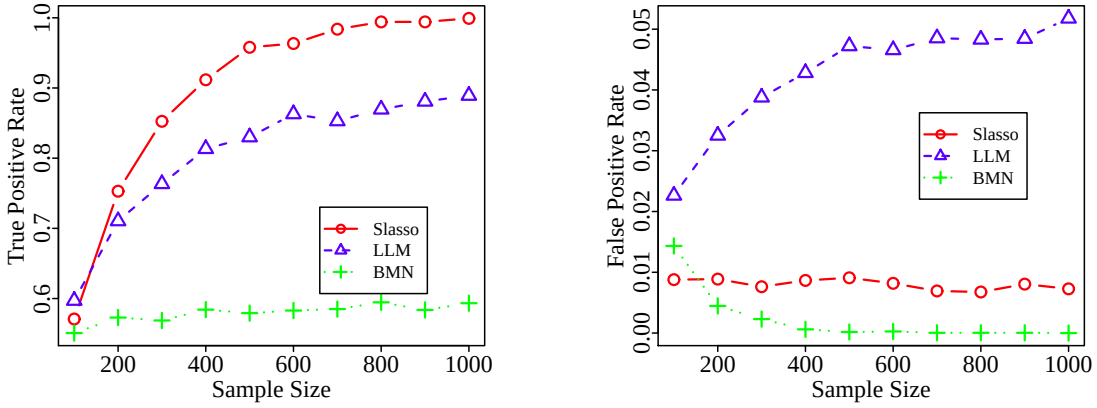
(a) True Positive Rate - Pattern

(b) False Positive Rate - Pattern

(c) True Positive Rate - Clique

(d) False Positive Rate - Clique

Figure 5.2 Graph 2, $p = 0$



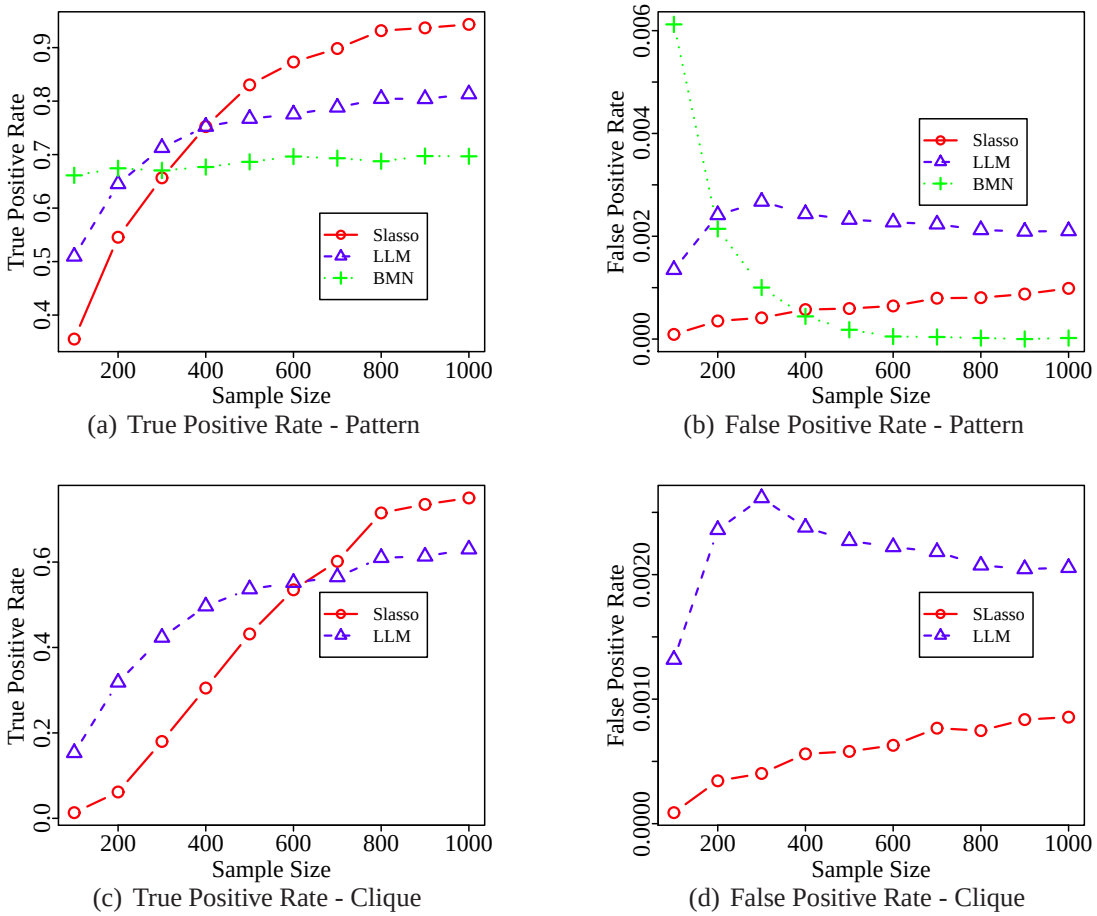
(a) True Positive Rate - Pattern

(b) False Positive Rate - Pattern

(c) True Positive Rate - Clique

(d) False Positive Rate - Clique

Figure 5.3 Graph 3, $p = 0$



(a) True Positive Rate - Pattern

(b) False Positive Rate - Pattern

(c) True Positive Rate - Clique

(d) False Positive Rate - Clique

Figure 5.4 Graph 4, $p = 0$

5.1.2.2 When $p = 5$

Here, we let the input feature be a vector of 5 dimensions. In Table 5.2, BMN and LLM are able to recover the pairwise interactions, but they cannot find higher order interactions effectively. In addition, BMN and LLM will detect many false positive patterns which are pairwise. This is mainly because of the effects of input features. In contrast, SLasso can effectively exploit the features to achieve good performance as it did when there is no feature.

The Figure (5.5) - (5.8) show the convergence of SLasso in estimating the graph structure. It achieves high TPR with FPR well controlled. LLM obtains high TPR at the cost of high FPR.

5.1.3 SLasso with Feature Selection

In this section, we evaluate the performance of SLasso with feature selection. The objective is in the following equation where the penalties are defined in Equation (3.9).

$$\mathcal{I}(f) = \mathcal{L}(f) + \lambda \left(\sum_{v \in \wp(\Omega)} p_v \sqrt{\sum_{\omega \in T_v} \|f^\omega\|^2_{\mathcal{H}_\omega}} + \bar{\lambda} \sum_{v \in \wp(\Omega)} \|c^v\|_1 \right) \tag{5.1}$$

The experiment is performed on Graph 3, with the same settings as before. The true parameters are the same as in Appendix B.2, except that c_1^ω and c_3^ω are set as 0 for all ω, i.e. X_1 and X_3 are irrelevant variables. The second tuning parameter $\bar{\lambda}$ is chosen to be 0.06 based on empirical results.

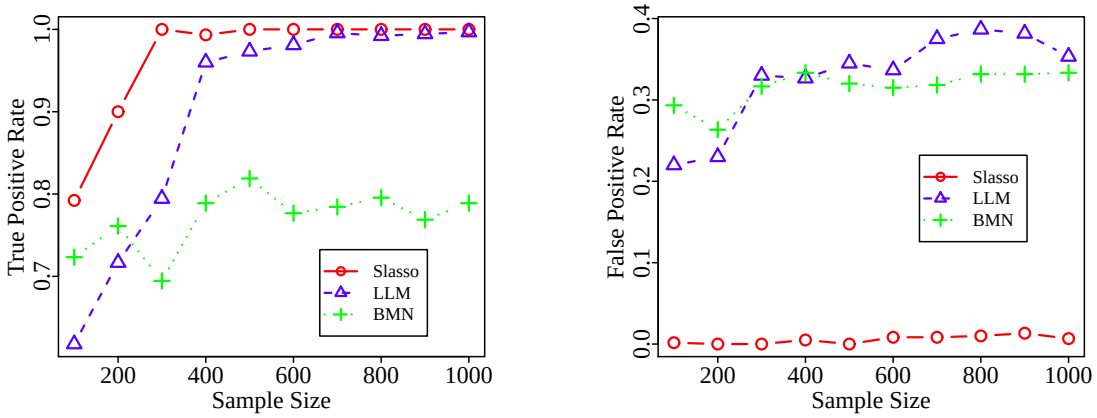
For each sample size ranging from 100 to 1000, we carry out 100 independent runs and average the results. We count the correctly and incorrectly discovered parameters and compare them to the true models. The true positive rate and false positive rate are ploted in the figure. We can see that with increasing sample size, SLasso recover the non-zero parameters more accurately.

5.1.4 Comparison with Ordinary Lasso

In this section, we compare the SLasso method with the ordinary Lasso for multivariate Bernoulli data, which we call "Vanilla" in Equation (5.2). For each of the sample sizes $100, 200, \cdots, 1000$,

Table 5.2 The average discovery rate of a selected set of the non-zero patterns when $p = 5$, $n = 1000$. The last column, FP (False Positive), is the average discovery rate of all the zero patterns in the true model. Note, the numbers of the zero patterns in the true model for the 6 graphs are 6, 51, 231, 994, 10^{30}, and 10^{30} respectively.

Graph	Method	$f^{1,2}$	$f^{1,3}$	$f^{2,3}$	$f^{3,4}$	$f^{1,2,3}$	$f^{5,7,8}$	$f^{5,6,7,8}$	FP
1	BMN	1.00	1.00	0.97	0.13	0	-	-	2.00
	LLM	1.00	1.00	1.00	0.97	1.00	-	-	2.12
	SLasso	1.00	1.00	1.00	1.00	1.00	-	-	0.04
2	BMN	0	1.00	1.00	1.00	0	-	-	3.14
	LLM	0.63	1.00	1.00	1.00	0	-	-	18.12
	SLasso	1.00	0.95	1.00	1.00	0.95	-	-	0.96
3	BMN	1.00	0.99	1.00	1.00	0	0	0	1.57
	LLM	1.00	1.00	1.00	1.00	0.09	0.76	0	17.17
	SLasso	1.00	1.00	1.00	0.98	1.00	1.00	1.00	0.35
4	BMN	1.00	1.00	1.00	1.00	0	0	0	9.10
	LLM	1.00	1.00	1.00	1.00	0.16	0	0	16.45
	SLasso	1.00	1.00	1.00	0.86	1.00	0.99	0.15	0.24
5	BMN	0.67	0.78	0.45	0.54	0	0	0	4.67
	SLasso	0.99	0.99	0.98	0.97	0.80	0.71	0	1.97
6	BMN	0.72	0.85	0.64	0.54	0	0	0	6.17
	SLasso	1.00	1.00	1.00	0.99	0.94	0.85	0	1.58



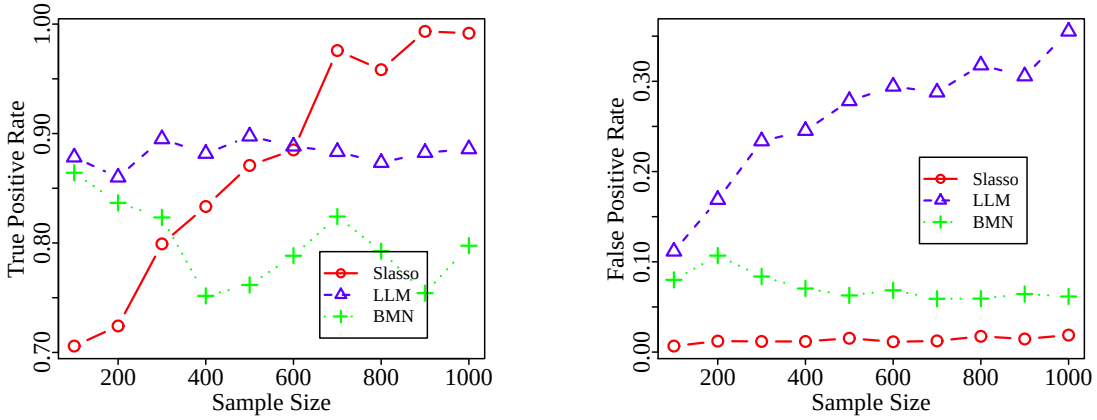
(a) True Positive Rate - Pattern

(b) False Positive Rate - Pattern

(c) True Positive Rate - Clique

(d) False Positive Rate - Clique

Figure 5.5 Graph 1, $p = 5$



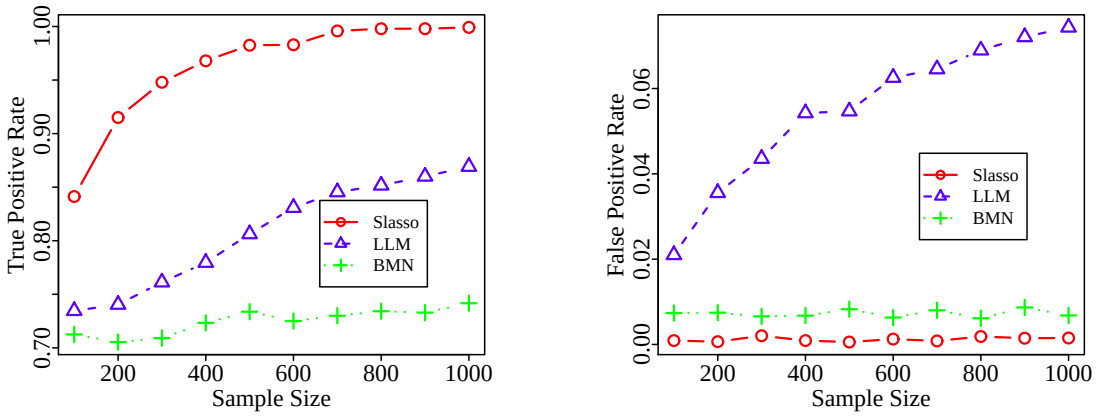
(a) True Positive Rate - Pattern

(b) False Positive Rate - Pattern

(c) True Positive Rate - Clique

(d) False Positive Rate - Clique

Figure 5.6 Graph 2, $p = 5$



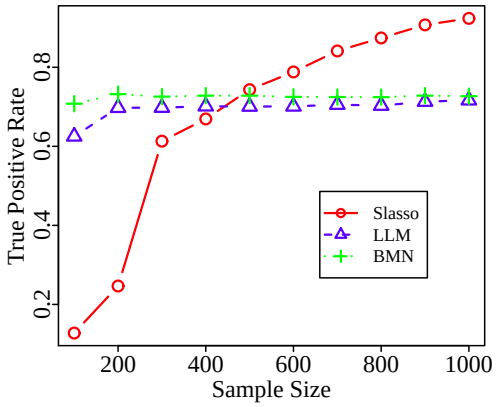
(a) True Positive Rate - Pattern

(b) False Positive Rate - Pattern

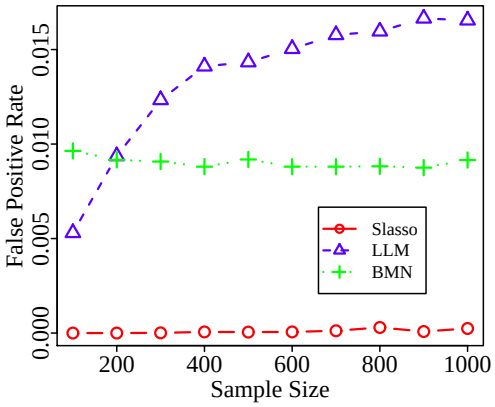
(c) True Positive Rate - Clique

(d) False Positive Rate - Clique

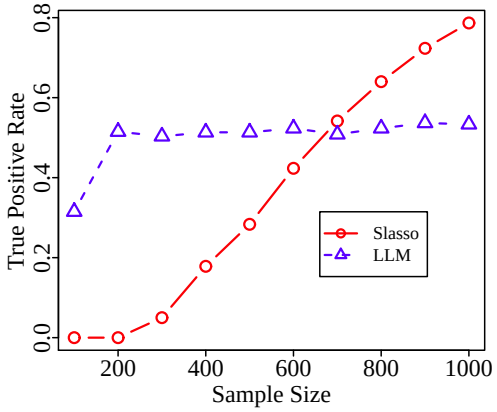
Figure 5.7 Graph 3, $p = 5$



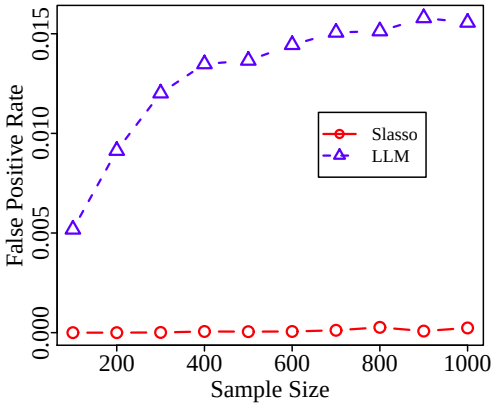
(a) True Positive Rate - Pattern



(b) False Positive Rate - Pattern



(c) True Positive Rate - Clique



(d) False Positive Rate - Clique

Figure 5.8 Graph 4, $p = 5$

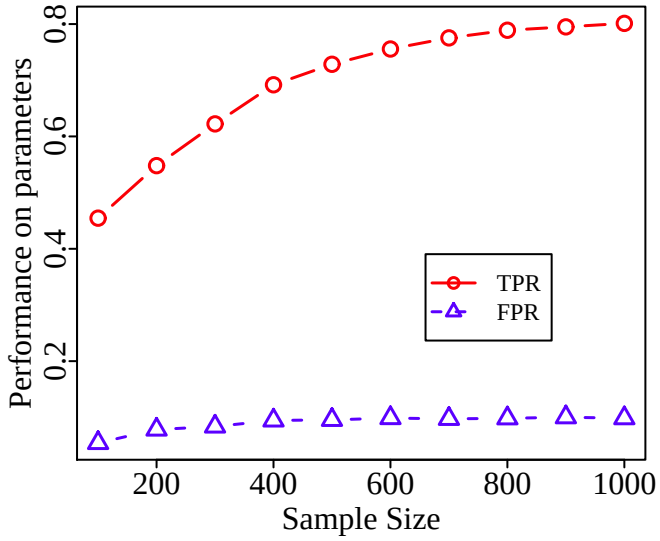


Figure 5.9 Performance of SLasso with feature selection on Graph 3, $p = 5$. The TPR and FPR are calculated on the unit of parameters. That is, we count the correctly and incorrectly discovered parameters and compare them to the true models.

we evaluate the methods on 100 separate runs and average the performance. In each run, we generate 2 datasets for the graph structure as in Figure (2.1(a))-(2.1(d)). One of the data set is used for training and the other for testing.

The true set of the model parameters, c_j^ω, $j = 1, \cdots, 5$, are shown in Appendix B.2. The features, X_j, are i.i.d uniform on [-1, 1]. Y is sampled according to the probability in equation (2.4).

We evaluate the following 5 models

$$1.Vanilla : \min_f \quad \mathcal{I}_\lambda(f) = \mathcal{L}(f) + \lambda \sum_{\omega,j} |c_j^\omega| \tag{5.2}$$

$$2.SLasso : \min_f \quad \mathcal{I}_\lambda(f) = \mathcal{L}(f) + \lambda \sum_{v \in \wp(\Omega)} p_v \sqrt{\sum_{\omega \in T_v} \|f^\omega\|^2_{\mathcal{H}_\omega}} \tag{5.3}$$

$$3.SLasso - Refit : \text{ Refit the logistic regression on the subset selected by SLasso} \tag{5.4}$$

$$4.Full : \text{ Fit the logistic regression without model selection} \tag{5.5}$$

$$5.Best : \text{ Use the true model parameters on test set} \tag{5.6}$$

We evaluate "BAC"(Balanced Accuracy, equals to (sensitivity + specificity)/2) and "Log-likelihood" on the test set and average the results over the 100 runs for each of the sample size. The results are plotted in Figure (5.10(a)), (5.11(a)) ,(5.12(a)), (5.13(a)), and Figure (5.10(b)), (5.11(b)), (5.12(b)), (5.13(b)).

In Figure (5.12(a)) and Figure (5.13(a)), "SLasso-Refit" achieves almost the same performance as "Best". This is because with increasing sample size, "SLasso" tends to select the true non-zero patterns, which makes the refitted models close to the true models. Since "SLasso" is itself a the model selection method, and thus provides biased estimators of the model parameters with finite sample size, the BAC performance is below "SLasso-Refit" and "Best". The "Vanilla" is below "SLasso" and "Full". One possible reason is using a single tuning parameter in the "Vanilla" model. In addition, the sparsity on the level of graph structure is more important than the sparsity on the level of model parameters. This might also be the reason that "Vanilla" is not as good as other methods.

In Figure (5.12(b)) and Figure (5.13(b)), similar results are observed. "SLasso-Refit" achieves almost the same performance as "Best" with increasing sample size. In terms of log-likelihood

on the test set, “Vanilla” and “SLasso” achieve very close performance, since they are both model selection methods.

For Graph 1 and Graph 2 which have simpler structure compared to Graph 3 and Graph 4, the performance of “SLasso” is not as good as “Vanilla” or “Full” (In Figure (5.10(a)), 5.11(a), and Figure 5.10(b), 5.11(b)). The reason may be the benifits of estimating the correct graph structure for a simple graphical model is overwhelmed by the bias brought by SLasso. But still, after re-fitting on the non-zero patterns estimated by SLasso , “SLasso-Refit” achieves almost the same performance as “Best” with increasing sample size.

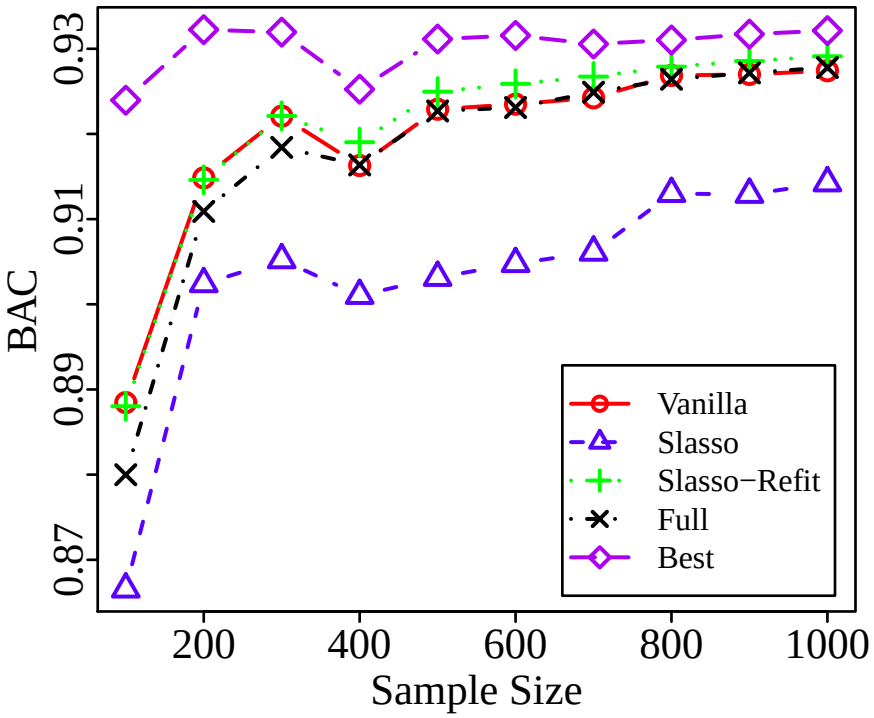
5.1.5 Consistency in Estimating the Cover of Non-zero Patterns

The grouping structure of the penalty in SLasso objective function in Equation (3.5) produces the consistency property of estimating the cover of non-zero patterns (see Section 4.1). If the true graphical model has the hierarchical structure, the SLasso method will recover the same set of non-zero patterns when sample size goes to infinity. However, if the true graphical model does not have the hierarchical structure, e.g. a fourth order interactions exists without some/all the lower order interactions, the SLasso method can recover the same graph structure eventually, but with all lower order interactions included.

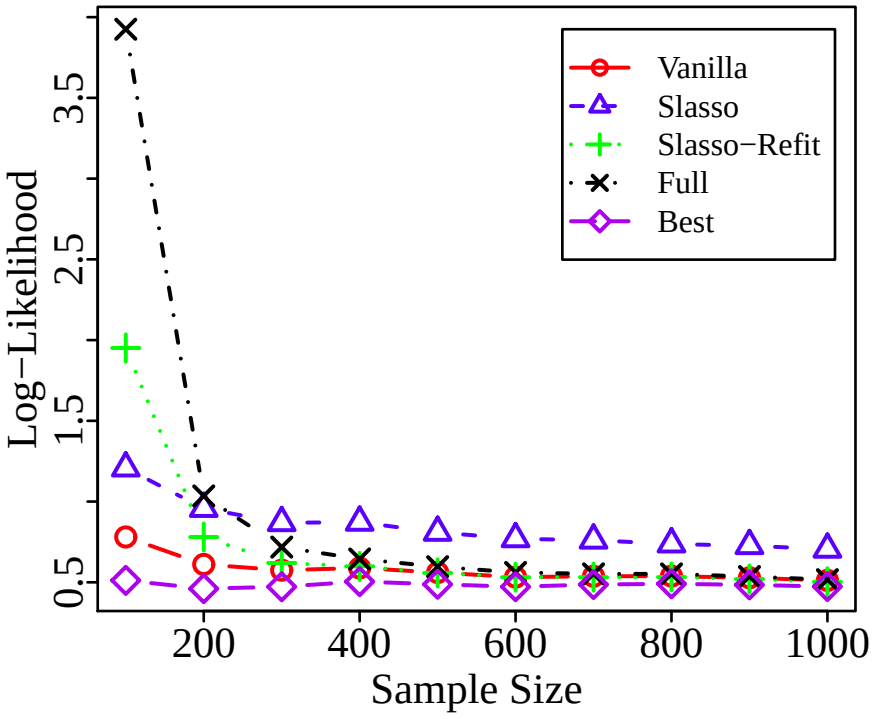
In this section, we will show the consistency of the SLasso method when the graph does not have the hierarchical structure. We carry out the experiments on Graph 3 when $p = 5$. Only the main effects and $f^{\omega}, \omega = \{3, 4\}, \{1, 2, 3\}, \{5, 6, 7, 8\}$ are non-zero. The true model parameters we use are listed below.

Graph 3, p=5

{1}	{2}	{3}	{4}	{5}	{6}	{7}	{8}
-0.5000	0.7500	-0.5000	0.5000	1.0000	1.0000	-0.7500	0.5000
0.7500	0.5000	-0.5000	-0.7500	1.0000	-0.7500	-1.0000	-0.5000
-0.7500	0.5000	-0.5000	0.7500	-0.7500	-1.0000	-1.0000	-1.0000
-0.5000	1.0000	-1.0000	-0.7500	0.5000	-0.5000	1.0000	0.5000
-0.7500	0.5000	0.5000	-0.7500	-0.5000	-1.0000	0.5000	0.5000
-0.4000	-0.4000	-0.8000	-0.8000	-0.4000	-0.4000	-0.4000	-0.4000

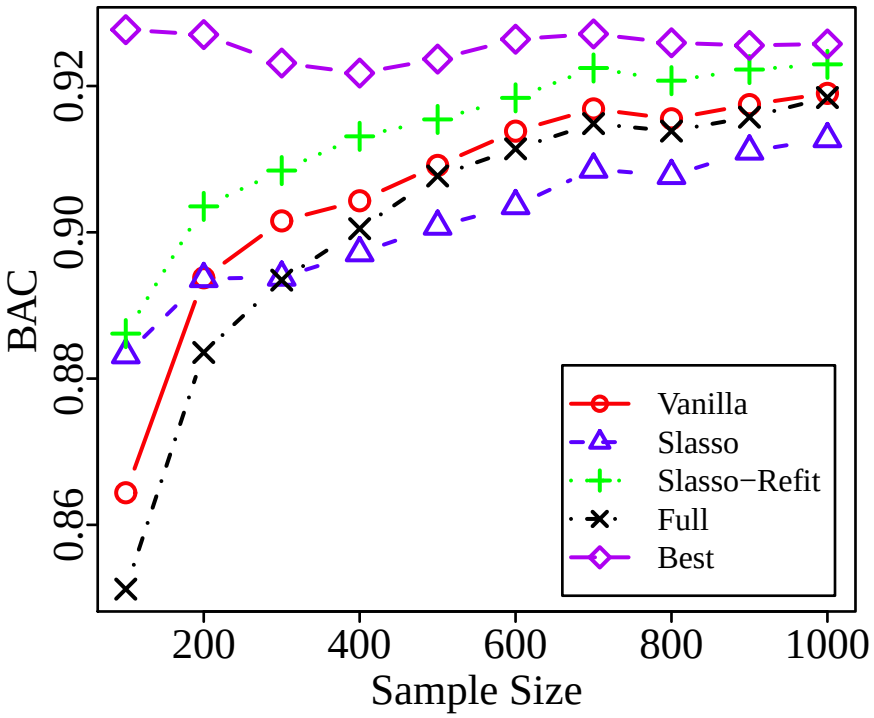


(a) Balanced Accuracy

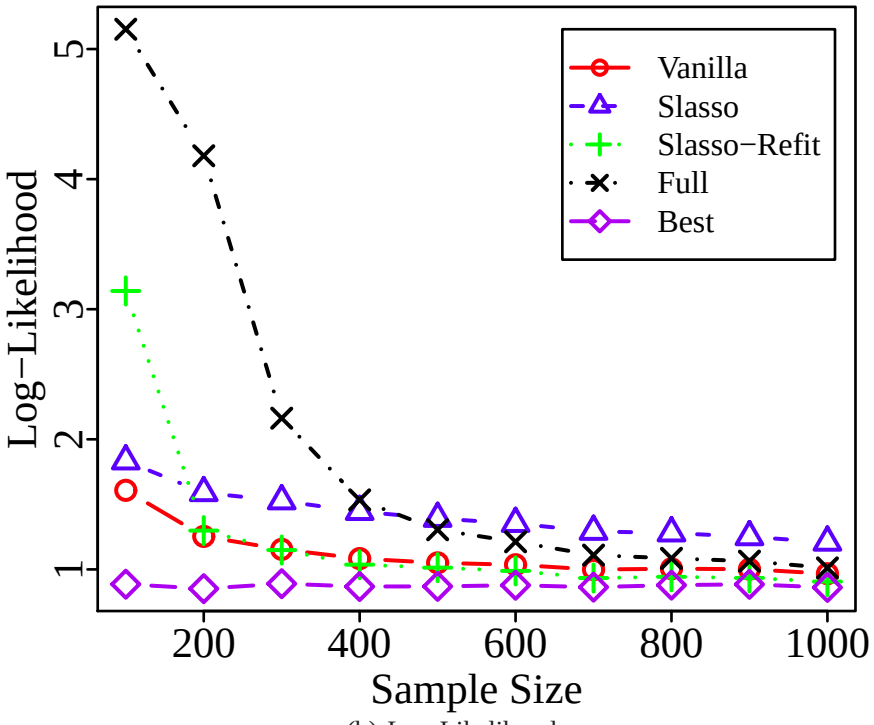


(b) Log-Likelihood

Figure 5.10 Comparison on Graph 1

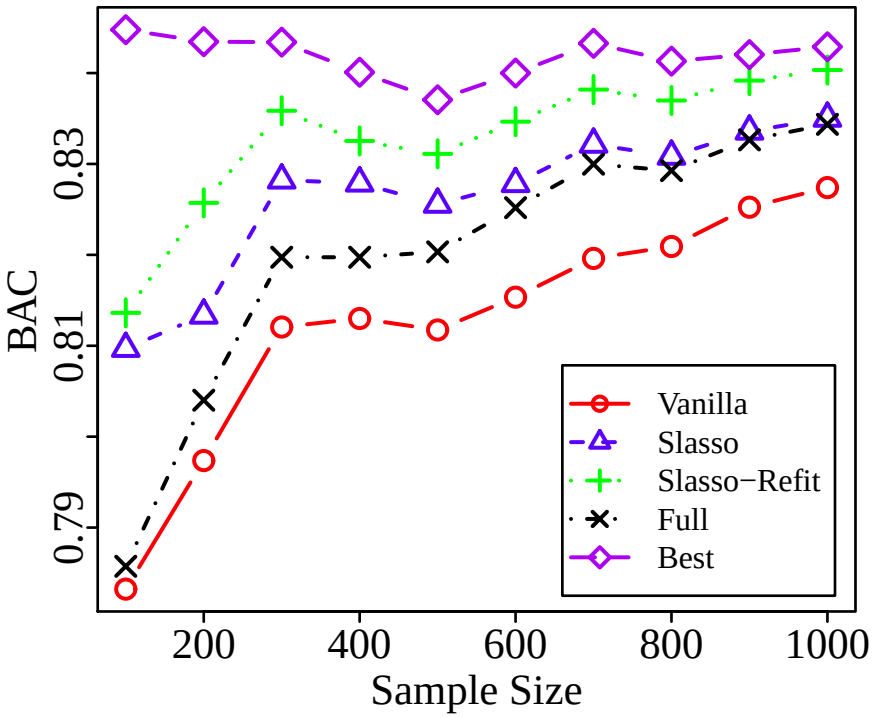


(a) Balanced Accuracy

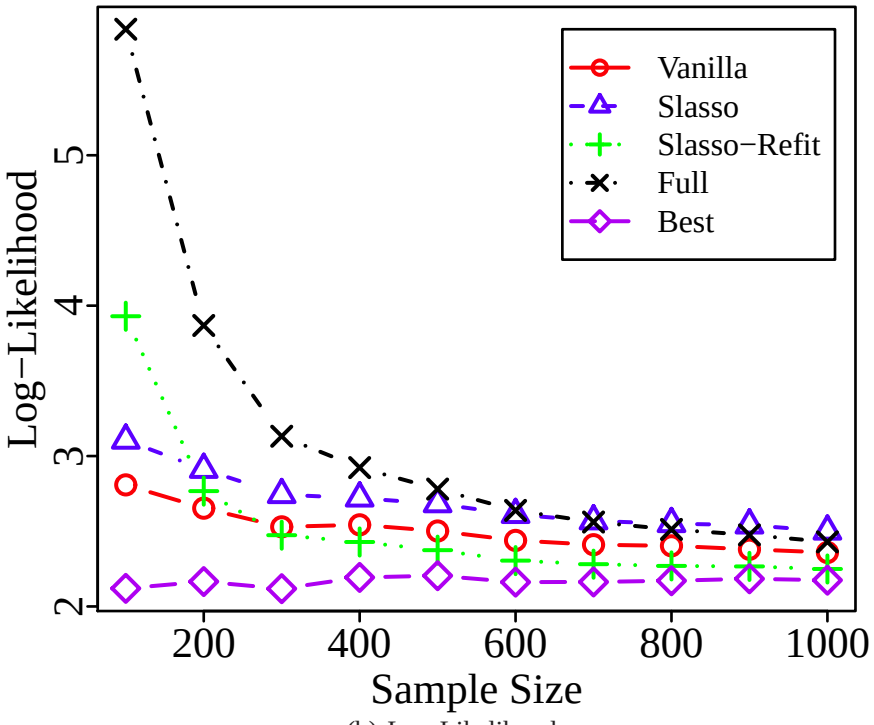


(b) Log-Likelihood

Figure 5.11 Comparison on Graph 2

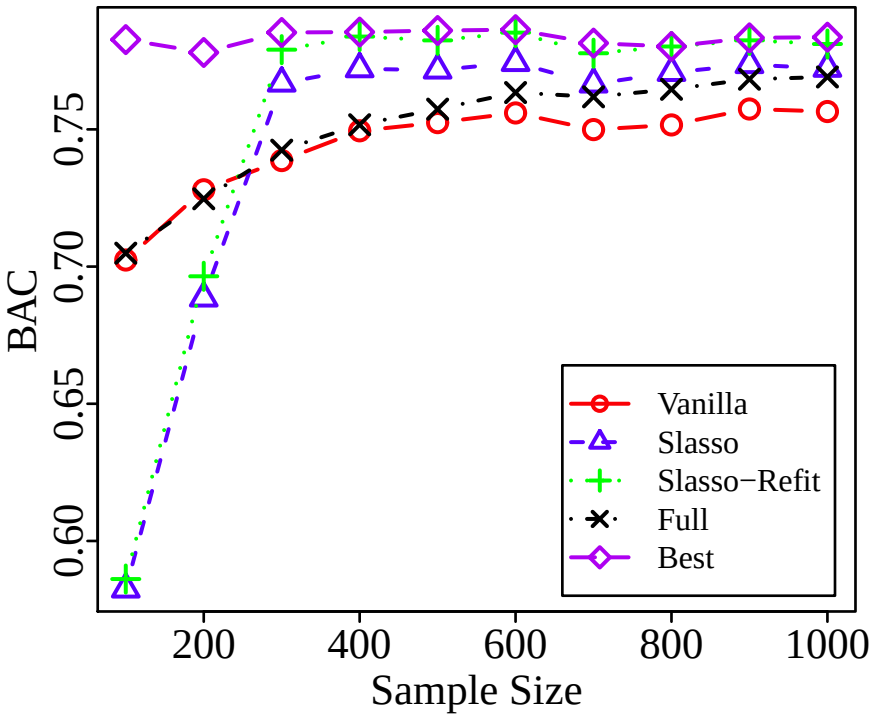


(a) Balanced Accuracy

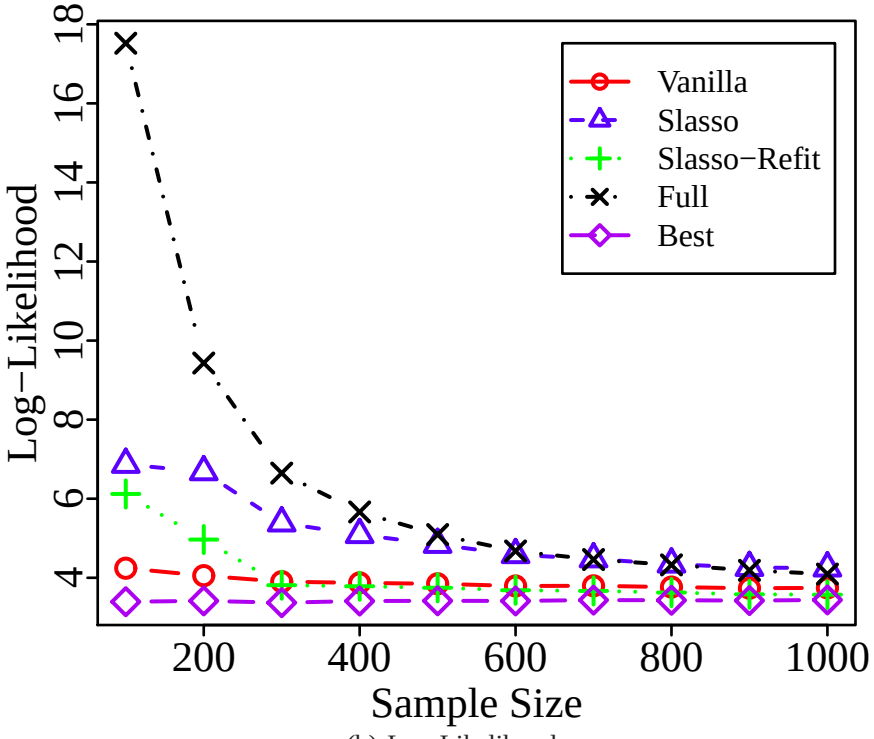


(b) Log-Likelihood

Figure 5.12 Comparison on Graph 3



(a) Balanced Accuracy



(b) Log-Likelihood

Figure 5.13 Comparison on Graph 4

```
 {3,4}     {1,2,3}   {5,6,7,8}
-0.6000    0.6000   -0.3000
 0.4500    0.4500   -0.4500
-0.3000   -0.6000    0.6000
 0.3000   -0.3000    0.6000
-0.6000   -0.6000   -0.4500
 2.0000    2.6000    3.6000
```

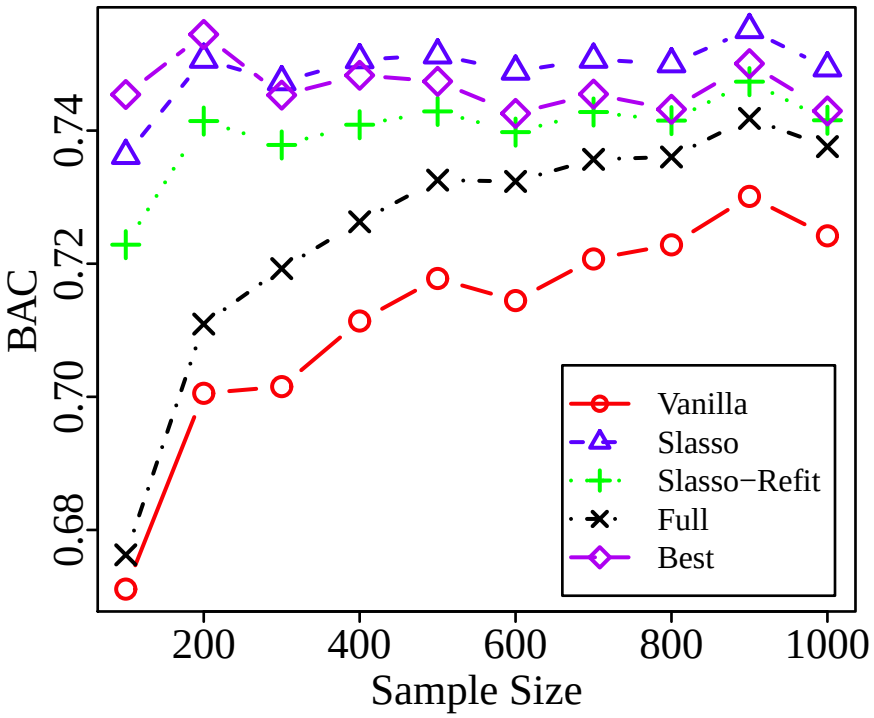
As we did in the previous section, we plot the performance of five methods in terms of balanced accuracy and log-likelihood in Figure (5.14). It is worthy of noting that "SLasso" is better than "Best" in terms of balanced accuracy. It may be because training the "SLasso" model alleviates the overfitting issue with the above chosen model parameters. But in terms of log-likelihood, "SLasso" is not as good as other methods, except for "Full" which tends to overfit in the setting. "SLasso-Refit" is getting very close to "Best" in both balanced accuracy and log-likelihood, when sample size becomes large. This shows that when the SLasso method recovers the true graph structure (in high probability when sample size is large), refitting the model achieves very similar results as those achieved in the oracle scenario.

Although SLasso discovers many lower order interactions belonging to the three cliques in the hierarchical structure, it performs very well in terms of structure learning. The TPR of SLasso recovering the cliques is 92.34%, with 1.54 false discovered cliques in one run on average, when the sample size is 1000.

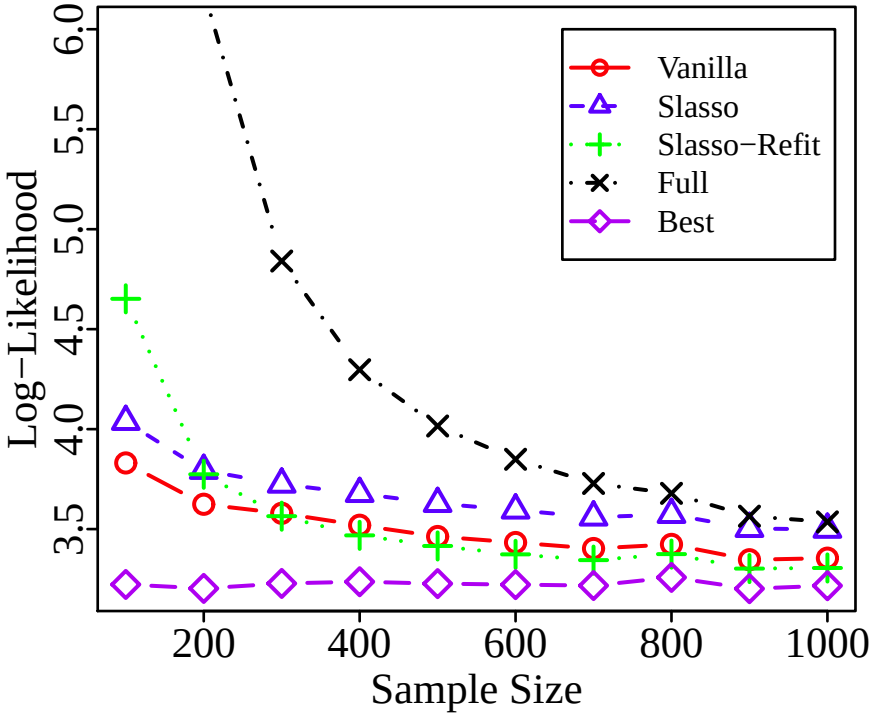
5.2 Case Study: Census Bureau County Data

We use the county data from U.S. Census Bureau[3] to validate our method. We remove the counties that have missing values and obtain 2668 entries in total. The outcomes of this study are summarized in Table 5.3. "Vote" Scammon et al. (2005) [74] is coded as 1 if the Republican candidate won in the 2004 presidential election. To dichotomize the remaining outcomes, the national mean is selected as a threshold. The data is standardized to mean 0 and variance 1. The following features are included: Housing unit change in percent from 2000-2006, percent of ethnic groups, percent foreign born, percent people over 65, percent people under 18, percent people with a high school education, percent people with a bachelors

[3] http://www.census.gov/statab/www/ccdb.html



(a) Balanced Accuracy



(b) Log-Likelihood

Figure 5.14 Comparison on Graph 3 where only the main effects and $f^{\omega}, \omega = \{3,4\}, \{1,2,3\}, \{5,6,7,8\}$ are non-zero. $p = 5$

Table 5.3 Selected response variables

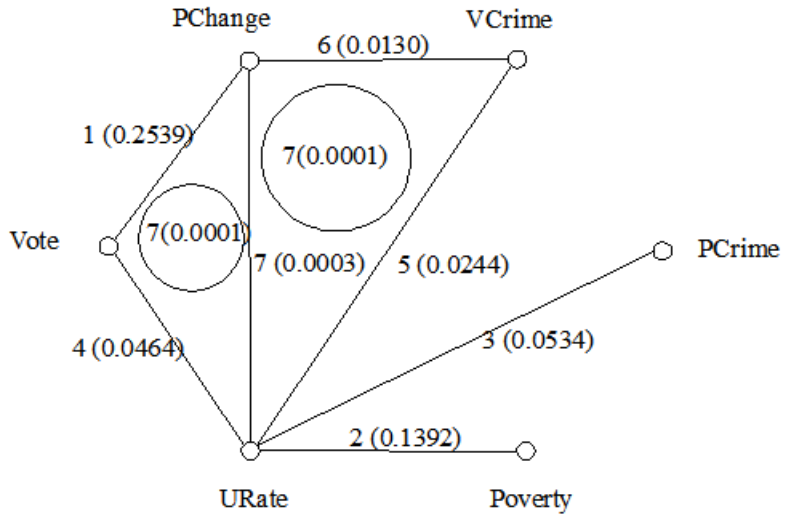
Response	Description	Positive%
Vote	2004 votes for Republican presidential candidate	81.11
Poverty	Poverty Rate	52.70
VCrime	Violent Crime Rate, eg. murder, robbery	23.09
PCrime	Property Crime Rate, eg. burglary	6.82
URate	Unemployment Rate	51.35
PChange	Population change in percent from 2000 to 2006	64.96

degree; birth rate, death rate, per capita government expenditure in dollars. By adjusting λ, we observe new interactions enter the model. The graph structure of $\lambda = 0.1559$ is shown in Figure 5.15(a). The results of BMN (the tuning parameter is 0.015) is in Figure 5.15(b). The unemployment rate plays an important role as a hub as discovered by SLasso , but not by BMN.

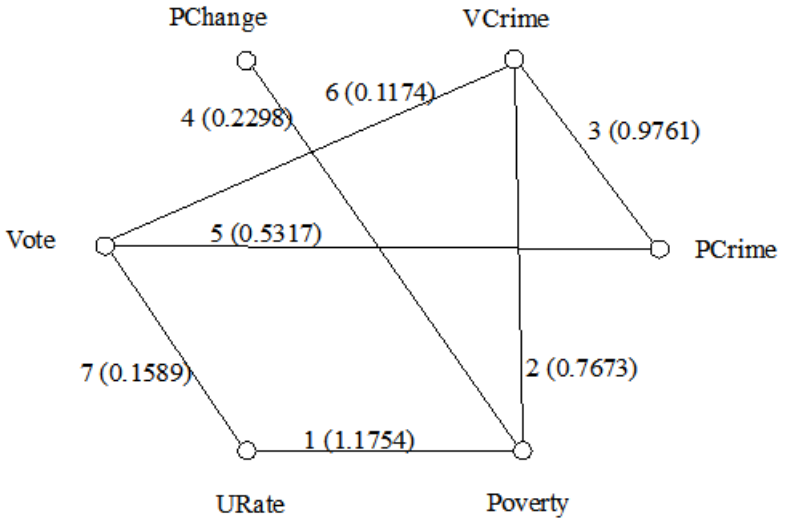
We analyze the link between "Vote" and "PChange". Though the marginal correlation between them (without X) is only 0.0389, which is the second lowest absolute pairwise correlation, the link is firstly recovered by SLasso . It has been suggested that there is indeed a connection[4]. This shows that after taking features into account, the dependence structure of response variables may change and hidden relations could be discovered. The main factors in this case are "percentage of housing unit change" (X_1) and "population percentage of people over 65" (X_2). The part of the fitted model shown below suggests that as housing units increase, the counties are more likely to have both positive results for "Vote" and "PChange". But this tendency will be counteracted by the increase of people over 65: the responses are less likely to take both positive values.

$$\hat{f}^{Vote} = 0.2913 \cdot X_1 + 0.3475 \cdot X_2 + \cdots$$
$$\hat{f}^{PChange} = 1.4726 \cdot X_1 - 0.3709 \cdot X_2 + \cdots$$
$$\hat{f}^{Vote,PChange} = 0.1358 \cdot X_1 - 0.0458 \cdot X_2 + \cdots$$

[4] http://www.ipsos-mori.com/researchpublications/researcharchive/2545/Analysis-Population-change-turnout-the-election.aspx



(a) SLasso-Complete



(b) BMN

Figure 5.15 Interactions of response variables in the Census Bureau data. The first number on the edge is the order at which the link is recovered. The number in bracket is the function norm on the clique and the absolute value of the elements in the concentration matrix, respectively. We note SLasso discovers at 7th step two third-order interactions which are displayed by two circles in (a).

Chapter 6

Concluding Remarks

The graphical models are very popular in modeling the relations in a set of discrete random variables Y. It is more interesting to estimate the distribution of Y conditioned on another set of predictive variables X. When the graph structure is given, the parameterization with potential functions are effective in estimating the model parameters. However, if we want to learn the graph structure and the functions of X that characterize the (higher-order) interactions among the Y's, this method could lead to over-parameterization.

Our first contribution is the reparameterization of the distribution of Y conditioned on X by multivariate discrete distributions. The conditional log odds ratios decompose the effect of X on Y to main effects and interactions of all orders. We prove the multivariate discrete distributions are equivalent to the graphical models parameterized by potential functions. The multivariate discrete model is easy to interpret the interactions among the nodes, since we prove the equivalence of the sparsity in the set of f^{ω}'s to the sparsity of the cliques in the graph. And the sparse estimation of the set of conditional log odds ratios leads to the conditional independence in the graphical model.

We propose the SLasso method to learn the graph structure that is specified by the conditional log odds ratios defined on the predictive variables X. The advantage is the combination of the graph structure learning and the flexible choice of the functional spaces on X. The method solves a maximum likelihood problem penalized by a structure penalty. The penalty is designed on groups of the conditional log odds ratios, following the hierarchical structure assumption. An efficient gradient descent algorithm is given to estimate the complete model. The global convergence of the algorithm is guaranteed. And a greedy approach is applied when the graph is large. The BGACV tuning method is derived to select the tuning parameter. It achieves satisfactory numerical results in simulation studies.

In addition, we allow the log odds ratios of the joint distribution conditioned on the predictive variables be functions in any separable Reproducing Kernel Hilbert Spaces. In this way, we extend the linear

models to teh non-parametric models. The asymptotic analysis shows that the SLasso method with parametric models and non-parametric models is consistent in terms of graph structure estimation, because of the special design of the structure penalty. That is, if the true model satisfies the hierarchical structure assumption, the SLasso method is consistent in estimating the set of non-zero conditional log odds ratios. If not, the SLasso method will recover a superset of the non-zero conditional log odds ratios in the true model. The superset will still give the same graph structure, so the estimation will still preserve the conditional independence structure.

This model can be applied to a variety of areas. One application is to discover the relations of multiple symptoms or clinical responses and how they are affected by the environmental and genetic covariates of the subjects. Smoking could be significant for many diseases and their interactions, but other covariates, such as taking Vitamin might be only related to a subset of the symptoms.

We can apply this method on Facebook data if available in the future. Say we have K ads (or ads categories), which will be clicked (1) or not (0) by the users (X). The observations of the clicks will be of multivariate Bernoulli distribution conditioned on X. The intuition is that those ads are related to each other. But these relations will depend on the features of the users, because different users have different browsing patterns. We are also interested in the prediction of a user's browsing behavior. In addition, we want to make better use of the social network between the users, e.g., friends, families, or subscriptions. This introduces another network, which is on X, and will be treated as known in the future work. The relationships between the users will provide additional information in the prediction, because friends are likely to share similar interests.

DISCARD THIS PAGE

Appendix A: Technical Proofs

A.1 Proof of Theorem 2.3

Proof. Given UGM (2.1), the corresponding parameterization in MVB model is shown in Equation (2.12) of Lemma 2.2.

The expression of $\exp(S^\omega(x))$ in Equation (2.11) follows from the definition of the conditional log odds ratios in (2.6).

Let y_C^ω be a realization of y_C such that $y_C^\omega = \{y_i^\omega \mid i \in C\}$ where $y_i^\omega = 1$ if $i \in \omega$ and $y_i^\omega = 0$ otherwise. Let the odd-even partition of the power set of ω defined as in Lemma 2.2. The conditional log odds ratios in MVB model are:

$$f^\omega(x) = \log \frac{\prod_{\kappa \in \wp_e(\omega)} \prod_{C \in \mathcal{C}} \Phi_C(y_C^\kappa; x)}{\prod_{\kappa \in \wp_o(\omega)} \prod_{C \in \mathcal{C}} \Phi_C(y_C^\kappa; x)} \quad \text{and} \quad b(f) = \log \frac{Z(x)}{\prod_{C \in \mathcal{C}} \Phi_C(0; x)} \tag{A.1}$$

Conversely, given the MVB model of (2.4), the cliques can be determined by the nonzero f^ω: clique C exists if $C = \omega$ and $f^\omega \neq 0$. Then the maximal cliques can be inferred from the graph structure. And suppose they are $C_1, \ldots, C_m$. Let $\omega_i = C_i$, for $i = 1, \ldots, m$, and $\kappa_1 = \emptyset$, $\kappa_i = C_i \cap (C_{i-1} \cup \cdots \cup C_1), i = 2, \ldots, m$. Then the parameterization is:

$$\Phi_{C_i}(y_{C_i}; x) = \exp\left(S^{\omega_i}(y; x) - S^{\kappa_i}(y; x)\right) \quad \text{and} \quad Z(x) = \exp(b(f)) \tag{A.2}$$

where $S^\omega(y; x) = \sum_{\kappa \subseteq \omega} y^\kappa f^\kappa(x)$. Thus, UGM (2.1) with bivariate nodes is equivalent to MVB (2.4).

In the latter part of the theorem, $1 \Rightarrow 2$ and $3 \Rightarrow 1$ follow naturally from the Markov property of graphical models. To show $2 \Rightarrow 3$, let y_C^ω be a realization of y_C such that $y_C^\omega = (y_i^\omega)_{i \in C}$ where $y_i^\omega = 1$ if $i \in \omega$ and $y_i^\omega = 0$ otherwise. Notice that whenever $\kappa \cap C = \kappa' \cap C$, we have $y_C^\kappa = y_C^{\kappa'}$. For any possible $v = \kappa \cap C$, $\kappa' \in \{\kappa | \kappa = v \cup u, \text{ s.t. } u \subseteq \omega - v\}$ will satisfy the condition: $\kappa' \cap C = v$. There are $2^{|\omega - v|}$ such κ' in total due to the choice of u. Also, they appear in the nominator and denominator of Equation (2.12) equally. So, for any $C \in \mathcal{C}$,

$$\prod_{\kappa \in \Psi^\omega_{even}} \Phi_C(y_C^\kappa; x) = \prod_{\kappa \in \Psi^\omega_{odd}} \Phi_C(y_C^\kappa; x) \tag{A.3}$$

It follows that $f^\omega = 0$ by (2.12). □

A.2 Proof of Theorem 3.2

Proof. The existence of the minimizer can be shown following the proof of Theorem 1 in Lin and Zhang (2006) [53]. Let the projection of f onto $span\{K(x_i,\cdot), i = 1,\ldots,n\} \subset \mathcal{H}_K$ be g, and $h = f - g$. Note, $\|f\|^2_{\mathcal{H}_K} = \|g\|^2_{\mathcal{H}_K} + \|h\|^2_{\mathcal{H}_K}$, and $f(x_i) = \langle K(x_i,\cdot), f\rangle = \langle K(x_i,\cdot), g\rangle$, then the objective function in Equation (3.5) is

$$\frac{1}{n}\sum_{i=1}^{n} L(\langle K(x_i,\cdot), g\rangle) + \lambda \sum_{v\in\wp(\Omega)} p_v \sqrt{\sum_{\omega\in T_v} \|g^\omega\|^2_{\mathcal{H}_\omega} + \sum_{\omega\in T_v} \|h^\omega\|^2_{\mathcal{H}_\omega}} \tag{A.4}$$

Therefore, we know the minimizer is in $span\{K(x_i,\cdot), i = 1,\ldots,n\}$.

□

A.3 Proof of Theorem 3.3

Proof. We give the proof for the linear case. The convexity of I_λ is easy to check, since L and $J(f^{T_v})$ are all convex in c. Suppose there is some $\omega_2 \supset \omega_1$ s.t. $\hat{c}^{\omega_2} \neq 0$ and $\hat{c}^{\omega_1} = 0$, by the groups constructed through Figure 3.1, $\|\hat{c}^{T_v}\| = \|(\hat{c}^\omega)_{v\subseteq\omega}\| \neq 0$ for all $v \subseteq \omega_1$. So the partial derivative of the objective (3.12) with respect to c^{ω_1} at $\hat{c}^{\omega_1}$ is

$$\left.\frac{\partial L}{\partial c^{\omega_1}}\right|_{c^{\omega_1}=\hat{c}^{\omega_1}} + \lambda \sum_{v\subseteq\omega_1} p_v \frac{\hat{c}^{\omega_1}}{\|\hat{c}^{T_v}\|} = 0 \tag{A.5}$$

Thus, the probability of $\{\hat{c}^{\omega_2} \neq 0\}$ equals to the probability of $\{\frac{\partial L}{\partial c^{\omega_1}}\big|_{c^{\omega_1}=\hat{c}^{\omega_1}} = 0\}$, which is 0. □

A.4 Proof of Lemma 4.3

Proof. Let

$$U_n(\delta) = \frac{1}{n}\sum_{i=1}^{n} \left[L_{Z_i}\left(f^* + \lambda_n\delta\right) - L_{Z_i}(f^*)\right], \tag{A.6}$$

$$V_n(\delta) = \lambda_n\left[\mathcal{J}\left(f^* + \lambda_n\delta\right) - \mathcal{J}(f^*)\right] \tag{A.7}$$

Note $\hat{\delta}_n = \frac{1}{\lambda_n}(\hat{f}_n - f^*)$ is the minimizer of $\frac{1}{\lambda_n{}^2}U_n(\delta) + \frac{1}{\lambda_n{}^2}V_n(\delta)$. We will first show the convergence of $U_n(\delta)$ and $V_n(\delta)$.

For $V_n(\delta)$, consider the cases where $v \in \mathcal{A}$ first.

$$\begin{aligned}\frac{1}{\lambda_n}\left[J_v\left(f^{*T_v}+\lambda_n\delta^{T_v}\right)-J_v(f^{*T_v})\right] &= \frac{p_v}{\lambda_n}\left[\|f^{*T_v}+\lambda_n\delta^{T_v}\|-\|f^{*T_v}\|\right] \\ &= \frac{p_v}{\lambda_n}\frac{2\left(f^{*T_v}\right)^T\lambda_n\delta^{T_v}+\|\lambda_n\delta^{T_v}\|^2}{\|f^{*T_v}+\lambda_n\delta^{T_v}\|+\|f^{*T_v}\|} \\ &\to \frac{p_v\left(f^{*T_v}\right)^T\delta^{T_v}}{\|f^{*T_v}\|}, \quad \text{as } n\to\infty \end{aligned} \tag{A.8}$$

For $v \in \mathcal{A}^c$, we have

$$\frac{1}{\lambda_n}\left[J_v\left(f^{*T_v}+\lambda_n\delta^{T_v}\right)-J_v(f^{*T_v})\right] \to J_v(\delta^{T_v}) \tag{A.9}$$

Then, we get the convergence result of $\frac{1}{\lambda_n{}^2}V_n(\delta)$:

$$\frac{1}{\lambda_n^2}V_n(\delta) \to (\gamma^{\mathcal{A}})^T\delta^{\mathcal{A}}+\mathcal{J}(\delta^{\mathcal{A}^c}), \quad \text{as } \lambda_n\to\infty \tag{A.10}$$

For $U_n(\delta)$, we have

$$\begin{aligned}\frac{1}{\lambda_n^2}U_n(\delta) &= \frac{1}{\lambda_n^2}\cdot\frac{1}{n}\sum_{i=1}^n\left[L_{Z_i}\left(f^*+\lambda_n\delta\right)-L_{Z_i}(f^*)\right] \\ &= \frac{1}{\lambda_n\sqrt{n}}\delta^T\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n\nabla L_{Z_i}(f^*)\right]+\delta^T\left[\frac{1}{2n}\sum_{i=1}^n\nabla^2L_{Z_i}(f^*)\right]\delta+o_p\left(\frac{\|\delta\|^2}{n}\right)\end{aligned} \tag{A.11}$$

Let $M(f^*)=\mathbb{E}\left[\nabla L_Z(f^*)\nabla L_Z(f^*)^T\right]$. By central limit theorem, $\frac{1}{\sqrt{n}}\sum_{i=1}^n\nabla L_{Z_i}(f^*)\xrightarrow{d}W=N(0,M(f^*))$, since $\mathbb{E}[\nabla L_{Z_i}(f^*)]=0$ and $\mathrm{var}(\nabla L_{Z_i}(f^*))=\mathbb{E}\left[\nabla L(Y;f^*(X))\nabla L(Y;f^*(X))^T\right]=M(f^*)$. So the first term converges to 0 in probability as $\lambda_n\sqrt{n}\to\infty$ when $n\to\infty$.

Note $\mathbb{E}\left[\nabla^2L_{Z_i}(f^*)\right]=\nabla^2\mathbb{E}\left[L_{Z_i}(f^*)\right]=H(f^*)$, we have

$$\frac{1}{n}\sum_{i=1}^n\delta^T\nabla^2L_{Z_i}(f^*)\delta\xrightarrow{a.s.}\delta^TH(f^*)\delta \tag{A.12}$$

Then, we get the convergence result of $\frac{1}{\lambda_n{}^2}V_n(\delta)$:

$$\frac{1}{\lambda_n^2}U_n(\delta)\xrightarrow{p}\frac{1}{2}\delta^TH(f^*)\delta \tag{A.13}$$

Therefore,

$$U_n(\delta)+V_n(\delta)\xrightarrow{p}W(\delta)=\frac{1}{2}\delta^TH(f)\delta+(\gamma^{\mathcal{A}})^T\delta^{\mathcal{A}}+\mathcal{J}(\delta^{\mathcal{A}^c}) \tag{A.14}$$

Finally, since $U_n(\delta) + V_n(\delta)$ is convex and $W(\delta)$ has a unique minimum, it follows Geyer (1996) [27] and that

$$\frac{1}{\lambda_n}\left(\hat{f}_n - f^*\right) = \arg\min_{\delta}(U_n(\delta) + V_n(\delta)) \tag{A.15}$$
$$\xrightarrow{d} \hat{\delta} = \arg\min_{\delta} \frac{1}{2}\delta^T H(f^*)\delta + (\gamma^{\mathcal{A}})^T \delta^{\mathcal{A}} + \mathcal{J}(\delta^{\mathcal{A}^c})$$

More general version of the convergence in minimization for random functions on finite dimensional Hilbert space can be found in Knight (1999) [41] and Rockafellar and Wets (1998) [73].

□

A.5 Proof of Lemma 4.5

Proof. We know $\hat{f}_n^{\mathcal{A}}$ is unique since the objective function in Equation (4.18) is strongly convex. Using the techniques when proving Lemma 4.3, the $\frac{1}{\lambda_n}$-consistent result of $\hat{f}_n^{\mathcal{A}}$ for $f^{*\mathcal{A}}$ which is similar to the one implied in Equation (4.8) means that $\hat{f}_n^{\mathcal{A}} \xrightarrow{p} f^{*\mathcal{A}}$. This implies that $\mathbb{P}\left(\mathcal{P} \subseteq \hat{\mathcal{P}}_{n\mathcal{A}}\right) \to 1$. Since $\mathcal{A} = \text{cover}(\mathcal{P})$, we have $\mathbb{P}\left(\mathcal{A} \subseteq \hat{\mathcal{A}}_{n\mathcal{A}}\right) \to 1$. Since $\hat{\mathcal{P}}_{n\mathcal{A}} = \hat{\mathcal{A}}_{n\mathcal{A}}$ almost surely as shown in Lemma 4.2, we have $\mathbb{P}\left(\mathcal{A} \subseteq \hat{\mathcal{P}}_{n\mathcal{A}}\right) \to 1$. $\hat{\mathcal{P}}_{n\mathcal{A}} \subseteq \mathcal{A}$ is always true, so $\mathbb{P}\left(\hat{\mathcal{P}}_{n\mathcal{A}} = \mathcal{A}\right) \to 1$. The conclusion of Lemma 4.5 follows from Lemma 4.2. □

A.6 Proof of Lemma 4.15

Proof. First, Volle and Hiriart-Urruty (2011) [89] showed that a weakly lower-semicontinuous function defined on a reflexive Banach space has a unique minimizer if and only if it is essentially strictly convex. Note, a RKHS is a reflexive Banach space. The properties of the objective function provides sufficient conditions for W_n and W attaining a unique minimizer.

Note, a sequence $\{F_n : \mathcal{H}_K \to \mathbb{R}\}$ is said to epi-converge to $F : \mathcal{H}_K \to \mathbb{R}$ at $f \in \mathcal{H}_K$ if for any $f_n \to f$, $\liminf F_n(f_n) \geq F(f)$ and $\exists f_n \to f$ such that $\limsup F_n(f_n) \leq F(f)$ (Dong and Wets (2000) [19]). Vogel and Lachout (2003) [88] showed that the point-wise convergence in probability for all $\delta \in \mathcal{H}_K$ implies that F_n epi-converges to F in probability. We introduce the notion of epi-convergence to utilize the general convergence results. More about the epi-convergence in probability can be found in Geyer (1994) [26], Hess (1996) [32], and Lachout (2006) [47]. It is worth noting that the continuity of W_n, W and point-wise convergence of W_n to W ensure W_n epi-converges to W in probability.

Since W_n epi-converge to W in probability, we can find another set of random elements W_n' and W' which are identically distributed as W_n and W, and W_n epi-converges to W almost surely (Van Der Vaart and Wellner (1996) [87]).

Let $\hat{\delta}_n' = \arg\min_{\delta \in \mathcal{H}_K} W_n'$ and $\hat{\delta}' = \arg\min_{\delta \in \mathcal{H}_K} W'$. Since $W_n(\delta) =_d W_n'(\delta)$ and $W(\delta) =_d W(\delta)$, it is easy to see that $\hat{\delta}_n'$ and $\hat{\delta}'$ have the same distribution as $\hat{\delta}_n$ and $\hat{\delta}$, which are the minimizers of W_n and W, respectively: $\hat{\delta}_n' =_d \hat{\delta}_n$ and $\hat{\delta}' =_d \hat{\delta}$.

It follows Theorem 2.11 and Corollary 2.13 in Attouch (1984) [5] that for $F_n, n = 1, 2, \cdots$ and F which are functionals defined on a separable Hilbert space $\mathcal{H}$, if F_n epi-converges to F and F has a unique minimizer, then $\arg\min F_n \to \arg\min F$. Since $W_n'(\cdot, e)$ epi-converges to $W'(\cdot, e)$ for almost all $e \in \mathcal{E}$, denote $\hat{\delta}_{n,e}' = \arg\min W_n'(\delta, e)$ and $\hat{\delta}_e = \arg\min W'(\delta, e)$, then $\hat{\delta}_{n,e}' \to \hat{\delta}_e'$ for almost all $e \in \mathcal{E}$, which implies that $\hat{\delta}_n' \to \hat{\delta}'$ almost surely. Therefore, $\arg\min_{\delta \in \mathcal{H}_K} W_n(\delta) \xrightarrow{d} \arg\min_{\delta \in \mathcal{H}_K} W(\delta)$.

□

A.7 Proof of Lemma 4.16

Proof. Similarly, let

$$U_n(\delta) = \frac{1}{n} \sum_{i=1}^{n} \left[L_{Z_i} \left(f^* + \lambda_n \delta \right) - L_{Z_i}(f^*) \right], \tag{A.16}$$

$$V_n(\delta) = \lambda_n \left[J \left(f^* + \lambda_n \delta \right) - \mathcal{J}(f^*) \right] \tag{A.17}$$

$$W_n(\delta) = \frac{1}{{\lambda_n}^2} [U_n(\delta) + V_n(\delta)] \tag{A.18}$$

From the Taylor's theorem in Banach space,

$$\begin{aligned} \frac{1}{\lambda_n^2} U_n(\delta) &= \frac{1}{\lambda_n^2} \cdot \frac{1}{n} \sum_{i=1}^{n} \left[L_{Z_i} \left(f^* + \lambda_n \delta \right) - L_{Z_i}(f^*) \right] \\ &= \frac{1}{\lambda_n} \left[\frac{1}{n} \sum_{i=1}^{n} \nabla L_{Z_i}(f^*)(\delta) \right] + \left[\frac{1}{2n} \sum_{i=1}^{n} \nabla^2 L_{Z_i}(f^*)(\delta, \delta) \right] + o_p \left(\frac{\|\delta\|_{\mathcal{H}_K}^2}{n} \right) \end{aligned} \tag{A.19}$$

Because $L_\cdot(f)$ is a bounded continuous operator of $Z \in \mathbb{Z}$, and f^* is optimal for the risk operator, we have for any $\delta \in \mathcal{H}_K$

$$\begin{aligned} \mathbb{E}_P[\nabla L_{Z_i}(f^*)(\delta)] &= E_P \circ \nabla L_{(\cdot)}(f^*)(\delta) \\ &= \nabla (E_P \circ L_{(\cdot)})(f^*)(\delta) \\ &= 0 \end{aligned} \tag{A.20}$$

So the first term $\frac{1}{\lambda_n}\left[\frac{1}{n}\sum_{i=1}^{n}\nabla L_{Z_i}(f^*)(\delta)\right] \xrightarrow{a.s.} 0$ by the law of large numbers in Banach space (Ledoux and Talagrand (1991) [49]).

For the second term,

$$\begin{aligned}\mathbb{E}_P\nabla^2 L_{Z_i}(f^*)(\delta,\delta) &= \nabla^2(E_P\circ L_{(\cdot)})(f^*)(\delta,\delta) \\ &= H(f^*)(\delta,\delta)\end{aligned} \tag{A.21}$$

So the second term $\left[\frac{1}{2n}\sum_{i=1}^{n}\nabla^2 L_{Z_i}(f^*)(\delta,\delta)\right] \xrightarrow{a.s.} \frac{1}{2}H(f^*)(\delta,\delta)$. Therefore,

$$\frac{1}{{\lambda_n}^2}U_n(\delta) \xrightarrow{p} \frac{1}{2}H(f^*)(\delta,\delta) \tag{A.22}$$

For $V_n(\delta)$, consider the cases where $v \in \mathcal{A}$ first.

$$\begin{aligned}\frac{1}{\lambda_n}\left[J_v\left(f^{*T_v}+\lambda_n\delta^{T_v}\right)-J_v(f^{*T_v})\right] &= \frac{p_v}{\lambda_n}\frac{2\left\langle f^{*T_v},\lambda_n\delta^{T_v}\right\rangle_{\mathcal{H}_K}+\|\lambda_n\delta^{T_v}\|^2_{\mathcal{H}_K}}{\|f^{*T_v}+\lambda_n\delta^{T_v}\|_{\mathcal{H}_K}+\|f^{*T_v}\|_{\mathcal{H}_K}} \\ &\to p_v\frac{\left\langle f^{*T_v},\delta^{T_v}\right\rangle_{\mathcal{H}_K}}{\|f^{*T_v}\|_{\mathcal{H}_K}}, \quad \text{as } n\to\infty\end{aligned} \tag{A.23}$$

For $v \in \mathcal{A}^{\mathsf{c}}$, we have

$$\frac{1}{\lambda_n}\left[J_v\left(f^{*T_v}+\lambda_n\delta^{T_v}\right)-J_v(f^{*T_v})\right] \to J_v(\delta^{T_v}) \tag{A.24}$$

So,

$$\frac{1}{\lambda_n^2}V_n(\delta) \to \left\langle\gamma^{\mathcal{A}},\delta^{\mathcal{A}}\right\rangle_{\mathcal{H}_K}+\mathcal{J}(\delta^{\mathcal{A}^{\mathsf{c}}}), \quad \text{as } n\to\infty \tag{A.25}$$

Therefore,

$$\begin{aligned}W_n(\delta) &= \frac{1}{{\lambda_n}^2}[U_n(\delta)+V_n(\delta)] \\ &\xrightarrow{p} \frac{1}{2}H(f^*)(\delta,\delta)+\left\langle\gamma^{\mathcal{A}},\delta^{\mathcal{A}}\right\rangle_{\mathcal{H}_K}+\mathcal{J}_{\mathcal{A}^{\mathsf{c}}}(\delta^{\mathcal{A}^{\mathsf{c}}})\end{aligned}$$

The conclusion of the lemma follows Lemma 4.15. □

A.8 B-spline

Given m knots, $t_0 \leq t_1 \leq \cdots \leq t_{m-1}$, the B-spline basis functions of degree d are defined recursively De Boor (1978) [17]:

$$b_{k,0} = \begin{cases} 1; & \text{if } t_k \leq t < t_{k+1} \\ 0; & \text{otherwise} \end{cases}, \text{ for } k = 0, \cdots, m-2$$

$$b_{k,l} = \frac{t - t_k}{t_{k+l} - t_k} b_{k,l-1}(t) + \frac{t_{k+l+1} - t}{t_{k+l+1} - t_{k+1}} b_{k+1,l-1}(t), \text{ for } k = 0, \cdots, m-d-2; l = 0, \cdots, d$$

Let $B_k(\cdot) = b_{k,d}(\cdot)$, then $\{B_k, k = 0, \cdots, m-d-2\}$ are $m-d-1$ basis functions, which span the functional space $\mathcal{B}$. The B-spline curve in $\mathcal{B}$ is:

$$g(t) = \sum_{k=0}^{m-d-2} c_k B_k(t) \tag{A.26}$$

where c_k's are the control points to be estimated. In our simulation studies, c_k's are assumed to be one dimensional scalers for simplicity.

Suppose $x \in \mathbb{R}^p$, we let each $f^\omega(x)$ be in $\mathcal{B}_0 \oplus \mathcal{B}_1 \oplus \cdots \oplus \mathcal{B}_p$. Here, $\mathcal{B}_0$ is a space of constant functions and $\mathcal{B}_j, j = 1 \cdots, p$ is a B-spline functional space on domain of x_j. Therefore,

$$f^\omega(x) = c_0^\omega + \sum_{j=1}^{p} g_j^\omega(x_j) \tag{A.27}$$

where $g_j^\omega \in \mathcal{B}_j$ are defined similarly as in (A.26): $g_j^\omega(x_j) = \sum_{k=1}^{D} c_{jk}^\omega B_k(x_j)$, and $D = m - d - 1$ is the number of basis functions.

Appendix B: True Model Parameters in the Experiment

Here, we list the true model paramters we used in the experiments, one column for one conditional log odds ratio. Without special notice, we use these parameters to generate the data in the experiments. For $p = 5$, we list the 6 paramters in one column in the order: $c_1^\omega, \cdots, c_5^\omega, c_0^\omega$, where c_0^ω is the intercept.

B.1 $p = 0$

Graph 1, p = 0

{1}	{2}	{3}	{4}	{1,2}	{1,3}	{2,3}	{3,4}
-2.0000	-2.0000	-2.0000	-2.0000	1.2000	1.2000	1.2000	1.2000

{1,2,3}
1.2000

Graph 2, p = 0

{1}	{2}	{3}	{4}	{5}	{6}	{1,2}	{1,3}
-0.3778	-0.2667	-0.0444	-0.3778	0.0667	0.0667	0.2889	-0.0444

{2,3}	{3,4}	{5,6}	{1,2,3}
-0.2667	-0.4889	-0.4889	1.0000

Graph 3, p = 0

{1}	{2}	{3}	{4}	{5}	{6}	{7}	{8}
-0.2000	-0.2000	-0.2000	0.2000	-0.2000	-0.2000	-0.2000	-0.2000

{1,2}	{1,3}	{2,3}	{3,4}	{5,6}	{5,7}	{5,8}	{6,7}
0.4000	0.4000	0.2000	0.5000	0.2000	0.3000	0.5000	0.6000

{6,8}	{7,8}	{1,2,3}	{5,6,7}	{5,6,8}	{5,7,8}	{6,7,8}	{5,6,7,8}
0.5000	0.5000	0.3000	-0.2000	-0.2000	-0.2000	-0.2000	1.0000

Graph 4, p = 0

{1}	{2}	{3}	{4}	{5}	{6}	{7}	{8}
-0.2000	-0.2000	-0.2000	-0.4000	-0.4000	-0.2000	-0.2000	-0.2000

{9}	{10}	{1,2}	{1,3}	{2,3}	{3,4}	{4,5}	{5,6}
-0.2000	-0.2000	0.6000	0.2000	0.4000	0.5000	0.6000	0.3000

{5,7}	{5,8}	{6,7}	{6,8}	{6,9}	{7,8}	{9,10}	{1,2,3}
0.6000	0.3000	0.4000	0.6000	0.4000	0.5000	0.6000	0.6000

{5,6,7}	{5,6,8}	{5,7,8}	{6,7,8}	{5,6,7,8}
-0.3000	-0.2000	-0.2000	-0.2000	1.4000

B.2 $p = 5$

Graph 1, p = 5

{1}	{2}	{3}	{4}	{1,2}	{1,3}	{2,3}	{3,4}
-2.0000	-3.0000	3.0000	3.0000	1.2000	-2.4000	1.8000	-2.4000
-4.0000	-3.0000	3.0000	3.0000	1.2000	-1.2000	-1.2000	-2.4000
4.0000	3.0000	-3.0000	-2.0000	-2.4000	2.4000	2.4000	2.4000
-3.0000	4.0000	-4.0000	3.0000	-1.2000	-2.4000	1.2000	1.8000
-2.0000	-3.0000	2.0000	2.0000	1.2000	-1.2000	-1.2000	2.4000
1.0000	1.0000	1.0000	1.0000	1.2000	1.2000	1.2000	1.2000

{1,2,3}
-2.4000
-1.2000
-2.4000
1.8000
-1.2000
1.2000

Graph 2, p = 5

{1}	{2}	{3}	{4}	{5}	{6}	{1,2}	{1,3}
-3.0000	-2.0000	3.0000	3.0000	-2.0000	3.0000	1.8000	-1.8000
-4.0000	3.0000	2.0000	3.0000	4.0000	-2.0000	1.8000	1.8000
4.0000	4.0000	-2.0000	-4.0000	4.0000	-3.0000	-1.2000	1.2000
-2.0000	4.0000	2.0000	2.0000	-2.0000	-3.0000	1.8000	1.2000
2.0000	-4.0000	-3.0000	4.0000	4.0000	3.0000	1.2000	2.4000
1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.2000	1.2000

{2,3}	{3,4}	{5,6}	{1,2,3}
2.4000	1.2000	1.2000	-1.2000
1.2000	-1.2000	-1.8000	-1.8000
-1.8000	-1.8000	1.8000	-2.4000
-1.2000	2.4000	1.8000	-1.2000
-1.8000	2.4000	2.4000	-1.2000
1.2000	1.2000	1.2000	1.2000

Graph 3, p = 5

{1}	{2}	{3}	{4}	{5}	{6}	{7}	{8}
-0.5000	-1.0000	-1.0000	-0.5000	-0.7500	0.7500	-1.0000	-0.7500
-0.5000	-1.0000	0.5000	0.7500	-1.0000	-0.7500	-1.0000	0.7500
-0.5000	1.0000	1.0000	0.5000	-1.0000	-0.7500	0.7500	0.5000
0.5000	0.5000	0.5000	-1.0000	0.5000	0.5000	-0.7500	-0.5000
-1.0000	1.0000	0.7500	1.0000	0.5000	0.7500	-0.5000	0.5000
-2.0000	-2.0000	-2.0000	-2.0000	-2.0000	-2.0000	-2.0000	-2.0000

{1,2}	{1,3}	{2,3}	{3,4}	{5,6}	{5,7}	{5,8}	{6,7}
0.6000	-0.3000	-0.6000	0.3000	0.4500	-0.4500	0.4500	-0.4500
-0.3000	0.6000	-0.3000	0.4500	0.4500	-0.6000	-0.4500	-0.3000
-0.6000	0.6000	-0.3000	0.3000	0.3000	0.4500	0.4500	-0.6000
0.6000	-0.3000	-0.3000	0.3000	0.4500	-0.3000	-0.3000	0.4500

-0.6000	-0.4500	-0.6000	-0.4500	0.6000	0.4500	-0.3000	-0.3000
1.2000	1.2000	1.2000	1.2000	1.2000	1.2000	1.2000	1.2000

{6,8}	{7,8}	{1,2,3}	{5,6,7}	{5,6,8}	{5,7,8}	{6,7,8}	{5,6,7,8}
0.3000	0.4500	0.6000	0.6000	-0.6000	0.3000	0.4500	-0.6000
-0.6000	-0.4500	-0.3000	0.4500	-0.6000	-0.6000	-0.6000	-0.4500
-0.4500	0.6000	0.6000	0.4500	-0.6000	-0.3000	0.6000	0.6000
0.3000	-0.6000	0.3000	-0.6000	0.3000	-0.6000	0.6000	-0.3000
-0.4500	-0.4500	0.3000	0.4500	-0.6000	-0.3000	0.3000	-0.6000
1.2000	1.2000	1.2000	1.2000	1.2000	1.2000	1.2000	1.2000

Graph 4, p = 5

{1}	{2}	{3}	{4}	{5}	{6}	{7}	{8}
-0.5000	-0.7500	0.5000	0.7500	0.7500	-1.0000	0.5000	-0.7500
-1.0000	-0.7500	-0.7500	0.5000	-1.0000	-0.7500	-0.5000	0.7500
0.5000	-1.0000	-1.0000	-1.0000	0.7500	0.5000	0.7500	-0.7500
-0.5000	-0.5000	0.5000	1.0000	0.5000	-1.0000	0.7500	-0.5000
0.7500	1.0000	1.0000	-0.7500	-0.7500	-0.5000	0.5000	1.0000
-1.5000	-1.5000	-1.5000	-1.2000	-1.2000	-1.5000	-1.5000	-1.5000

{9}	{10}	{1,2}	{1,3}	{2,3}	{3,4}	{4,5}	{5,6}
0.7500	0.5000	-0.6000	-0.3000	-0.6000	0.4500	-0.4500	0.3000
-0.7500	0.7500	-0.6000	-0.4500	-0.3000	0.4500	-0.6000	-0.3000
-1.0000	-1.0000	-0.6000	-0.3000	-0.3000	0.6000	-0.4500	0.3000
0.5000	-0.5000	0.6000	-0.6000	0.4500	-0.3000	0.4500	-0.4500
0.7500	-0.7500	-0.3000	0.6000	-0.3000	-0.4500	-0.6000	0.3000
-1.5000	-1.5000	1.2000	1.2000	1.2000	1.5000	1.8000	1.2000

{5,7}	{5,8}	{6,7}	{6,8}	{6,9}	{7,8}	{9,10}	{1,2,3}
-0.4500	0.4500	-0.4500	0.4500	-0.4500	0.4500	-0.6000	-0.3000
-0.6000	-0.4500	0.3000	0.4500	0.6000	0.4500	0.6000	0.3000

```
-0.4500    0.4500   -0.3000    0.4500    0.6000    0.6000   -0.3000   -0.3000
-0.4500    0.3000   -0.3000    0.4500   -0.3000   -0.4500    0.3000   -0.6000
-0.4500    0.3000   -0.3000    0.6000    0.6000    0.3000    0.6000    0.3000
 1.2000    1.2000    1.2000    1.2000    1.2000    1.2000    1.8000    2.0000

 {5,6,7}   {5,6,8}   {5,7,8}   {6,7,8}   {5,6,7,8}
-0.3000    0.4500   -0.3000    0.3000   -0.3000
-0.6000   -0.6000   -0.6000    0.6000   -0.3000
 0.3000   -0.6000    0.3000   -0.4500    0.6000
-0.6000   -0.3000    0.3000   -0.6000    0.4500
-0.6000   -0.4500    0.4500   -0.6000   -0.3000
-0.9000   -0.5000   -0.5000   -0.5000    2.4000
```

Bibliography

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Second international symposium on information theory*, volume 1, pages 267–281. Springer Verlag, 1973.

[2] R. Akerkar. *Nonlinear functional analysis*. Narosa Publishing House, 1999.

[3] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73 (3):243–272, 2008.

[4] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68, 1950.

[5] H. Attouch. *Variational convergence for functions and operators*, volume 1. Pitman Advanced Pub. Program, 1984.

[6] F.R. Bach. Consistency of the group lasso and multiple kernel learning. *The Journal of Machine Learning Research*, 9:1179–1225, 2008.

[7] F.R. Bach, G.R.G. Lanckriet, and M.I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *Proceedings of the twenty-first international conference on Machine learning*. ACM New York, NY, USA, 2004.

[8] O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516, 2008. ISSN 1532-4435.

[9] V. Barbu and T. Precupanu. *Convexity and optimization in Banach spaces*. Springer Verlag, 2012.

[10] C.M. Bishop. *Pattern recognition and machine learning*. Springer New York, 2006.

[11] J.K. Bradley and C. Guestrin. Learning tree conditional random fields. In *Proceedings of the 27th International Conference on Machine learning*, pages 127–134, 2010.

[12] D. Buchmann, M. Schmidt, S. Mohamed, D. Poole, and N. de Freitas. On sparse, spectral and other parameterizations of binary probabilistic models. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

[13] A. Caponnetto, C.A. Micchelli, M. Pontil, and Y. Ying. Universal multi-task kernels. *The Journal of Machine Learning Research*, 9:1615–1646, 2008.

[14] C. Carmeli, E. De Vito, and A. Toigo. Vector valued reproducing kernel hilbert spaces of integrable functions and mercer theorem. *Analysis and Applications*, 4(4):377, 2006.

[15] R. Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[16] P. Craven and G. Wahba. Smoothing noisy data with spline functions. *Numerische Mathematik*, 31 (4):377–403, 1979.

[17] C. De Boor. *A practical guide to splines*. Applied Mathematical Sciences, 1978.

[18] C. De Boor. *B (asic)-spline basics*. Mathematics Research Center, University of Wisconsin-Madison, 1986.

[19] M.X. Dong and R.J.B. Wets. Estimating density functions: a constrained maximum likelihood approach. *Journal of Nonparametric Statistics*, 12(4):549–595, 2000.

[20] Z. Duan, L. Lu, and C. Zhang. Collective annotation of music from multiple semantic sategories. In *Proceedings of 9th International Conference on Music Information Retrieval*, pages 237–242, Philadelphia, USA, 2008.

[21] B. Efron. The estimation of prediction error. *Journal of the American Statistical Association*, 99 (467):619–632, 2004.

[22] T. Evgeniou, C.A. Micchelli, and M. Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6(1):615, 2005.

[23] J. Friedman, T. Hastie, and R. Tibshirani. A note on the group lasso and a sparse group lasso. *Arxiv preprint arXiv:1001.0736*, 2010.

[24] J.H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19:1–67, 1991.

[25] F. Gao, G. Wahba, R. Klein, and B. Klein. Smoothing Spline ANOVA for multivariate Bernoulli observations, with application to ophthalmology data. *Journal of the American Statistical Association*, 96(453):127, 2001.

[26] C.J. Geyer. On the asymptotics of constrained m-estimation. *The Annals of Statistics*, 22:1993–2010, 1994.

[27] C.J. Geyer. On the asymptotics of convex stochastic optimization. *Unpublished manuscript*, 1996.

[28] G.H. Golub, M. Heath, and G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, pages 215–223, 1979.

[29] I.J. Good and R.A. Gaskins. Nonparametric roughness penalties for probability densities. *Biometrika*, 58(2):255–277, 1971.

[30] P.R. Halmos. *Introduction to Hilbert space and the theory of spectral multiplicity*. Chelsea Pub Co, 1998.

[31] T.J. Hastie and R.J. Tibshirani. *Generalized additive models*. Chapman & Hall/CRC, 1990.

[32] C. Hess. Epi-convergence of sequences of normal integrands and strong consistency of the maximum likelihood estimator. *The Annals of Statistics*, 24(3):1298–1315, 1996.

[33] H. Höfling and R. Tibshirani. Estimation of sparse binary pairwise markov networks using pseudo-likelihoods. *The Journal of Machine Learning Research*, 10:883–906, 2009.

[34] J. Huang, J.L. Horowitz, and F. Wei. Variable selection in nonparametric additive models. *Annals of statistics*, 38(4):2282, 2010.

[35] L. Jacob, G. Obozinski, and J.P. Vert. Group Lasso with overlap and graph Lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 433–440, 2009.

[36] R. Jenatton, J.Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. *Journal of Machine Learning Research*, 12:2777–2824, 2011.

[37] J. Jia and B. Yu. On model selection consistency of elastic net when p $>>$ n. *Statistica Sinica*, 20: 595–611, 2010.

[38] S. Kim and E.P. Xing. Tree-guided group lasso for multi-task regression with structured sparsity. In *Proceedings of 27th International Conference on Machine Learning*, pages 543–550, Haifa, Israel, 2010.

[39] G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95, 1971.

[40] R. Kindermann, J.L. Snell, and American Mathematical Society. *Markov random fields and their applications*. American Mathematical Society Providence, RI, 1980.

[41] K. Knight. Epi-convergence in distribution and stochastic equi-semicontinuity. *Unpublished manuscript*, 1999.

[42] K. Knight and W. Fu. Asymptotics for lasso-type estimators. *Annals of Statistics*, pages 1356–1378, 2000.

[43] K. Koh, S.J. Kim, and S. Boyd. An interior-point method for large-scale l1-regularized logistic regression. *Journal of Machine learning research*, 8(8):1519–1555, 2007.

[44] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009. ISBN 0262013193.

[45] V. Koltchinskii and M. Yuan. Sparsity in multiple kernel learning. *The Annals of Statistics*, 38(6): 3660–3695, 2010.

[46] L.A. Korf and R.J.B. Wets. Random lsc functions: An ergodic theorem. *Mathematics of Operations Research*, pages 421–445, 2001.

[47] P. Lachout. Epi-convergence almost surely, in probability and in distribution. *Annals of Operations Research*, 142(1):187–214, 2006.

[48] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289, 2001.

[49] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer, 1991.

[50] K.C. Li. From stein's unbiased risk estimates to the method of generalized cross validation. *The Annals of Statistics*, 13(4):1352–1377, 1985.

[51] K.C. Li. Asymptotic optimality of c_l and generalized cross-validation in ridge regression with application to spline smoothing. *The Annals of Statistics*, 14(3):1101–1112, 1986.

[52] K.C. Li. Asymptotic optimality for cp, cl, cross-validation and generalized cross-validation: discrete index set. *The Annals of Statistics*, 15(3):958–975, 1987.

[53] Y. Lin and H.H. Zhang. Component selection and smoothing in smoothing spline analysis of variance models. *Annals of Statistics*, 34(5):2272–2297, 2006.

[54] H. Liu and J. Zhang. Estimation consistency of the group lasso and its applications. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, 2009.

[55] H. Liu, X. Chen, J. Lafferty, and L Wasserman. Graph-valued regression. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1423–1431. 2010.

[56] J. Liu, S. Ji, and J. Ye. *SLEP: Sparse Learning with Efficient Projections*. Arizona State University, 2009. URL `http://www.public.asu.edu/~jye02/Software/SLEP`.

[57] D.G. Luenberger. *Optimization by vector space methods*. Wiley-Interscience, 1997.

[58] X. Ma. *Penalized Regression in Reproducing Kernel Hilbert Spaces With Randomized Covariate Data*. PhD thesis, Department of Statistics, University of Wisconsin-Madison, 2010.

[59] J. Mairal, R. Jenatton, G. Obozinski, and Bach F. Network flow algorithms for structured sparsity. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1558–1566. 2010.

[60] P. McCullagh and J.A. Nelder. *Generalized linear models*. Chapman & Hall/CRC, 1989.

[61] L. Meier, S. Van De Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

[62] L. Meier, S. Van De Geer, and P. Bühlmann. High-dimensional additive modeling. *The Annals of Statistics*, 37(6B):3779–3821, 2009.

[63] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

[64] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1):246–270, 2009.

[65] C.A. Micchelli and M. Pontil. On Learning Vector-Valued Functions. *Neural Computation*, 17(1): 177–204, 2005.

[66] J. Nocedal and S.J. Wright. *Numerical optimization*. Springer verlag, 1999.

[67] J. Peng, P. Wang, N. Zhou, and J. Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009.

[68] D. Percival. Theoretical properties of the overlapping groups lasso. *Electronic Journal of Statistics*, 6:269–288, 2012.

[69] P. Radchenko and G.M. James. Variable selection using adaptive nonlinear interaction structures in high dimensions. *Journal of the American Statistical Association*, 105(492):1541–1553, 2010.

[70] P. Ravikumar, J. Lafferty, H. Liu, and L. Wasserman. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030, 2009.

[71] P. Ravikumar, M.J. Wainwright, and J. Lafferty. High-dimensional Ising model selection using l1-regularized logistic regression. *Annals of Statistics*, 38(3):1287–1319, 2010.

[72] G.V. Rocha, X. Wang, and B. Yu. Asymptotic distribution and sparsistency for l1-penalized parametric m-estimators with applications to linear svm and logistic regression. *Arxiv preprint arXiv:0908.1940*, 2009.

[73] R.T. Rockafellar and R.J.B. Wets. *Variational analysis*, volume 317. Springer Verlag, 1998.

[74] R.M. Scammon, A.V. McGillivray, and R. Cook. *America Votes 26: 2003-2004, Election Returns By State*. CQ Press, 2005. ISBN 9781568029740.

[75] M. Schmidt and K. Murphy. Convex structure learning in log-linear models: Beyond pairwise potentials. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

[76] M. Schmidt, K. Murphy, G. Fung, and R. Rosales. Structure learning in random fields for heart motion abnormality detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[77] P. Schnitzspan, M. Fritz, S. Roth, and B. Schiele. Discriminative structure learning of hierarchical representations for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2238–2245, 2009.

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

[79] W. Shi, G. Wahba, S. Wright, K. Lee, R. Klein, and B. Klein. LASSO-Patternsearch algorithm with application to ophthalmology and genomic data. *Statistics and its Interface*, 1(1):137, 2008.

[80] T.P. Speed and H.T. Kiiveri. Gaussian Markov distributions over finite graphs. *The Annals of Statistics*, 14(1):138–150, 1986. ISSN 0090-5364.

[81] C.M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9 (6):1135–1151, 1981.

[82] C. Sutton, A. McCallum, and K. Rohanimanesh. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. *The Journal of Machine Learning Research*, 8:693–723, 2007. ISSN 1532-4435.

[83] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. A comparative study of energy minimization methods for markov random fields with smoothness-based priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30: 1068–1080, June 2007. ISSN 0162-8828.

[84] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[85] A. Torralba, K.P. Murphy, and W.T. Freeman. Sharing features: efficient boosting procedures for multiclass object detection. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 762 – 769, 2004.

[86] S.A. Van De Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645, 2008.

[87] A.W. Van Der Vaart and J.A. Wellner. *Weak convergence and empirical processes*. Springer Verlag, 1996.

[88] S. Vogel and P. Lachout. On continuous convergence and epi-convergence of random functions. *Kybernetika*, 39(1):75–98, 2003.

[89] M. Volle and J.B. Hiriart-Urruty. A characterization of essentially strictly convex functions on reflexive banach spaces. *Nonlinear Analysis: Theory, Methods & Applications*, 2011.

[90] G. Wahba. *Spline Models for Observational Data*. Society for Industrial Mathematics, 1990.

[91] G. Wahba. Multivariate Function and Operator Estimation, Based on Smoothing Splines and Reproducing Kernels. In M. Casdagli and S. Eubank, editors, *Nonlinear Modeling and Forecasting, Santa Fe Institue Studies in the the Sciences of Complexity*, volume 12, pages 95–112. ADDISON-WESLEY PUBLISHING CO, 1992.

[92] G. Wahba, Y. Wang, C. Gu, R. Klein, and B. Klein. Smoothing spline anova for exponential families, with application to the wisconsin epidemiological study of diabetic retinopathy: the 1994 neyman memorial lecture. *The Annals of Statistics*, 23(6):1865–1895, 1995.

[93] M.J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.

[94] M.J. Wainwright and M.I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1:1–305, 2008. ISSN 1935-8237.

[95] J. Whittaker. *Graphical models in applied multivariate statistics*. Wiley (Chichester England and New York), 1990.

[96] S.J. Wright. Accelerated block-coordinate relaxation for regularized optimization. Technical report, Department of Computer Science, University of Wisconsin-Madison, 2010.

[97] D. Xiang and G. Wahba. A generalized approximate cross validation for smoothing splines with non-Gaussian data. *Statistica Sinica*, 6:675–692, 1996. ISSN 1017-0405.

[98] L. Xue, H. Zou, and T. Cai. Non-concave penalized composite likelihood estimation of sparse ising models. Technical report, Department of Statistics, University of Minnesota, 2010.

[99] Lei Yuan, Jun Liu, and Jieping Ye. Efficient methods for overlapping group lasso. In J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F.C.N. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 352–360. 2011.

[100] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[101] M. Yuan and Y. Lin. On the non-negative garrotte estimator. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):143–161, 2007.

[102] P. Zhao and B. Yu. On Model Selection Consistency of Lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

[103] P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics*, 37(6A):3468–3497, 2009.

[104] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.