Blessing of Massive Scale: Spatial Graphical Model Estimation with a Total Cardinality Constraint

Ethan X. Fang* Han Liu† Mengdi Wang‡

November 12, 2015

Abstract

We consider the problem of estimating high dimensional spatial graphical models with a total cardinality constraint (i.e., the ℓ_0-constraint). Though this problem is highly nonconvex, we show that its primal-dual gap diminishes linearly with the dimensionality and provide a convex geometry justification of this 'blessing of massive scale' phenomenon. Motivated by this result, we propose an efficient algorithm to solve the dual problem (which is concave) and prove that the solution achieves optimal statistical properties. Extensive numerical results are also provided.

Keywords: graphical models, ℓ_0-constraint method, nonconvex combinatorial optimization, high-dimensional data, computational complexity

1 Introduction

We consider the problem of estimating high dimensional spatial graphical models. More specifically, let $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ be a d-dimensional random vector on a spatial field (e.g., a lattice). We aim to find an undirected graph $G = (V, E)$ with vertex set $V = \{1, 2, ..., d\}$ and edge set $E \subset V \times V$ to encode the conditional independence of $\boldsymbol{X}$, i.e., $(j, k) \in E$ if and only if X_j and X_k are conditionally independent given the remaining variables. A spatial graphical model also requires the graph G to be conformed with the spatial proximity. In other words, a necessary condition for the existence of edge $(j, k) \in E$ is that vertices j and k are spatially closed (more details are provided later.).

1.1 Motivating Applications of Spatial Graphical Models

Spatial graphical models find important real-world applications. We provide two concrete motivating examples. The first application is to infer the topology of sensor network on a 2D surface. The

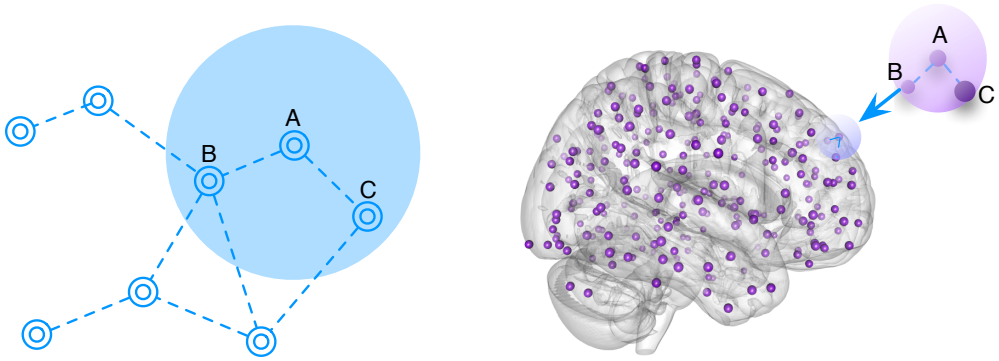
*Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA; e-mail: xingyuan@princeton.edu

†Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA; e-mail: hanliu@princeton.edu

‡Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA; e-mail: mengdiw@princeton.edu

wireless sensor network is widely used in various applications including agriculture (Langendoen et al., 2006), military (Lee et al., 2009) and environmental science (Howard et al., 2002). See Yick et al. (2008) for a survey. In these applications, it is important to understand how the sensors interact with one another. In practice, each sensor can only communicate with other sensors that are geographically close. Also, in applications such as agricultural and environmental studies, all sensors' corresponding locations are known. Thus, the spatial proximity information of these sensors are available a priori. See Figure 1(a) for a simple illustration.

Another example is to estimate the short-range brain network. In these networks, the vertices are voxels or ROIs (region of interest) embedded in the 3D space. Given the brain imaging data and the spatial information, we aim to estimate the graphical model under the constraint that each vertex can only connect to the nodes that are physically close (Cao and Fei-Fei, 2007). See Figure 1(b) for an illustrative example. The estimated graph serves a first step for more sophisticated downstream analysis (e.g., long-range brain network construction or functional region partition (Bullmore and Sporns, 2009; Bullmore and Bassett, 2011)).



(a) (b)

Figure 1: (a) Sensor network example: In the network, each sensor can only connect to another sensor if they are physically close on the plane. In the figure, each dashed line represents a possible connection. For example, sensor A can only possibly connect to B or C, but not others. (b) Brain network example: In a brain network, each vertex can only connect to another if they are close in the 3-D space. Taking vertex A for example, it can only possibly connect to vertices B or C.

There are many other applications of spatial graphical models. For example, in global weather analysis, we are interested in how the weathers at various locations interact with one another, and two different locations can be conditionally dependent only if they are sufficiently close.

1.2 Main Contributions

Under many statistical models, such as Gaussian or Ising models, we formulate the spatial graphical model estimation problems as

$$\min_{\boldsymbol{\beta}_j \in \mathcal{C}_j} \frac{1}{d} \sum_{j=1}^{d} \mathcal{L}_j(\boldsymbol{\beta}_j), \text{ subject to } \sum_{j=1}^{d} \mathcal{R}_j(\boldsymbol{\beta}_j) \leq K. \tag{1.1}$$

For each vertex j, $\mathcal{L}_j : \mathbb{R}^{d_j} \to \mathbb{R}$ is some convex loss function associated with the statistical model. $\mathcal{R}_j(\cdot)$ is some nonconvex function, such as smoothly clipped absolute deviation (SCAD), minimax concave penalty (MCP) or the ℓ_0-(pseudo)norm function. The feasible set $\mathcal{C}_j$ is a closed set, and K is some tuning parameter representing the desired sparsity level. Denote the global minimizer of the problem by $\{\widetilde{\boldsymbol{\beta}}_j\}_{j=1}^d$. The corresponding estimated graph is $\widetilde{G} = (V, \widetilde{E})$, where $(j,k) \in \widetilde{E}$ if and only if $\widetilde{\beta}_{jk} \neq 0$ or $\widetilde{\beta}_{kj} \neq 0$. Given the spatial proximity information, we have that each node j can only connect to a set of vertices $\mathcal{N}_j \subset \{1, ..., d\}$, where $|\mathcal{N}_j| = d_j \ll d$. Then, each set $\mathcal{C}_j \subset \mathbb{R}^{d_j}$ is of small dimensions, and this makes each subproblem $\min_{\boldsymbol{\beta}_j \in \mathcal{C}_j} \mathcal{L}_j(\boldsymbol{\beta}_j)$ small dimensional. For example, under the Gaussian model, we let $\mathcal{L}_j(\boldsymbol{\beta}_j) = n^{-1}\|\mathbb{X}_j - \mathbb{X}_{\mathcal{N}_j}\boldsymbol{\beta}_j\|_2^2$, where $\mathbb{X}_j \in \mathbb{R}^n$ is the data vector corresponding to vertex X_j, and $\mathbb{X}_{\mathcal{N}_j} \in \mathbb{R}^{n \times d_j}$ is the data which corresponds to the potential neighbors of vertex X_j.

Our first contribution is the characterization of the complexity of problem (1.1). Suppose that the number of potential neighbors per-vertex is fixed. We prove that solving problem (1.1) is NP-complete and thus difficult in general. However, for the special case $\mathcal{R}_j(\boldsymbol{\beta}_j) = \|\boldsymbol{\beta}_j\|_0$, we discover that the problem becomes polynomial-time solvable by a dynamic programming algorithm. We further prove that if $\mathcal{R}_j$ is a vector-valued constraint, the problem becomes fundamentally more difficult. For example, we cannot solve the problem if $\mathcal{R}_j(\boldsymbol{\beta}_j) = (\|\boldsymbol{\beta}_j\|_0, \sum_{k=1}^{d_j-1} \|\beta_{j,k+1} - \beta_{jk}\|_0)^T$ unless P = NP.

Our second contribution is a scalable algorithm to solve problem (1.1). Although there exists a polynomial-time dynamic programming algorithm to solve the problem when $\mathcal{R}_j(\boldsymbol{\beta}_j) = \|\boldsymbol{\beta}_j\|_0$, the algorithm is not tractable in practice. To achieve a more practical algorithm, we develop a Splitting-Communicating (SPICA) algorithm which solves the Lagrangian dual of the primal problem (1.1). The algorithm utilizes the separable structure of the dual problem and converges to a dual optimal solution geometrically. Since problem (1.1) is nonconvex, there exists a positive duality gap between the primal and dual optimal solutions. We prove that the average-per-vertex duality gap diminishes at the rate of $\mathcal{O}(d^{-1})$ as the graph dimension d increases. As a result, if the dimension d is large, the dual optimal solution is close to the primal optimal solution, and achieves optimal statistical properties. This reveals a "blessing of massive scale" phenomenon.

1.3 Relationship with Existing Literature

The problem (1.1) is highly nonconvex and raises computational challenges. To overcome the challenges, several existing works rely on solving optimization problems derived from convex relaxations, such as the ℓ_1 or semidefinite relaxations (Tibshirani, 1996; d'Aspremont et al., 2007). The motivation of different convex relaxations is to avoid solving nonconvex or combinatorial optimization problems while still achieves fast statistical rates. Meanwhile, some efficient greedy methods are proposed (Tropp et al., 2004). Extensive works study the theoretical guarantees of different relaxations or greedy methods in various models, and achieve optimal minimax lower bounds under certain regularity assumptions. See Meinshausen and Bühlmann (2006); Zhao and Yu (2006); Meinshausen and Yu (2009); Zhang (2009); Meinshausen and Bühlmann (2010); Liu and Wang (2012); Vu et al. (2013) and Lei and Vu (2014). However, some results prove that there are some unavoidable statistical losses for the estimators derived from these methods in several important problems, such

as sparse PCA (Amini and Wainwright, 2009; Berthet and Rigollet, 2013a,b; Krauthgamer et al., 2013), low-rank matrix problems (Oymak et al., 2013), and linear regression (Zhang et al., 2014). For example, in linear regression, it is well-known that the ℓ_0-constrained estimator $\widehat{\boldsymbol{\beta}}_0$ obtains optimal rate of convergence that $n^{-1}\mathbb{E}\big\{\|\mathbb{X}\widehat{\boldsymbol{\beta}}_0 - \mathbb{X}\boldsymbol{\beta}^*\|_2^2\big\} = \mathcal{O}(n^{-1}s\log d)$, where $s = \|\boldsymbol{\beta}^*\|_0$. In comparison, without restricted eigenvalue-type assumptions, other polynomial-time methods can only achieve a slower rate (Zhang et al., 2014).

Our work develops a novel framework to estimate high-dimensional spatial graphical models by directly attacking the nonconvex problem under the total cardinality constraint (1.1). The proposed algorithm produces a near-optimal solution to the optimization problem, which achieves optimal statistical properties.

Notations. Let $\mathbb{X} \in \mathbb{R}^{n\times d}$ be the data matrix, and $\mathbb{X}_j$ denotes the j-th column of $\mathbb{X}$. Also, $\mathbb{X}_{\mathcal{N}_j}$ denotes the columns of possible neighbors of X_j, where $\mathcal{N}_j$ is the set of possible neighbors of X_j. For a matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, we denote its maximum eigenvalue by $\Lambda_{\max}(\mathbf{A})$, and its minimum eigenvalue by $\Lambda_{\min}(\mathbf{A})$. We denote by $[d] = \{1, \ldots, d\}$.

Paper Organization. The rest of this paper is organized as follows. Section 2 characterizes the complexity of problem (1.1). Section 3 describes the SPICA algorithm and illustrates its main idea using geometric nonconvex duality. Section 4 provides theoretical guarantees of the "blessing of massiveness" phenomenon, and we analyze the statistical properties of the estimators. Section 5 provides extensive numerical experiments using both synthetic and sensor network data

2 The Complexity of Spatial-Graphical Model Problem

In this section, we study the complexity of problem (1.1) by relating it to a classical discrete NP-complete problem - the knapsack problem. For general constraints $\mathcal{R}_j$'s, we show that problem (1.1) is NP-complete. In the special case of $\mathcal{R}_j(\boldsymbol{\beta}_j) = \|\boldsymbol{\beta}_j\|_0$, the problem admits a polynomial-time solution. In the general setting where the nonconvex constraints $\mathcal{R}_j$'s are vector-valued, we prove that the problem does not admit a fully polynomial-time approximation scheme unless P = NP, and the problem is fundamentally more difficult. For example, we cannot solve problems under the constraint $\sum_{j\in[d]} \mathcal{R}_j(\boldsymbol{\beta}_j) \le (K_1, K_2)^T$, where $\mathcal{R}_j(\boldsymbol{\beta}_j) = (\|\boldsymbol{\beta}_j\|_0, \sum_{k=1}^{d_j-1} \|\beta_{j,k+1} - \beta_{jk}\|_0)^T$.

2.1 Knapsack Problem and Complexity

The knapsack problem plays an important role in combinatorics. It is motivated from applications in resource allocation, where the goal is to maximize the total utility under capacity constraints. Its simplest form is the following 0-1 knapsack problem:

$$\max_{x_j} \sum_{j=1}^{d} c_j x_j, \text{ subject to } \sum_{j=1}^{d} b_j x_j \le b_0, \ x_j \in \{0,1\}, \quad \text{for } j = 1, .., d, \tag{2.1}$$

where c_j's, b_j's and b_0 are positive integers. The input to the 0-1 knapsack problem includes: the constant c_j which is the value of the j-th item; the constant b_j which is the cost of the j-th item,

and the constant b_0 which is the total budget. Let $\mathbf{c} = (c_1, ..., c_d)^T$ and $\mathbf{b} = (b_1, ..., b_d)^T \in \mathbb{R}^d$. We refer to problem (2.1) as the 0-1 knapsack problem with input $(\mathbf{c}, \mathbf{b}, b_0)$. This problem is known to be NP-complete (Williamson and Shmoys, 2011).

An important variant of the 0-1 knapsack problem is the *multiple-row knapsack problem* (also known as the multiple-dimensional knapsack problem):

$$\max_{x_j} \sum_{j=1}^{d} c_j x_j, \quad \text{subject to } \sum_{j=1}^{d} b_j^{(\ell)} x_j \le b_0^{(\ell)}, \quad \text{for all } \ell = 1, ..., L, \ x_j \in \{0, 1\}. \tag{2.2}$$

In comparison with the 0-1 knapsack problem, this problem has multiple-row constraints. The multiple-row knapsack problem is fundamentally more difficult than the 0-1 knapsack problem. It is NP-hard to solve the problem to an arbitrary precision. More specifically, it is shown that finding a fully polynomial time approximation scheme for the multiple-row knapsack problem is NP-hard (Magazine and Chern, 1984), which is defined below.

Definition 2.1. An approximation scheme for a maximization problem (P) is an algorithm that takes two inputs: One is the problem instance P, and the other is a desired numerical accuracy $\epsilon > 0$. Denote by $f^* > 0$ the optimal value of P. The algorithm produces a solution for P with objective value $f(P)$ such that $\{f^* - f(P)\}/f^* \le \epsilon$. If the running time for the algorithm is bounded by a polynomial function of $1/\epsilon$ and the problem size, it is a fully polynomial time approximation scheme.

To facilitate our later discussion, we briefly review some definitions in computational complexity theory in Appendix Section A. We refer to Williamson and Shmoys (2011) for more detailed discussion about the knapsack problem and computational complexity theory.

2.2 NP-Completeness of Problem (1.1)

In this subsection, we prove that problem (1.1) is NP-complete. For ease of presentation, we assume all $\boldsymbol{\beta}_j$'s are of identical dimensions d_0, i.e., $\boldsymbol{\beta}_j \in \mathbb{R}^{d_0}$ for all j, where d_0 is a given constant.

To prove that problem (1.1) is NP-complete, we shall construct a two-way polynomial time reduction between problem (1.1) and 0-1 knapsack problem (2.1). We show that given one instance of the 0-1 knapsack problem or problem (1.1), we can construct another instance of the other problem within a polynomial-time, and by solving the other instance we can recover the solution to the original instance. As we discussed in the introduction, the form of loss functions $\mathcal{L}_j$'s depends on the specific statistical model. Without loss of generality, we assume all $\mathcal{L}_j$'s are of a same form (least square or logistic loss for example), and each $\mathcal{L}_j$ only depends on some input data $(\mathbb{X}_j, \mathbb{Y}_j)$. Thus, each $\mathcal{L}_j(\boldsymbol{\beta}_j)$ can be represented as $\mathcal{L}(\boldsymbol{\beta}_j; \mathbb{X}_j, \mathbb{Y}_j)$. We consider finding an ϵ-optimal solution to the problem

$$\min_{\boldsymbol{\beta}_j} \sum_{j=1}^{d} \mathcal{L}(\boldsymbol{\beta}_j; \mathbb{X}_j, \mathbb{Y}_j), \ \text{subject to } \sum_{j=1}^{d} \mathcal{R}_j(\boldsymbol{\beta}_j) \le b_0, \tag{2.3}$$

with input $\big(\{\mathbb{X}_j, \mathbb{Y}_j\}_{j\in[d]}, b_0, \epsilon\big)$, where we say a solution is ϵ-optimal if its corresponding objective value is within ϵ of the optimal value, and the solution is feasible. Problem (2.3) can be continuous since both objective and constraint functions in (2.3) can be continuous, and 0-1 knapsack problem

is discrete. To connect the two problems, we need to "discretize" problem (1.1). We first consider the loss functions. We impose the following assumption.

- (A.1) Given positive constants c_j's for all $j \in [d]$, we can find $\mathbb{X}_j$, $\mathbb{Y}_j$ and a constant c_0 within a polynomial time such that

$$\mathcal{L}(0; \mathbb{X}_j, \mathbb{Y}_j) = c_0 \quad \text{and} \quad \min_{\beta_j \in \mathbb{R}} \mathcal{L}(\beta_j; \mathbb{X}_j, \mathbb{Y}_j) = -c_j + c_0.$$

This assumption is satisfied for most statistical models. For example, if $\mathcal{L}(\beta_j; \mathbb{X}_j, \mathbb{Y}_j) = \|\mathbb{Y}_j - \mathbb{X}_j\beta_j\|_2^2/n$, it is easy to verify that letting $\mathbb{Y}_j = (\sqrt{c_0}, \sqrt{c_0})^T$ and $\mathbb{X}_j = \left(\sqrt{c_j + c_j'}, \sqrt{c_j - c_j'}\right)^T$, where $c_j' = \sqrt{2c_jc_0 - c_0^2}$, satisfy this assumption.

Next, we look at constraint functions $\mathcal{R}_j$'s. Given a problem instance of (2.3), we need to efficiently construct a knapsack problem of which the constraint is similar to the problem instance (2.3). Since the knapsack problem is discrete, and problem (2.3) is possibly continuous, we assume that we can efficiently "discretize" the constraint functions $\mathcal{R}_j$'s, where we impose the following assumption.

- (A.2) For any j, given any $\delta > 0$ and any set $[-r, r]^{d_0}$ for some $r > 0$, we can find a finite discretization $\mathcal{B}$ of the set that for any point $\boldsymbol{\beta} \in [-r, r]^{d_0}$, there exists a point $\mathbf{p} \in \mathcal{B}$ such that $\|\mathbf{p} - \boldsymbol{\beta}\|_2 \leq \delta$ and $\mathcal{R}_j(\mathbf{p}) \leq \mathcal{R}_j(\boldsymbol{\beta})$ in a polynomial time.

This assumption holds for most common $\mathcal{R}_j$'s in statistical applications. For example, suppose $\mathcal{R}_j$'s are SCAD functions. We have that the discretization $\{0, \pm\sqrt{\delta/d_0}, \pm 2\sqrt{\delta/d_0}, ..., \pm p^*\sqrt{\delta/d_0}\}^{d_0}$ satisfies the assumption, where $p^* = \text{argmax}_{p \in \mathbb{N}}\{p\sqrt{\delta/d_0} \leq r\}$, and $\mathbb{N}$ is the set of natural numbers.

Next, we provide the main theorem of this section. We show that given one instance of 0-1 knapsack problem (2.1), we can construct an instance of problem (2.3) within a polynomial-time, and by solving the instance of problem (2.3) we can recover the solution to the original instance of 0-1 knapsack problem, and vice versa. This proves that problem (2.3) is NP-complete since 0-1 knapsack problem (2.1) is known to be NP-complete.

Theorem 2.2. Under assumptions (A.1)-(A.2), the nonconvex constrained optimization problem (2.3) is NP-complete.

Proof. See Appendix B for the detailed proof. □

2.3 Polynomial-Time Algorithm in the Case of ℓ_0-Constrained Problem

Though problem (2.3) is NP-complete, we show that the special case of problem (2.3) under a total cardinality constraint, i.e., the case where $\mathcal{R}_j(\boldsymbol{\beta}_j) = \|\boldsymbol{\beta}_j\|_0$, admits a polynomial-time algorithm. In particular, given an instance of problem (2.3), we can map it to an instance of multiple-choice knapsack problem, and by solving the instance of multiple-choice knapsack problem efficiently, we can recover the solution to the instance of problem (2.3).

Let us first introduce multiple-choice knapsack problem. Denote by $\mathbf{C} = (\mathbf{c}_1, ..., \mathbf{c}_d)^T \in \mathbb{R}^{d \times d_0}$ and $\mathbf{B} = (\mathbf{b}_1, ..., \mathbf{b}_d)^T \in \mathbb{R}^{d \times d_0}$, where $\mathbf{c}_j = (c_{j1}, ..., c_{jd_0})^T, \mathbf{b}_j = (b_{j1}, ..., b_{jd_0})^T$. Consider the multiple-choice knapsack problem with input $(\mathbf{C}, \mathbf{B}, b_0)$, where all b_{jk}'s and b_0 are positive integers:

$$\max_{x_{jk}} \sum_{j=1}^{d} \sum_{k=1}^{d_0} c_{jk} x_{jk}, \text{ subject to } \sum_{j=1}^{d} \sum_{k=1}^{d_0} b_{jk} x_{jk} \leq b_0, \sum_{k=1}^{d_0} x_{jk} \leq 1, \ x_{jk} \in \{0, 1\}, \tag{2.4}$$

for all $j \in [d]$ and all $k \in [d_0]$. Given an instance of problem (2.3) under the ℓ_0-constraint, we map the instance to an instance of multiple-choice knapsack problem (2.4). For each j, we solve the subproblems

$$\widehat{\boldsymbol{\beta}}_j^{(k)} = \underset{\boldsymbol{\beta}_j \in \mathcal{C}_j}{\operatorname{argmin}} \mathcal{L}(\boldsymbol{\beta}_j; \mathbb{X}_j, \mathbb{Y}_j), \quad \text{subject to } \|\boldsymbol{\beta}_j\|_0 \leq k, \text{ for } k = 0, 1, ..., d_0.$$

Since we assume that d_0 is a constant, the cost of computing all $\widehat{\boldsymbol{\beta}}_j^{(k)}$'s increases linearly as d increases. Let $b_{jk} = k$, $b_0 = K$ and $c_{jk} = -\mathcal{L}\big(\widehat{\boldsymbol{\beta}}_j^{(k)}; \mathbb{X}_j, \mathbb{Y}_j\big) + c_0$, where $c_0 > \max_{j,k} \mathcal{L}\big(\widehat{\boldsymbol{\beta}}_j^{(k)}; \mathbb{X}_j, \mathbb{Y}_j\big)$ for $j \in [d]$ and $k \in [d_0]$. We obtain a multiple-choice knapsack problem of the form (2.4). Denote by $\{x_{jk}^*\}$ an optimal solution to the multiple-choice knapsack problem. We have that $\{x_{jk}^*\}$ recovers an optimal solution to the ℓ_0-constrained problem by setting $\widehat{\boldsymbol{\beta}}_j = \widehat{\boldsymbol{\beta}}_j^{(k)}$ if $x_{jk}^* = 1$.

Next, we present a dynamic programming approach to solve the multiple-choice knapsack problem, which is a variant of Pisinger (1995). We formulate a dynamic program with the state variable (d', k'), where $1 \leq d' \leq d$ and $0 \leq k' \leq K$. The dimension of the state space is $d(K+1)$. We define the value function of a state (d', k') to be the optimal value for the multiple-choice knapsack problem considering only multiple-choice sets 1 to d' with constraint k'. In another words, let

$$V(d', k') = \max_{x_{jk}} \sum_{j=1}^{d'} \sum_{k=0}^{d_0} c_{jk} x_{jk}, \text{ subject to } \sum_{j=1}^{d'} \sum_{k=0}^{d_0} k x_{jk} \leq k', \sum_{k=0}^{d_0} x_{jk} \leq 1, \ x_{jk} \in \{0, 1\}.$$

Thus, $V(d, K)$ is the optimal value for the original multiple-choice knapsack problem. To facilitate our discussion, fixing c_{jk}'s, we denote the knapsack problem with first d' multiple choice set and constraint variable k' by $(MK_{d',k'})$, i.e., we let

$$(MK_{d',k'}): \quad \max_{x_{jk}} \sum_{j=1}^{d'} \sum_{k=0}^{d_0} c_{jk} x_{jk}, \text{ subject to } \sum_{j=1}^{d'} \sum_{k=0}^{d_0} k x_{jk} \leq k', \sum_{k=0}^{d_0} x_{jk} \leq 1, \ x_{jk} \in \{0, 1\}.$$

Denote by $\{x_{jk}^*\}$ the optimal solution to the problem $(MK_{d,K})$. Let $k' = \sum_{j=1}^{d'} \sum_{k=0}^{d_0} k x_{jk}^*$. To efficiently solve the problem, a key observation is that the partial solution $\{x_{jk}^*\}$ for $j = 1, ..., d'$ and $k = 0, ..., d_0$ is the optimal solution solution for the problem $(MK_{d',k'})$. This can be proved by contradiction that if the assertion does not hold, we can replace the partial solution with the optimal solution to $(MK_{d',k'})$, and we keep the rest of the original optimal solution to $(MK_{d,K})$ the same. The sum of the corresponding objectives of the two partial solutions is greater than the original optimal objective. This leads to a contradiction.

Based on this observation, we find the optimal value $V(d, K)$ by a recursive algorithm based on following recursive equations:

$$V(1, k') = \max\Big\{V(1, k'-1), \max\{c_{1k} : k \le k'\}\Big\},$$

and

$$V(d', k') = \max_k \Big\{V(d', k'-1), \max\{V(d'-1, k'-k) + c_{d'k} : k \le k'\Big\} \text{ for } d' > 1. \tag{2.5}$$

The dynamic programming algorithm for solving the problem $(MK_{d,K})$ is summarized in Algorithm 1. The total number of states is $d(K+1)$ for the problem, and the computational

Algorithm 1 Dynamic Programming Algorithm for Problem (2.4)

Input: $c_{jk} \in \mathbb{R}_+$, K

Output: x^*_{jk}

$V(d', -1) \leftarrow 0$, $V(0, k') \leftarrow 0$, $\mathcal{S}(d', k') = 0$ for all d' and k'. $d' \leftarrow 0$.
Let $\mathcal{S}(1, k') \leftarrow \text{argmax}_k\{c_{1k} : k \le k'\}$.
while $d' < d$ **do**
  Let $d' \leftarrow d' + 1$. Solve (2.5) for $1 \le k' \le K$.
  for $k' = 0 : K$ **do**
    if $\max_k \big\{V(d'-1, k'-k) + c_{d'k} : k \le k', V(d'-1, k'-k) > 0\big\} > V(d', k'-1)$ **then**
      Let $\mathcal{S}(d', k') \leftarrow \text{argmax}_k \big\{V(d'-1, k'-k) + c_{d'k} : k \le k', V(d'-1, k'-k) > 0\big\}$.
    end if
  end for
end while
$k' \leftarrow K$.
while $d' > 0$ **do**
  Let $x^*_{d', \mathcal{S}(d', k')} = 1$, $k' \leftarrow k' - \mathcal{S}(d', k')$, $d' \leftarrow d' - 1$.
end while

complexity for computing each $V(d', k')$ is $\mathcal{O}(d_0 + 1)$. Thus, the complexity of computing $V(d, K)$ is of the order $\mathcal{O}(dd_0K)$. In our problem, the number K is upper-bounded by dd_0, so the computational complexity is of the order $\mathcal{O}(d^2 d_0^2)$. Note that this does not include the computation for the coefficients c_{jk}'s. To compute all c_{jk}'s, for each sub-problem $\mathcal{L}_j$, we need to enumerate all 2^{d_0} possible combinations of the support of $\boldsymbol{\beta}_j \in \mathbb{R}^{d_0}$. Thus, applying dynamic programming techniques, the total computational complexity for solving the ℓ_0-constraint problem is of the order $\mathcal{O}(2^{d_0}d + d^2 d_0^2)$, which is still a polynomial order of the dimension d.

In summary, Algorithm 1 is a dynamic programming algorithm that runs in a polynomial-time. However, it can be very expensive in practice as it requires enumerating and solving all subproblems. In the next section, we will propose a more practical algorithm.

Meanwhile, we point out that the dynamic programming approach becomes significantly more expensive when $\mathcal{R}_j$'s are some continuous functions instead of the ℓ_0 norm. When $\mathcal{R}_j$ is continuous, our reduction to the multiple-choice knapsack problem requires a fine discretization of $\mathcal{R}_j$. This

may result in a large number of choices in the constructed knapsack problem, making the dynamic programming approach inefficient. In comparison, when $\mathcal{R}_j$ is the ℓ_0-constraint, the values of $\mathcal{R}_j$ are naturally discrete. The resulting knapsack problem has at most d_0 choices, which is a relatively small number. In general, the dynamic programming approach to problem (2.3) is practically slow, even though it is a polynomial-time algorithm.

2.4 A "Harder" Result in the Case of Vector-Valued Constraint

In this subsection, we consider the case where the functions $\mathcal{R}_j(\boldsymbol{\beta}_j)$'s are vector-valued. Specifically, we consider the problem:

$$\min_{\beta_j} \sum_{j=1}^{d} \mathcal{L}(\boldsymbol{\beta}_j; \mathbb{X}_j, \mathbb{Y}_j), \text{ subject to } \sum_{j=1}^{d} \mathcal{R}_j^{(\ell)}(\boldsymbol{\beta}_j) \leq b_0^{(\ell)}, \text{ for } \ell = 1, ..., L, \tag{2.6}$$

where $L \geq 1$. This problem contains (2.3) as a special case. In practice, one of the row-constraints can be the total sparsity constraint, and the other can be a fused-type constraint. Intuitively speaking, finding an ϵ-optimal solution to the problem (2.6) should not be more difficult than problem (2.3). However, the next theorem proves that the multiple-row constraints case (2.6) is fundamentally more difficult.

Theorem 2.3. Under assumptions (A.1)-(A.2), if $L > 1$, finding a fully polynomial-time approximation scheme for problem (2.6) is NP-hard.

Proof. The proof is based on constructing a two-way polynomial-time reduction between the multiple-row knapsack problem (2.2) and the problem (2.6). The argument is analogous to the proof of Theorem 2.2, and we omit it to avoid repetition. Consequently, as shown in Magazine and Chern (1984), there does not exist a fully polynomial-time approximation scheme to solve the two-dimensional multiple-choice knapsack problem unless we assume P = NP. □

This theorem establishes one of the strongest forms of complexity, and shows the problem (2.6) is fundamentally hard to solve. In comparison, when there exists only one total cardinality constraint, we can solve the problem within a polynomial-time by dynamic programming.

3 SPICA Algorithm for Spatial-Graph Estimation

In this section, we describe an efficient duality-based algorithm to directly attack the nonconvex problem (1.1). We illustrate the geometric intuition on why this algorithm generates a near optimal solution when the dimension d is large. We focus our discussion on the case where the problem (1.1) is subject to the ℓ_0-constraint

$$\sum_{j=1}^{d} \mathcal{R}_j(\boldsymbol{\beta}_j) = \sum_{j=1}^{d} \|\boldsymbol{\beta}_j\|_0 \leq K.$$

Note that all the analyses of this and the next sections can be generalized to other nonconvex $\mathcal{R}_j$'s. We focus our discussion on the case of ℓ_0-constraint, in which the solution achieves optimal statistical properties.

3.1 SPICA Algorithm

In this subsection, we propose an alternative algorithm to solve problem (1.1) subject to total cardinality constraint. It is practically more efficient than the dynamic programming approach and can handle problems with larger dimensions. We consider the Lagrangian dual of (1.1) when $\mathcal{R}_j(\boldsymbol{\beta}_j) = \|\boldsymbol{\beta}_j\|_0$,

$$\max_{\lambda \geq 0} \sum_{j=1}^{d} \mathcal{Q}_j(\lambda) - \lambda K, \text{ where } \mathcal{Q}_j(\lambda) = \inf_{\boldsymbol{\beta}_j \in \mathcal{C}_t} \{\mathcal{L}_j(\boldsymbol{\beta}_j) + \lambda \|\boldsymbol{\beta}_j\|_0\} \text{ for all } j = 1, ..., d. \tag{3.1}$$

The variable λ is the Lagrangian multiplier. According to literatures on duality theory (Bertsekas, 1999), even though the primal problem (1.1) is nonconvex, letting $\mathcal{Q}(\lambda) = \sum_{j \in [d]} \mathcal{Q}_j(\lambda)$, the dual $\widehat{\mathcal{Q}}(\lambda) = \mathcal{Q}(\lambda) - \lambda K$ is a concave function of λ. We aim to obtain the dual optimal solution-multiplier pair defined as

$$\begin{aligned} \widehat{\lambda} &= \operatorname*{argmax}_{\lambda \geq 0} \sum_{j=1}^{d} \mathcal{Q}_j(\lambda) - \lambda K, \text{ where } \mathcal{Q}_j(\lambda) = \mathcal{L}_j(\widehat{\boldsymbol{\beta}}_j) + \lambda \|\widehat{\boldsymbol{\beta}}_j\|_0, \\ \text{and } \widehat{\boldsymbol{\beta}}_j &= \operatorname*{argmin}_{\boldsymbol{\beta}_j \in \mathcal{C}_j} \mathcal{L}_j(\boldsymbol{\beta}_j) + \widehat{\lambda} \|\boldsymbol{\beta}_j\|_0, \text{ where } \sum_{j=1}^{d} \|\boldsymbol{\beta}_j\|_0 \leq K. \end{aligned} \tag{3.2}$$

We adopt the "golden section search" method to solve the dual problem (3.1), which runs iteratively. Let $\xi = (-1 + \sqrt{5})/2$. Given two initial points λ_1 and λ_2, let $\lambda_3 = \lambda_2 + \xi(\lambda_1 - \lambda_2)$ and $\lambda_4 = \lambda_1 + \xi(\lambda_2 - \lambda_1)$. During each iteration, if $\widehat{\mathcal{Q}}(\lambda_3) > \widehat{\mathcal{Q}}(\lambda_4)$, then we move the points $\{\lambda_2, \widehat{\mathcal{Q}}(\lambda_2)\}$ to $\{\lambda_4, \widehat{\mathcal{Q}}(\lambda_4)\}$, and $\{\lambda_4, \widehat{\mathcal{Q}}(\lambda_4)\}$ to $\{\lambda_3, \widehat{\mathcal{Q}}(\lambda_3)\}$, and we update λ_3 to $\lambda_2 + \xi(\lambda_1 - \lambda_2)$. Otherwise, let $\{\lambda_1, \widehat{\mathcal{Q}}(\lambda_1)\}$ be $\{\lambda_3, \widehat{\mathcal{Q}}(\lambda_3)\}$, and $\{\lambda_3, \widehat{\mathcal{Q}}(\lambda_3)\}$ be $\{\lambda_4, \widehat{\mathcal{Q}}(\lambda_4)\}$, and we update λ_4 to $\lambda_1 + \xi(\lambda_2 - \lambda_1)$. Specifically, at each iteration, we first compute the values of $\{\widehat{\mathcal{Q}}_j(\lambda_i)\}_{i=1}^4$. This can be conducted efficiently since the dual problem (3.1) "splits" the Lagrangian minimization problem into d small dimensional nonconvex problems, and we can compute $\mathcal{Q}_j(\lambda_i)$'s in parallel. We call this a "splitting" step. Next, we centrally update λ_i's according to the golden section search method, which is a "communicating" step. Thus we call it a "splitting-communicating" (SPICA) algorithm, which is summarized in Algorithm 2. This algorithm finds a narrow interval that contains the optimal multiplier of the problem (3.1) after some iterations, and the output solution is the midpoint of the interval. It is well known that the golden section search method converges ξ-geometrically to the dual optimal solution $(\{\widehat{\boldsymbol{\beta}}_j\}_{j \in [d]}, \widehat{\lambda})$ (Bertsekas, 1999), i.e., we have

$$\widehat{\lambda}^{(t)} - \widehat{\lambda} \leq \xi^t |\lambda_2^{(0)} - \lambda_1^{(0)}|,$$

where $\lambda_i^{(0)}$'s denote the initial points, $\widehat{\lambda}^{(t)} = |\lambda_1^{(t)} - \lambda_2^{(t)}|/2$, and $(\lambda_1^{(t)}, \lambda_2^{(t)})$ denotes the corresponding point after t iterations.

The SPICA algorithm provides significant computational advantages. However, instead of a global optimal solution to problem (1.1), it generates an optimal solution to dual problem (3.1). Since the total cardinality constraint is nonconvex, there exists some duality gap between the dual

Algorithm 2 SPICA Algorithm

Input: $\lambda_1, \lambda_2 \in \mathbb{R}_+$, $\epsilon > 0$, $\xi = (-1+\sqrt{5})/2$.

Output: $\widehat{\lambda}$, $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$.

$\lambda_3 \leftarrow \lambda_2 + \xi(\lambda_1 - \lambda_2)$, $\lambda_4 \leftarrow \lambda_1 + \xi(\lambda_2 - \lambda_1)$.
while $|\lambda_1 - \lambda_2| > \epsilon$ **do**
  if $\mathcal{Q}(\lambda_3) - \lambda_3 K > \mathcal{Q}(\lambda_4) - \lambda_4 K$ **then**
    $\{\lambda_2, \mathcal{Q}(\lambda_2) - \lambda_2 K\} \leftarrow \{\lambda_4, \mathcal{Q}(\lambda_4) - \lambda_4 K\}$, $\{\lambda_4, \mathcal{Q}(\lambda_4) - \lambda_4 K\} \leftarrow \{\lambda_3, \mathcal{Q}(\lambda_3) - \lambda_3 K\}$.
    $\lambda_3 \leftarrow \lambda_2 + \xi(\lambda_1 - \lambda_2)$.
  else
    $\{\lambda_1, \mathcal{Q}(\lambda_1) - \lambda_1 K\} \leftarrow \{\lambda_3, \mathcal{Q}(\lambda_3) - \lambda_3 K\}$, $\{\lambda_3, \mathcal{Q}(\lambda_3) - \lambda_3 K\} \leftarrow \{\lambda_4, \mathcal{Q}(\lambda_4) - \lambda_4 K\}$.
    $\lambda_4 \leftarrow \lambda_1 + \xi(\lambda_2 - \lambda_1)$.
  end if
end while
$\widehat{\lambda} \leftarrow (\lambda_1 + \lambda_2)/2$, $\widehat{\boldsymbol{\beta}}_j = \operatorname{argmin}_{\boldsymbol{\beta}_j} \mathcal{L}_j(\boldsymbol{\beta}_j) + \widehat{\lambda}\|\boldsymbol{\beta}_j\|_0$ for all $j \in [d]$.

and the primal optimal solutions. In the next section, we illustrate a convexification phenomenon that, as d increases, the duality gap diminishes and does not impair any statistical loss in a wide range of problems.

3.2 The Convexification Phenomenon

Before rigorously proving that the dual optimal solution obtained by the SPICA algorithm is close to the primal optimal solution of problem (1.1), we illustrate some geometric intuition. The intuition traces back to some early convex geometry work, namely, the Shapley-Folkman Lemma (Starr, 1969). Consider the averaged Minkowski sum of d sets $\mathcal{A}_1$,...,$\mathcal{A}_d$ defined as $\{d^{-1}\sum_{j\in[d]} a_j : a_j \in \mathcal{A}_j \text{ for } j \in [d]\}$. The lemma reveals a geometric fact that the average of many nonconvex sets tends to be convex. In particular, letting $\rho(\mathcal{A})$ be a metric of the nonconvexity of the set $\mathcal{A}$, we have

$$\rho\Big(\frac{\mathcal{A}_1 + \mathcal{A}_2 + \cdots + \mathcal{A}_d}{d}\Big) \to 0, \text{ as } d \to \infty.$$

We provide an example to illustrate this convexification effect. Let the maximum distance between two sets $\mathcal{A}$ and $\mathcal{B}$ be $d(\mathcal{A}, \mathcal{B}) = \sup_{b\in\mathcal{B}} \inf_{a\in\mathcal{A}}\{\|a-b\| : a \in \mathcal{A}, b \in \mathcal{B}\}$. We measure the nonconvexity of a set $\mathcal{A}$ by the maximum distance between $\mathcal{A}$ and its convex hull. Since this distance is 0 if and only if $\mathcal{A}$ is convex, the maximum distance is a reasonable measure of how convex a set is. Considering the discrete set $\mathcal{A} = \{0, 1\}$ and its convex hull $\overline{\mathcal{A}} = [0, 1]$, we have $\rho(\mathcal{A}) = d(\mathcal{A}, \overline{\mathcal{A}}) = 1/2$. The maximum distance between the average of the Minkowski sum of two $\mathcal{A}$'s, which is $\mathcal{A}_2 = \{0, 1/2, 1\}$, and its convex hull is $\rho(\mathcal{A}_2) = d(\mathcal{A}_2, \overline{\mathcal{A}}) = 1/4$. Let the average of d $\mathcal{A}$'s be $\mathcal{A}_d$. We have $\rho(\mathcal{A}_d) = 1/2d$, which converges to 0 as d increases. We thus conclude that the average of d $\mathcal{A}$'s tend to be more convex as d increases. In Figure 2, we provide an geometric illustration of such increase of convexity, and we provide the mathematical description of Shapley-Folkman Lemma in Appendix C.

Let us return to problem (1.1). The duality gap between the primal problem (1.1) and its dual can be bounded by the nonconvexity of $d^{-1}\sum_{j\in[d]} \mathcal{A}_j$, where each $\mathcal{A}_j = \big\{\big(\mathcal{R}_j(\boldsymbol{\beta}_j), \mathcal{L}_j(\boldsymbol{\beta}_j)\big) : \boldsymbol{\beta}_j \in \mathcal{C}_j\big\}$

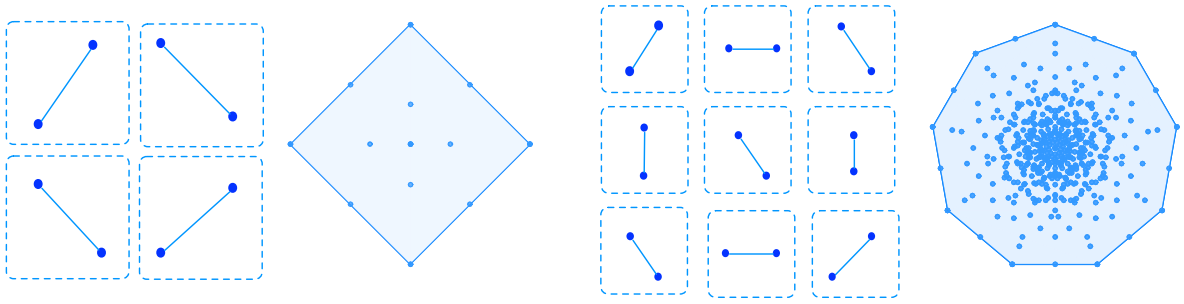


Figure 2: Left two: The shaded area of the second figure is the convex hull of the averaged Minkowski sum of four sets illustrated on the first figure. Each of the four sets contains two points, and the line between them represents the convex hull. Right two: The shaded area on the right is the convex hull of the averaged Minkowski sum of nine sets. The maximum distance between the averaged Minkowski sum and its convex hull decreases as the number of sets increases.

characterizes the joint nonconvexity of $\big(\mathcal{R}_j(\boldsymbol{\beta}_j), \mathcal{L}_j(\boldsymbol{\beta}_j)\big)$. By the intuition above, as d increases, the set $d^{-1}\sum_{j\in[d]}\mathcal{A}_j$ tends to be convex, and we expect a diminishing duality gap. This convexification phenomenon provides a hint that solving the dual problem might be as good as solving the (possibly) NP-complete primal problem.

4 Theoretical Justification of SPICA Algorithm

In this section, we provide theoretical justifications for the SPICA algorithm. We prove that the average-per-vertex duality gap diminishes as d increases. We analyze the statistical properties of the estimators computed by the SPICA algorithm. We also discuss the computational complexities of the dynamic programming approach and the SPICA algorithm in Appendix E.

4.1 Diminishing Duality Gap

In this subsection, we prove that the average-per-vertex duality gap diminishes at a rate of $\mathcal{O}(1/d)$. This result provides the theoretical justification that the estimator obtained by the SPICA algorithm (Alg. 2) is near-optimal, i.e., it is close to the primal optimal solution $\{\widetilde{\boldsymbol{\beta}}_j\}_{j\in[d]}$ of problem (1.1) defined as

$$\{\widetilde{\boldsymbol{\beta}}_j\}_{j=1}^d = \operatorname*{argmin}_{\boldsymbol{\beta}_j\in\mathcal{C}_j} \frac{1}{d}\sum_{j=1}^d \mathcal{L}_j(\boldsymbol{\beta}_j), \text{ subject to } \sum_{j=1}^d \|\boldsymbol{\beta}_j\|_0 \le K,$$

As we discussed in the previous section, we consider the Lagrangian dual problem (3.1), and the SPICA algorithm (Alg. 2) finds the dual optimal solution-multiplier pair $\big(\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}, \widehat{\lambda}\big)$ as defined in (3.2). Since the problem is nonconvex, strong duality does not hold. In this case, we only have weak duality that

$$\frac{1}{d}\sum_{j=1}^d \mathcal{L}_j(\widehat{\boldsymbol{\beta}}_j) + \widehat{\lambda}\Big(\sum_{j=1}^d \|\widehat{\boldsymbol{\beta}}_j\|_0 - K\Big) \le \frac{1}{d}\sum_{j=1}^d \mathcal{L}_j(\widetilde{\boldsymbol{\beta}}_j), \tag{4.1}$$

where both $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$ and $\{\widetilde{\boldsymbol{\beta}}_j\}_{j\in[d]}$ satisfy the total cardinality constraint, but some duality gap might exist in this case. Note that the duality gap is the difference between primal and dual optimal objective values. We provide an example to illustrate that the primal and dual optimal solutions $\widetilde{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}$ do not necessarily match, which results in a positive duality gap. Let

$$\mathbb{X} = \begin{pmatrix} 1 & 6 & 5 & 5 \\ 8 & 9 & 3 & 2 \\ 7 & 10 & 8 & 8 \end{pmatrix} \text{ and } \mathbf{y} = \begin{pmatrix} 12 \\ 20 \\ 25 \end{pmatrix}.$$

Considering the ℓ_0-constrained problem,

$$\min_{\boldsymbol{\beta}} \|\mathbb{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2, \text{ subject to } \|\boldsymbol{\beta}\|_0 \le 2,$$

the primal optimal solution is $\widetilde{\boldsymbol{\beta}} = (857/497, 0, 292/165, 0)^T$. For the dual solution,

$$\widehat{\boldsymbol{\beta}}(\lambda) = \|\mathbb{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_0.$$

it is not difficult to check that when $\lambda \ge 1169 - 1669/217$, $\widehat{\boldsymbol{\beta}}(\lambda) = \mathbf{0}$, if $\lambda \in [1669/434, 1669/217)$, $\widehat{\boldsymbol{\beta}}(\lambda) = (0, 502/217, 0, 0, 0)^T$, if $\lambda < 1669/434$, $\widehat{\boldsymbol{\beta}}(\lambda) = (1, 1, 1, 0)^T$. This implies that the primal and dual optimal solutions do not match, and there exists a strictly positive duality gap equals 449/894.

The next theorem proves that, as d increases, the average-per-vertex duality gap vanishes at the rate of $\mathcal{O}(1/d)$. This gives a strong evidence that the dual solution obtained by the SPICA algorithm is a fairly good approximation to the primal solution, especially when d is large. Usually, such a large d would cause the "curse of dimensionality" in nonconvex optimization, but our result reveals a "blessing of massive scale" phenomenon.

Theorem 4.1. The solution $\big(\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}, \widehat{\lambda}\big)$ obtained by the SPICA algorithm (Alg. 2) is a dual optimal solution-multiplier pair, which solves the dual problem (3.1), and satisfies

$$\frac{1}{d}\sum_{j=1}^{d} \mathcal{L}_j(\widehat{\boldsymbol{\beta}}_j) \le \frac{1}{d}\sum_{j=1}^{d} \mathcal{L}_j(\widetilde{\boldsymbol{\beta}}_j) + \frac{C_g}{d}, \text{ and } \sum_{j=1}^{d} \|\boldsymbol{\beta}_j\|_0 \le K, \tag{4.2}$$

where $\{\widetilde{\boldsymbol{\beta}}\}_{j\in[d]}$ is the primal optimal solution, and the constant C_g is

$$C_g = \max_{j\in[d]} \big|\mathcal{L}_j(\mathbf{0}) - \min_{\boldsymbol{\beta}_j\in\mathcal{C}_j} \mathcal{L}_j(\boldsymbol{\beta}_j)\big|. \tag{4.3}$$

Proof. First, we need the existence of the optimal dual solution. This is proved in Lemma D.1. Next, if $\widehat{\lambda} = 0$, we have $\widehat{\boldsymbol{\beta}}_j = \operatorname{argmin}_{\boldsymbol{\beta}_j} \mathcal{L}_j(\boldsymbol{\beta}_j)$ for all j's as defined in (3.2). Since $\sum_{j\in[d]} \|\widehat{\boldsymbol{\beta}}_j\|_0 \le K$ by the feasibility of $\widehat{\boldsymbol{\beta}}_j$'s, we have $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$ is also the primal optimal solution. This implies $\widehat{\boldsymbol{\beta}}_j = \widetilde{\boldsymbol{\beta}}_j$ for all j, and our claim follows as desired.

If $\widehat{\lambda} > 0$, we prove in Lemma D.3 that one of the two cases must hold:

(i) There exists a dual optimal solution-multiplier pair $(\widehat{\lambda}, \{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]})$, such that $\sum_{j\in[d]} \|\widehat{\boldsymbol{\beta}}_j\|_0 = K$.

(ii) Case (i) does not hold, and there exist at least two solutions achieve dual optimal objective, denoted as $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$ and $\{\widehat{\boldsymbol{\beta}}'_j\}_{j\in[d]}$, such that $\sum_{j\in[d]} \|\widehat{\boldsymbol{\beta}}_j\|_0 < K$ and $\sum_{j\in[d]} \|\widehat{\boldsymbol{\beta}}'_j\|_0 > K$.

Next, we consider the two cases separately. For case (i), there exists a dual optimal solution $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$ satisfying $\sum_{j\in[d]}\|\widehat{\boldsymbol{\beta}}_j\|_0 = K$. By the weak duality (4.1), we have

$$\frac{1}{d}\sum_{j=1}^{d}\mathcal{L}_j(\widehat{\boldsymbol{\beta}}_j) \le \frac{1}{d}\sum_{j=1}^{d}\mathcal{L}_j(\widetilde{\boldsymbol{\beta}}_j) + \widehat{\lambda}\big(\sum_{j=1}^{d}\|\widehat{\boldsymbol{\beta}}_j\|_0 - K\big) = \frac{1}{d}\sum_{j=1}^{d}\mathcal{L}_j(\widetilde{\boldsymbol{\beta}}_j),$$

where the first inequality holds by the definition of dual optimal solution that $\widehat{\boldsymbol{\beta}}_j = \operatorname{argmin}_{\boldsymbol{\beta}_j}\mathcal{L}_j(\boldsymbol{\beta}_j) + \widehat{\lambda}\|\boldsymbol{\beta}_j\|_0$, and the assertion of the theorem follows as desired. We also point out that since $\{\widetilde{\boldsymbol{\beta}}_j\}_{j\in[d]}$ is the primal optimal solution, the above inequality and the feasibility of $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$ guarantee the primal optimality of $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$, i.e., the dual optimal solution $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$ is also a primal optimal solution. This also leads to the certificate of primal optimality result stated in Corollary 4.2.

In the remaining proof, we focus our discussion on case (ii). This case is more complicated and requires more careful analysis due to the existence of multiple solutions. Recall that, given the multiplier $\widehat{\lambda}$, a dual solution is obtained by solving d subproblems of the ℓ_0-penalized form:

$$\min_{\boldsymbol{\beta}_j}\mathcal{L}_j(\boldsymbol{\beta}_j) + \widehat{\lambda}\|\boldsymbol{\beta}_j\|_0, \text{ for all } j = 1, ..., d.$$

Since there are multiple dual optimal solutions, as shown in Lemma D.3, we have that there is at least one j, such that the above ℓ_0-penalized optimization problem has multiple solutions, i.e., for some j, there exist $\widehat{\boldsymbol{\beta}}_j^{(1)}$ and $\widehat{\boldsymbol{\beta}}_j^{(2)}$ such that

$$\mathcal{L}_j(\widehat{\boldsymbol{\beta}}_j^{(1)}) + \widehat{\lambda}\|\widehat{\boldsymbol{\beta}}_j^{(1)}\|_0 = \mathcal{L}_j(\widehat{\boldsymbol{\beta}}_j^{(2)}) + \widehat{\lambda}\|\widehat{\boldsymbol{\beta}}_j^{(2)}\|_0. \tag{4.4}$$

In addition, any combination of the optimal solutions of the subproblems provides a dual optimal objective without satisfying the feasibility. In what follows, we show that we can select a dual optimal solution from all possible combinations, such that the selected solution achieves the error bound (4.2).

Suppose there exist m solutions achieve dual optimal objective. Let $\{\boldsymbol{\beta}_j^{(1)}\}_{j\in[d]}$, $\{\boldsymbol{\beta}_j^{(2)}\}_{j\in[d]}$,..., $\{\boldsymbol{\beta}_j^{(m)}\}_{j\in[d]}$ be the sequence of solutions ranked by their corresponding primal objective values, i.e.,

$$\sum_{j=1}^{d}\mathcal{L}_j(\boldsymbol{\beta}_j^{(1)}) \le \sum_{j=1}^{d}\mathcal{L}_j(\boldsymbol{\beta}_j^{(2)}) \le, ..., \le \sum_{j=1}^{d}\mathcal{L}_j(\boldsymbol{\beta}_j^{(m)}). \tag{4.5}$$

Meanwhile, by the dual optimality, we have,

$$\begin{aligned}\sum_{j=1}^{d}\Big[\mathcal{L}_j\{\boldsymbol{\beta}_j^{(1)}\} + \widehat{\lambda}\|\boldsymbol{\beta}_j^{(1)}\|_0\Big] &= \sum_{j=1}^{d}\Big[\mathcal{L}_j\{\boldsymbol{\beta}_j^{(2)}\} + \widehat{\lambda}\|\boldsymbol{\beta}_j^{(2)}\|_0\Big] \\ &=, ..., = \sum_{j=1}^{d}\Big[\mathcal{L}_j\{\boldsymbol{\beta}_j^{(m)}\} + \widehat{\lambda}\|\boldsymbol{\beta}_j^{(m)}\|_0\Big].\end{aligned}$$

Since $\widehat{\lambda} > 0$ by assumption, we have

$$\sum_{j=1}^{d}\|\boldsymbol{\beta}_j^{(1)}\|_0 \ge \sum_{j=1}^{d}\|\boldsymbol{\beta}_j^{(2)}\|_0 \ge, ..., \ge \sum_{j=1}^{d}\|\boldsymbol{\beta}_j^{(m)}\|_0.$$

Consequently, by the assumption that case (ii) holds, we have

$$\sum_{j=1}^{d} \|\boldsymbol{\beta}_j^{(1)}\|_0 > K > \sum_{j=1}^{d} \|\boldsymbol{\beta}_j^{(m)}\|_0.$$

To prove our claim, a key observation is that, for any $m' \in \{1, ..., m-1\}$, $\sum_{j\in[d]} \mathcal{L}_j\{\boldsymbol{\beta}_j^{(m'+1)}\} - \sum_{j\in[d]} \mathcal{L}_j\{\boldsymbol{\beta}_j^{(m')}\} \le C_g$, where C_g is defined in (4.3). This is proved in Lemma D.5.

Thus, by the assumption that case (ii) holds, there exist two consecutive solutions $\boldsymbol{\beta}_j^{(m_1)}$ and $\boldsymbol{\beta}_j^{(m_1+1)}$, such that

$$\Big|\sum_{j=1}^{d} \mathcal{L}_j\{\boldsymbol{\beta}_j^{(m_1)}\} - \sum_{j=1}^{d} \mathcal{L}_j\{\boldsymbol{\beta}_j^{(m_1+1)}\}\Big| \le C_g,$$
$$\text{and } \sum_{j=1}^{d} \|\boldsymbol{\beta}_j^{(m_1)}\|_0 > K > \sum_{j=1}^{d} \|\boldsymbol{\beta}_j^{(m_1+1)}\|_0.$$

In addition, by the dual optimality of the two solutions, it holds that

$$\begin{aligned}&\sum_{j=1}^{d} \mathcal{L}_j\{\boldsymbol{\beta}_j^{(m_1)}\} + \widehat{\lambda}\{\sum_{j=1}^{d} \|\boldsymbol{\beta}_j^{(m_1)}\|_0 - K\}\\ =&\sum_{j=1}^{d} \mathcal{L}_j\{\boldsymbol{\beta}_j^{(m_1+1)}\} + \widehat{\lambda}(\sum_{j=1}^{d} \|\boldsymbol{\beta}_j^{(m_1+1)}\|_0 - K).\end{aligned}$$

Consequently, as $\sum_{j\in[d]} \|\boldsymbol{\beta}_j^{(m_1)}\|_0 > K > \sum_{j\in[d]} \|\boldsymbol{\beta}_j^{(m_1+1)}\|_0$, and $\widehat{\lambda} > 0$, we further obtain that

$$\begin{aligned}0 \le -\widehat{\lambda}(\sum_{j=1}^{d} \|\boldsymbol{\beta}_j^{(m_1+1)}\|_0 - K) \le \widehat{\lambda}(\sum_{j=1}^{d} \|\boldsymbol{\beta}_j^{(m_1)}\|_0 - \sum_{j=1}^{d} \|\boldsymbol{\beta}_j^{(m_1+1)}\|_0)\\ = \sum_{j=1}^{d} \mathcal{L}_j\{\boldsymbol{\beta}_j^{(m_1+1)}\} - \sum_{j=1}^{d} \mathcal{L}_j\{\boldsymbol{\beta}_j^{(m_1)}\} \le C_g.\end{aligned} \tag{4.6}$$

We have

$$\sum_{j=1}^{d} \mathcal{L}_j\{\boldsymbol{\beta}_j^{(m_1)}\} \le \sum_{j=1}^{d} \mathcal{L}_j(\widetilde{\boldsymbol{\beta}}_j) - \widehat{\lambda}(\sum_{j=1}^{d} \|\boldsymbol{\beta}_j^{(m_1)}\|_0 - K) \le \sum_{j=1}^{d} \mathcal{L}_j(\widetilde{\boldsymbol{\beta}}_j) + C_g,$$

where the first inequality holds by the weak duality (4.1), and the second inequality holds by (4.6).

To conclude, in both cases (i) and (ii), we prove that there exists an dual optimal solution $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$ that achieves the total cardinality constraint, and approximates the primal solution within a constant error bound even if d increases. □

To interpret the constant C_g, each $\big|\mathcal{L}_j(\mathbf{0}) - \min_{\boldsymbol{\beta}_j\in\mathcal{C}_j} \mathcal{L}_j(\boldsymbol{\beta}_j)\big|$ is some "divergence" related to vertex j. It essentially measures the information gain by using neighboring vertices to explain uncertainties of vertex j. The constant C_g is the maximumal divergence among all vertices.

This result indicates that when the maximal divergence C_g is bounded, the average-per-vertex duality gap decreases to 0 as d increases. By the proof in Appendix D, the next corollary follows immediately, which provides a criterion to determine if the primal optimality holds for $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$.

Corollary 4.2. (Certificate for Primal Optimality) Let $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$ be the dual optimal solution for problem (1.1) obtained by the SPICA algorithm (Alg. 2). When the equality $\sum_{j\in[d]}\|\widehat{\boldsymbol{\beta}}_j\|_0 = K$ holds, it holds that the dual optimal solution also achieves the primal optimality, i.e.,

$$\sum_{j=1}^{d}\mathcal{L}_j(\widehat{\boldsymbol{\beta}}_j) = \sum_{j=1}^{d}\mathcal{L}_j(\widetilde{\boldsymbol{\beta}}_j), \text{ if } \sum_{j=1}^{d}\|\widehat{\boldsymbol{\beta}}_j\|_0 = K.$$

In the later simulation section, we find that empirically, the dual solution $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$ satisfies the certificate $\sum_{j\in[d]}\|\widehat{\boldsymbol{\beta}}_j\|_0 = K$ with high probability.

4.2 Statistical Rate

In this subsection, we provide the statistical guarantee that under weak assumptions, the duality gap does not sacrifice any statistical efficiency when d is large. This matches Theorem 4.1. We discuss the rates of convergence for the estimator provided by the SPICA algorithm under Gaussian and Ising graphical models. All technical proofs are provided in Appendix F.

4.2.1 Gaussian Graphical Model

We first apply the SPICA algorithm to estimate Gaussian graphical model. Consider a d-dimensional random vector $\boldsymbol{X} = (X_1, X_2, ..., X_d)^T \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. Under the Gaussian assumption, the conditional independence between X_j and X_k holds if and only if $\boldsymbol{\Theta}_{jk} = 0$, where $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$. Extensive literatures study this problem (Meinshausen and Bühlmann, 2006; Yuan and Lin, 2007; Cai et al., 2011; Liu and Wang, 2012). Under the spatial graphical modeling setting, taking a neighborhood pursuit approach, we formulate the graph estimation problem as

$$\min_{\{\boldsymbol{\beta}_j\}_{j\in[d]}}\frac{1}{dn}\sum_{j=1}^{d}\|\mathbb{X}_j - \mathbb{X}_{\mathcal{N}_j}\boldsymbol{\beta}_j\|_2^2, \text{ subject to } \sum_{j=1}^{d}\|\boldsymbol{\beta}_j\|_0 \le K,$$

where $\mathbb{X}\in\mathbb{R}^{n\times d}$ is the data matrix; $\mathbb{X}_{\mathcal{N}_j}\in\mathbb{R}^{n\times d_j}$ denotes the columns of $\mathbb{X}$ which correspond to the potential neighbors of X_j, and K is a pre-specified total cardinality. Given a solution $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$, we obtain the connected neighbors of each X_j by taking the corresponding nonzero components of $\widehat{\boldsymbol{\beta}}_j$. Consequently, we construct the graph estimator by either "OR" or "AND" rule on combining the neighborhoods for all X_j's. This approach is based on the fact that

$$\begin{aligned} X_j = \boldsymbol{X}_{\mathcal{N}_j}^T\boldsymbol{\beta}_j^* + \epsilon_j, \text{ where } \boldsymbol{\beta}_j^* = \left(\boldsymbol{\Sigma}_{\mathcal{N}_j,\mathcal{N}_j}\right)^{-1}\boldsymbol{\Sigma}_{\mathcal{N}_j,j}\in\mathbb{R}^{d-1}, \epsilon_j\sim N\left(0,\sigma_j^2\right), \\ \text{and } \sigma_j^2 = \boldsymbol{\Sigma}_{jj} - \boldsymbol{\Sigma}_{j,\mathcal{N}_j}\left(\boldsymbol{\Sigma}_{\mathcal{N}_j,\mathcal{N}_j}\right)^{-1}\boldsymbol{\Sigma}_{\mathcal{N}_j,j}, \end{aligned} \tag{4.7}$$

and by the block matrix inversion formula, it holds that

$$\boldsymbol{\Theta}_{jj} = \{\mathrm{Var}(\epsilon_j)\}^{-1} = \sigma_j^{-2}, \text{ and } \boldsymbol{\Theta}_{\mathcal{N}_j,j} = -\{\mathrm{Var}(\epsilon_j)\}^{-1}\boldsymbol{\beta}_j^* = -\sigma_j^{-2}\boldsymbol{\beta}_j^*.$$

Thus, $\boldsymbol{\Theta}_{jk} = 0$ if and only if the corresponding component of $\boldsymbol{\beta}_j^*$ is 0.

We point out that there are several advantages of the total cardinality approach over the ℓ_1 or other penalized approaches: (i) Imposing the total cardinality constraint directly handles the estimator's sparsity level. This provides a more intuitive approach than penalized methods, where tuning parameters do not give very interpretable meanings. (ii) Total cardinality constraint approach does not incur any estimation bias. In comparison, penalized approach induces some estimation biases. Although such biases are asymptotically negligible under appropriate scaling, the finite-sample behavior of the penalized approach is indeed outperformed by the total cardinality approach as demonstrated in later simulation studies.

Next, we analyze the statistical properties of the estimator obtained by the SPICA algorithm. As the neighborhood pursuit approach formulates the problem as a regression problem, we first bound the "prediction risk" of the estimator. This leads to the estimator's fast rate of convergence.

In the following discussion, for ease of presentation, we assume that the numbers of potential neighbors of the vertices are the same, i.e., $|\mathcal{N}_1| =, ..., = |\mathcal{N}_d| = d_0$. The next corollary of Theorem 4.1 guarantees that the average-per-vertex risk of our estimator converges at the minimax optimal rate, and justifies the vanishing gap does not incur statistical loss if the dimension d is large.

Corollary 4.3. Suppose that we have n independent samples of $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma}) \in \mathbb{R}^d$, and the spatial information that each vertex j can only connect to a set of vertices $\mathcal{N}_j \subset \{1, ..., d\}$ and $|\mathcal{N}_j| = d_0$. Let $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$ be the estimator obtained by the SPICA algorithm. Assume $s \leq K$, and $2K \leq dd_0$, where $\sum_{j\in[d]} \|\boldsymbol{\beta}_j^*\|_0 = s$, and $\boldsymbol{\beta}_j^*$'s are defined in (4.7). We further assume $\mathrm{diag}(\boldsymbol{\Sigma}) \leq \sigma^2$. Then, with probability at least $1 - \mathcal{O}(d^{-1})$, we have

$$\frac{1}{dn}\sum_{j=1}^{d} \|\mathbb{X}_{\mathcal{N}_j}\widehat{\boldsymbol{\beta}}_j - \mathbb{X}_{\mathcal{N}_j}\boldsymbol{\beta}_j^*\|_2^2 \leq C_1 \cdot \frac{K \log d}{dn} + C_2 \cdot \frac{\log d}{d}, \tag{4.8}$$

where C_1 and C_2 are two constants, and do not depend on K, d and n.

Proof. See Appendix F.1 for the detailed proof. □

This theorem proves that if $n < d$ and if the average-per-vertex degree is larger than 1, the estimator $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$ obtains the optimal rate of convergence. Note that this result does not require any restricted-eigenvalue type assumptions on $\mathbb{X}$. In comparison, it is shown in Zhang et al. (2014) that if we do not impose such assumptions, other estimators based on convex relaxations, such as the Lasso estimator, cannot achieve the optimal rate unless P = NP. In addition, if we impose the sparse eigenvalue condition that the minimum eigenvalue of the subcovariance matrices $\boldsymbol{\Sigma}_{\mathcal{N}_j,\mathcal{N}_j}$'s are all bounded below, i.e., there exists a constant $\rho > 0$, such that

$$\Lambda_{\min}\big(\boldsymbol{\Sigma}_{\mathcal{N}_j,\mathcal{N}_j}\big) > \rho, \text{ for all } j = 1, ..., d.$$

We have that the estimator $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$ obtains the fast rate of convergence that

$$\frac{1}{d}\sum_{j=1}^{d} \|\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\|_2^2 \leq \underbrace{C_1 \cdot \frac{K \log d}{dn}}_{\text{statistical error}} + \underbrace{C_2 \cdot \frac{\log d}{d}}_{\text{duality gap}},$$

where C_1 and C_2 are two constants, and do not depend on K, d and n. Note that if the certificate of primal optimality (Cor. 4.2) holds, the duality gap term disappears.

In graphical model estimation, support recovery is of significant practical interest. The next corollary provides the support recovery guarantee of the estimator.

Corollary 4.4. Assume that all the assumptions in Corollary 4.3 and the sparse eigenvalue condition hold and $K = s$, where $\sum_{j\in[d]} \|\boldsymbol{\beta}_j^*\|_0 = s$. Suppose that we have the minimal signal strength that for all j,

$$\|\boldsymbol{\beta}_j^*(\mathcal{S}_j)\|_{\min} > C \cdot \sqrt{\frac{\log d}{n}}, \tag{4.9}$$

where $\mathcal{S}_j$ denotes the support of $\boldsymbol{\beta}_j^*$, and $\boldsymbol{\beta}_j^*(\mathcal{S}_j)$ denotes the corresponding components of $\boldsymbol{\beta}_j^*$; $\|\mathbf{v}\|_{\min} = \min_j |v_j|$, and C is a constant which does not depend on K, d and n. We have, with probability at least $1 - \mathcal{O}(d^{-1})$,

$$\big|\mathrm{supp}\big(\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}\big) \cap \mathrm{supp}\big(\{\boldsymbol{\beta}_j^*\}_{j\in[d]}\big)\big| \geq s - d_0.$$

Proof. See Appendix F.2 for the detailed proof. □

This corollary proves that the SPICA algorithm almost exactly recovers the support of the graph with high probability. As d and s increase, if d_0 is fixed, the ratio between the number of correctly estimated support over the number of true support converges to 1 with high probability. Also, similar to the estimation results, if the certificate of primal optimality (Cor. 4.2) holds, we have the estimator exactly recovers the true support with high probability.

4.2.2 Ising Graphical Model

In this subsection, we consider the spatial Ising graphical model. Ising graphical model studies the conditional independences among random variables $X_j \in \{\pm 1\}$ for $j \in [d]$. Under Ising graphical model, the joint distribution of $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ is

$$\mathbb{P}(X_1 = x_1, ..., X_d = x_d) = \frac{1}{Z(\boldsymbol{\beta})} \exp\Big(\sum_{j\neq k} \frac{\beta_{jk} x_j x_k}{4}\Big),$$

where $Z(\boldsymbol{\beta})$ is some unknown partition function; each β_{jk} describes the interaction between vertex j and vertex k, and $\beta_{jk} = \beta_{kj}$.

Since the function $Z(\boldsymbol{\beta})$ is not given, directly estimating β_{jk}'s is not tractable. For the i-th observation $\mathbf{x}_i = (x_{i1}, ..., x_{id})^T \in \{\pm 1\}^d$, let $\theta_{ij} = \mathbb{P}(X_j = x_{ij} | \boldsymbol{X}_{\setminus j} = \mathbf{x}_{i,\setminus j})$ be the conditional distribution of the j-th vertex given others. Adopting the composite likelihood idea, we have

$$\theta_{ij} = \frac{\exp\big(\sum_{k:k\neq j} \beta_{jk} x_{ij} x_{ik}\big)}{\exp\big(\sum_{k:k\neq j} \beta_{jk} x_{ij} x_{ik}\big) + 1}.$$

We have that the negative conditional log-likelihood of the j-th vertex is

$$\mathcal{L}_j(\boldsymbol{\beta}_j) = -\frac{1}{n} \sum_{i=1}^{n} \log(\theta_{ij}),$$

Incorporating the spatial information, we have the prior information that $\boldsymbol{\beta}_{jk} = 0$ if $(j,k) \notin \mathcal{N}_j$ for each j, where $|\mathcal{N}_j| = d_0$. Adopting the total cardinality approach, we estimate $\boldsymbol{\beta}_j$'s by solving the following problem

$$\min_{\boldsymbol{\beta}_j} \frac{1}{d} \sum_{j=1}^{d} \mathcal{L}_j(\boldsymbol{\beta}_j), \text{ subject to } \sum_{j=1}^{d} \|\boldsymbol{\beta}_j\|_0 \leq K.$$

Next, we analyze the statistical properties of the estimators $\{\widehat{\boldsymbol{\beta}}_j\}_{j=1}^d$ obtained by the SPICA algorithm. We impose the following mild assumptions:

Assumption 4.5. Under Ising graphical model with parameters $\{\boldsymbol{\beta}_j^*\}_{j\in[d]}$, assume:

- (B.1) $\|\boldsymbol{\beta}_j^*\|_\infty \leq R$ for some $R \in (0, \infty)$.
- (B.2) The population Hessian matrix with respect to any subset $\mathcal{K} \subset \{1, ..., dd_0\}$, satisfies the local sparse eigenvalue condition that $\Lambda_{\min}\{\mathbb{E}[\nabla^2_{\mathcal{K}\mathcal{K}}\mathcal{L}(\boldsymbol{\beta}^*)]\} > 2\rho$, where $|\mathcal{K}| = K$, $\mathcal{L}(\boldsymbol{\beta}^*) = \sum_{j\in[d]} \mathcal{L}_j(\boldsymbol{\beta}_j^*)$, $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^{*T}, ..., \boldsymbol{\beta}_d^{*T})^T$, and $\rho > 0$ is a constant.

Note that assumption (B.1) is used in most literatures. For assumption (B.2), we only assume such a sparse eigenvalue condition at the point $\boldsymbol{\beta}^*$. This is essential for the identifiability of $\boldsymbol{\beta}^*$. Existing work (Ravikumar et al., 2010; Xue et al., 2012) imposes additional assumptions such as incoherence condition on the population Hessian matrix. Thus, our assumption is weaker than existing work. The next corollary of Theorem 4.1 provides the fast rate of convergence of the estimator obtained by the SPICA algorithm.

Corollary 4.6. Suppose that Assumption 4.5 holds, and assume that the samples are generated from a Ising model with parameters $\boldsymbol{\beta}_j^* \in \mathbb{R}^{d_0}$ for all j and $\sum_{j\in[d]} \|\boldsymbol{\beta}_j^*\|_0 = s \leq K$. We have, with probability at least $1 - \mathcal{O}(d^{-1})$,

$$\frac{1}{d}\sum_{j=1}^{d} \|\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\|_2^2 \leq \underbrace{C_1 \cdot \frac{K \log d}{dn}}_{\text{statistical error}} + \underbrace{C_2 \cdot \frac{1}{d}}_{\text{duality gap}},$$

where C_1 and C_2 are two constants, and do not depend on K, d and n.

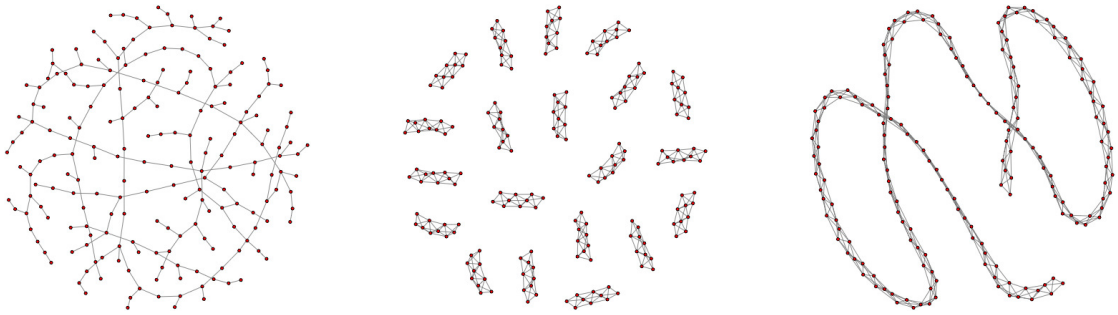
Proof. See Appendix F.3 for the detailed proof. □

5 Numerical Results

In this section, we conduct extensive numerical experiments to test the SPICA algorithm in comparison with ℓ_1-penalized method. We compare the parameter estimation and graph recovery performances of these two methods using both synthetic and real datasets. For ease of presentation, we provide the numerical performances under the Gaussian graphical model.

5.1 Synthetic Data

We first use synthetic data. We consider three different sets of parameters: (i) $n = 100$, $d = 1,000$; (ii) $n = 100$, $d = 2,000$; (iii) $n = 100$, $d = 5,000$, and we let the number of potential neighbors $d_0 = 10$. We further consider three different models for generating undirected graphs and precision matrices. Figure 3 illustrates sample graphs under these models. We repeat each setting for 100 times and report the averaged performance.



(a) Scale-Free (b) Block (c) Band

Figure 3: Examples of the three graph patterns we consider in the simulation study.

$*$ **Scale-free graph**. We generate the graph by the preferential attachment mechanism. We begin with a graph with a small chain of 2 vertices. At iteration j, we add a new vertex to the graph. The new vertex j connects to one of the previous d_0 vertices, with a probability which is proportional to the number of degrees of the existing vertex. Mathematically, let p_i be the probability that the new vertex j will connect to the existing vertex i is, $p_i = k_i / \sum_{i'=\min\{1,j-d_0\}}^{j-1} k_{i'}$, where k_i is the current degree of the vertex i. Thus, the resulting graph has $d-1$ edges. Given the graph, we generate the corresponding adjacency matrix $\mathbf{A}$ by setting the diagonal elements to be 0, and we set the nonzero off-diagonal elements to be $\rho = 0.1, 0.3$, or 0.5. Then, we construct the precision matrix $\mathbf{\Theta}$ as

$$\mathbf{\Theta} = \mathbf{D}\big[\mathbf{A} + \big\{\big|\Lambda_{\min}(\mathbf{A})\big| + 0.2\big\} \cdot \mathbf{I}_d\big]\mathbf{D}, \tag{5.1}$$

where $\Lambda_{\min}(\mathbf{A})$ denotes the smallest eigenvalue of $\mathbf{A}$; $\mathbf{I}_d$ denotes the identity matrix, and $\mathbf{D}$ is a diagonal matrix with $D_{jj} = 1$ for $j = 1, ..., d/2$ and $D_{jj} = 3$ for $j = d/2 + 1, ..., d$. Finally, we generate the multivariate Gaussian samples: $\mathbf{x}_1, ..., \mathbf{x}_n \sim N_d(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = \mathbf{\Theta}^{-1}$.

$*$ **Block graph**. We construct the precision matrix $\mathbf{A}$ as a block diagonal matrix. Each block is of the size 8. We set the nonzero off-diagonal entries to be ρ and diagonal entries to be 1. This matrix is positive definite. The graph has $3.5d$ edges, and we let the precision matrix be $\mathbf{\Theta} = \mathbf{DAD}$.

$*$ **Band graph**. Given d vertices indexed by $j = 1, ..., d$, we generate edges between the vertices whose corresponding coordinates are at distance less than or equal to 3. The resulting graph has $3d - 6$ edges. Given the graph, we construct the precision matrix same as (5.1).

We first consider the graph recovery performances of SPICA algorithm and the ℓ_1-penalized method. In particular, we evaluate the graph recovery performance by looking at the false positive and false negative rates. In particular, let $\widehat{G}^K = (V, \widehat{E}^K)$ be an estimated graph under the total cardinality constraint with tuning parameter K. The number of false positive discoveries using tuning parameter K is $\mathrm{FP}(K) = |\widehat{E}^K \backslash E|$, where $A \backslash B = \{a : a \in A \text{ and } a \notin B\}$, and the number of false negative discoveries with K is $\mathrm{FP}(K) = |E \backslash \widehat{E}^K|$. Consequently, we define the corresponding false positive rate (FPR) and the false negative rate (FNR) as

$$\mathrm{FPR}(K) = \frac{\mathrm{FP}(K)}{\binom{d}{2} - |E|} \text{ and } \mathrm{FNR}(K) = \frac{\mathrm{FN}(K)}{|E|}.$$

We plot the receiver operating characteristic (ROC) curves using $\big\{\mathrm{FNR}(K), 1 - \mathrm{FPR}(K)\big\}$ for the SPICA algorithm. Also, we plot the averaged ROC curves for the ℓ_1-penalized method for comparisons.

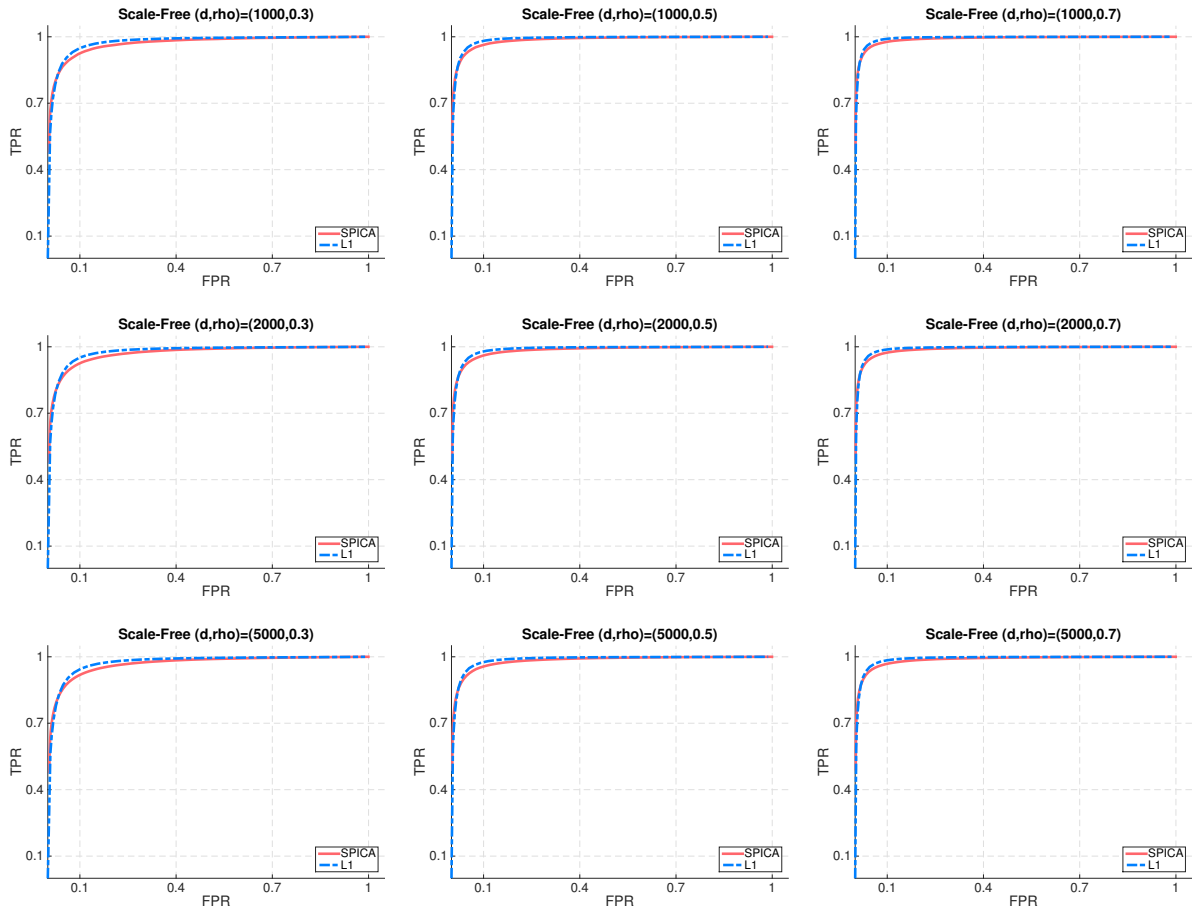


Figure 4: ROC curves for Scale-Free Model under different settings.

By Figures 4, 5 and 6, we see that the SPICA algorithm performs better than ℓ_1-penalized method under the block and band models, and the two methods perform similarly under the

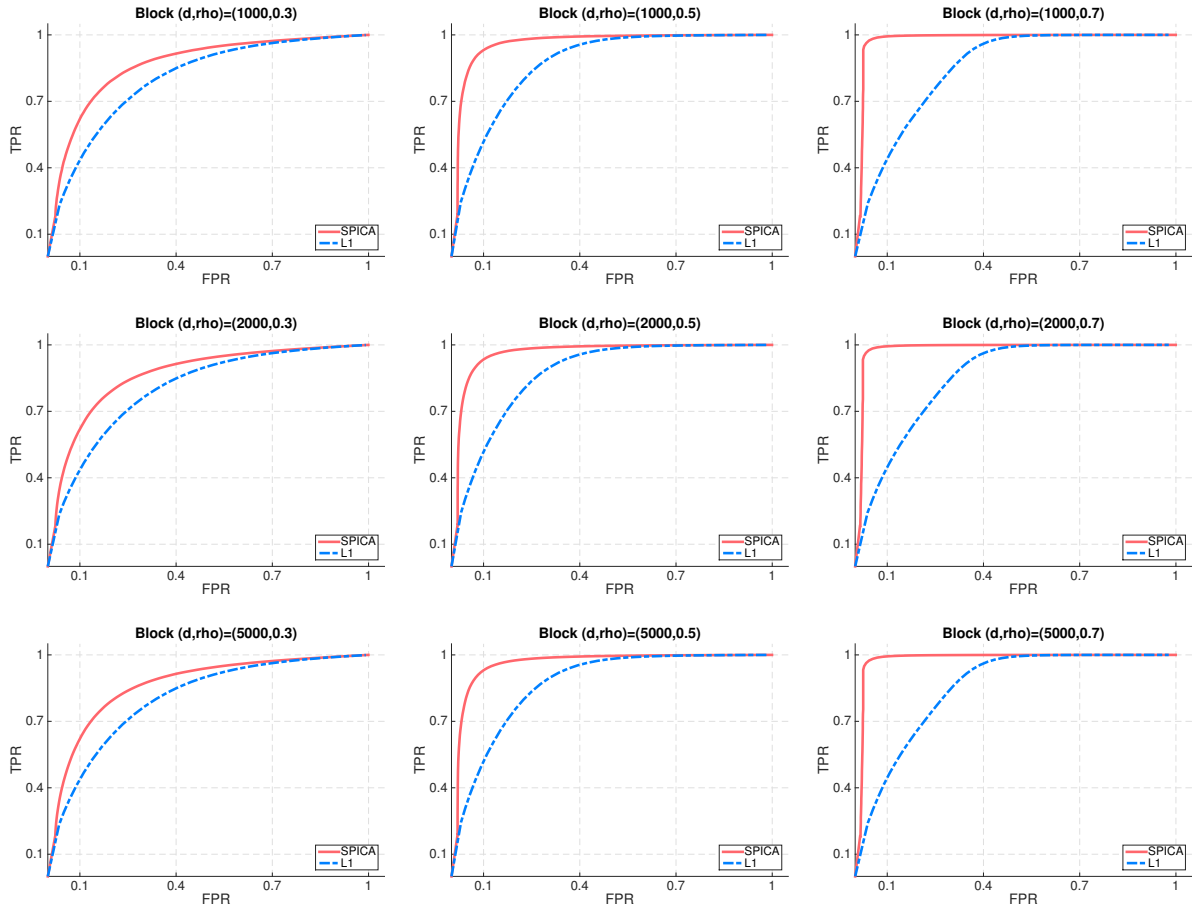


Figure 5: ROC curves for Block Model under different settings.

scale-free model. Thus, we conclude that the SPICA method works better than the ℓ_1-penalized method when the number of edges of the graph is larger. Also, for block and band models, we observe that the margin of the SPICA method over the ℓ_1-penalized method increases when ρ increases. This phenomenon has an intuitive explanation that the penalization term $\lambda \sum_j \|\boldsymbol{\beta}_j\|_1$ increases with the signal strength of $\boldsymbol{\beta}_j^*$'s, which induces more estimation bias, and results a worse performance in graph recovery.

We then compare the SPICA algorithm with the ℓ_1-penalized method from the perspective of parameter estimation. We select the tuning parameters by stability selection (Meinshausen and Bühlmann, 2010), and we report the error $\sum_{j=1}^d \|\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\|_2^2$ under all settings mentioned above. We observe that the SPICA algorithm performs better than the ℓ_1-penalized method as the degree or the signal strength increases. In addition, we observe some interesting phenomenon. In the scale-free and block models, the errors decrease as ρ increases for both methods. This is intuitive that as the increase of signal strength helps graph recovery, and consequently it also helps parameters estimation. In the band model, same as the scale-free and block models, the errors decrease as ρ increases for the SPICA algorithm. However, the errors increase as ρ increases for the ℓ_1-penalized method. This

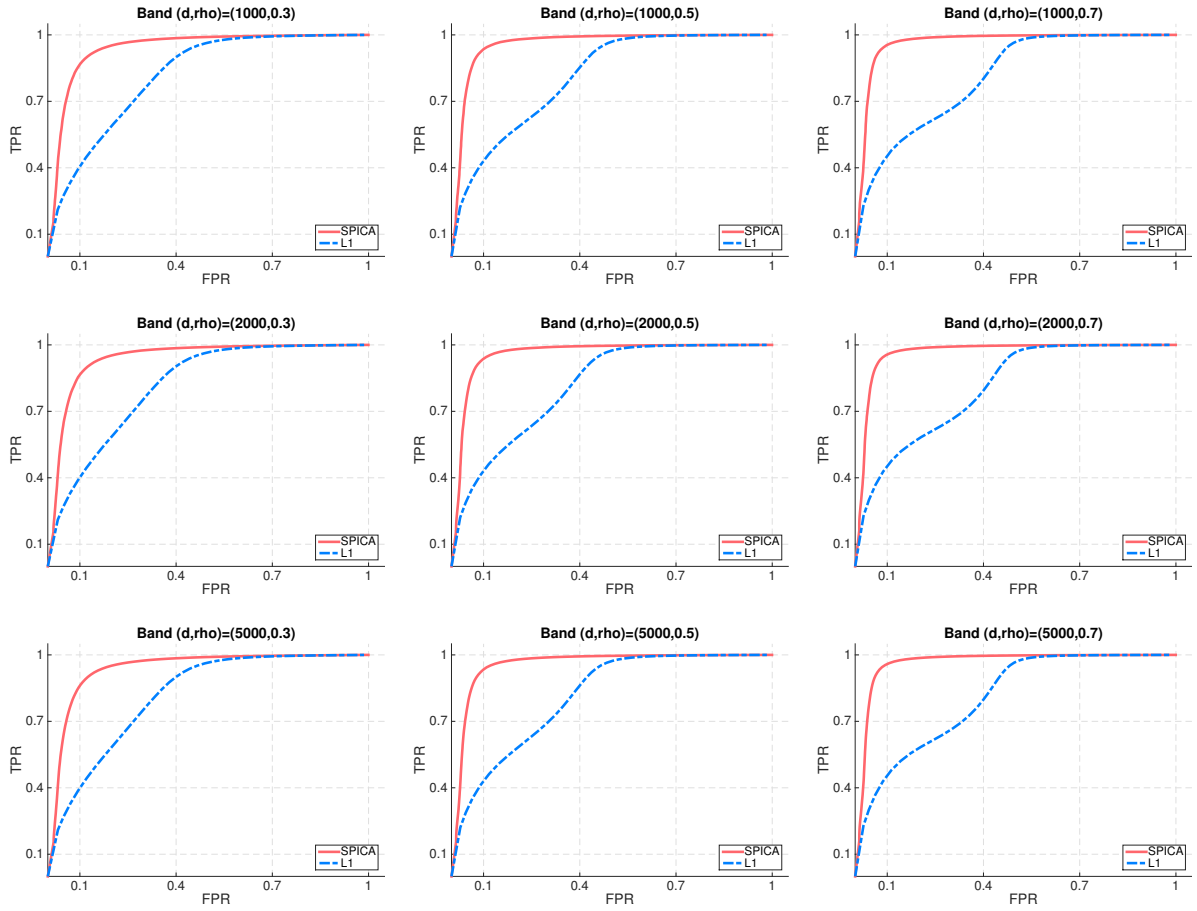


Figure 6: ROC curves for Band Model under different settings.

again confirms the intuition that as the ℓ_1-norm $\sum_{j\in[d]} \|\boldsymbol{\beta}_j^*\|_1$ increases, the penalization terms induces more biases. In comparison, the total cardinality constraint approach does not induce any biases.

To summarize, the superiority of SPICA over the ℓ_1-penalized method is well illustrated from the perspectives of both graph support recovery and parameters estimation. We also point out that the certificate of primal optimality, i.e., $\sum_{j\in[d]} \|\widehat{\boldsymbol{\beta}}_j\|_0 = K$, holds in more than 98% cases, which means that the SPICA algorithm generates the optimal solution to the problem with total cardinality constraint in these cases. This further shows the reliability of the SPICA algorithm.

5.2 Sensor Network Data

We also use wireless sensor network data to conduct tests. Our goal is to estimate how the sensors are connected. In practical applications, depending on the sensor type, the communication network of sensors might be known or unknown. In our data, the communication network is given. The reason we choose this type of data is that our primary goal is to evaluate the two different methods, and without such information, it is difficult to tell which method works better. In the implementation of

different methods, we do not use the information of how the sensors are connected, and we only use such information to evaluate the results at a later stage.

As discussed in the introduction, in a sensor network, each sensor can only connect with another if they are sufficiently close. Thus, estimating the network of sensors fits into the spatial graphical model framework. In our data, we have $d = 3,592$ sensors, and each sensor can only connect with another if they are within 3 meters. On average, each sensor has 24 potential neighbors. We have in total $n = 98$ samples. Each sample contains a signal strength of each sensor. Taking a Gaussian graphical model approach, we test the SPICA method and ℓ_1-penalized methods. We plot the ROC curves of the SPICA algorithm and the ℓ_1-penalized method in Figure 7 . It is clear that the SPICA method performs significantly better than the ℓ_1-penalized method. This shows that the SPICA method is capable of estimating spatial graphical models in practice.

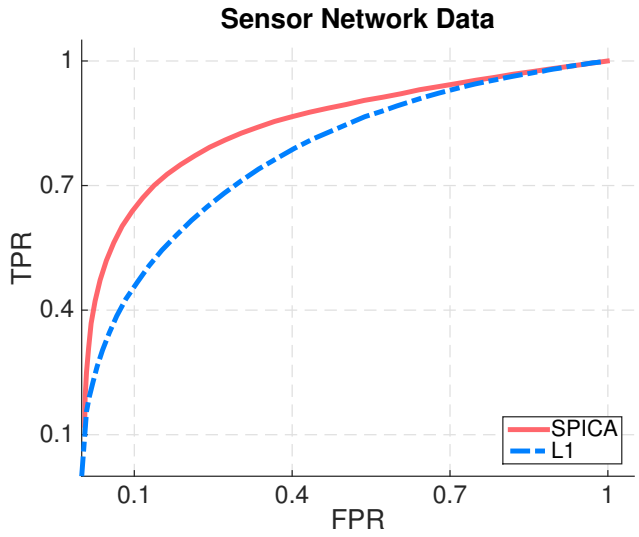


Figure 7: ROC curve for sensor network data.

6 Conclusion

To conclude, we provide several new fundamental results to better understand the total cardinality constraint approach for the spatial graphical model. We prove that problem (1.1) is NP-complete. We also show that for the case $\mathcal{R}_j(\boldsymbol{\beta}_j) = \|\boldsymbol{\beta}_j\|_0$, the problem is polynomial-time solvable. We further propose a more practical SPICA algorithm to solve the problem by considering the Lagrangian dual problem. Though the problem is nonconvex, we prove that the average-per-vertex duality gap decreases as the dimension d increases, and we achieve optimal statistical properties if the dimension is sufficiently large. We conduct thorough numerical experiments to backup our theory. For future work, we will continue to develop efficient algorithms to attack different cardinality constrained problems without sacrificing statistical efficiencies. In addition, we plan to apply the proposed method to conduct some real applications, such as brain functional region partition.

Table 1: Quantitative comparisons of the SPICA and ℓ_1-penalized method on different models. We report the averaged Frobenius norm $\sum_1^d \|\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\|_2^2$ with sample variance in the parenthes after repeating the simulation 100 times.

			$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.7$	
Model	n	d	SPICA	ℓ_1	SPICA	ℓ_1	SPICA	ℓ_1
Scale-Free	100	1,000	61.329	61.670	54.871	54.656	51.3893	51.7357
			(1.886)	(1.881)	(2.045)	(2.150)	(2.092)	(2.195)
		2,000	122.402	122.635	110.085	110.168	104.119	105.225
			(3.170)	(3.291)	(3.084)	(3.276)	(2.448)	(2.756)
		5,000	310.646	312.157	280.662	280.588	266.137	263.420
			(4.266)	(3.843)	(5.722)	(5.353)	(9.237)	(11.315)
Block	100	1,000	102.757	101.108	86.691	93.019	74.326	80.351
			(2.319)	(1.953)	(2.684)	(2.507)	(1.930)	(1.994)
		2,000	205.958	204.684	172.149	185.602	147.552	159.882
			(4.228)	(3.923)	(2.964)	(3.952)	(3.531)	(4.019)
		5,000	519.001	519.991	434.613	467.328	368.587	399.727
			(5.806)	(5.541)	(5.650)	(5.913)	(6.207)	(6.108)
Band	100	1,000	89.666	91.822	83.516	97.139	80.624	103.440
			(2.598)	(1.788)	(2.798)	(2.407)	(2.405)	(3.390)
		2,000	180.204	183.039	164.144	189.883	160.774	206.789
			(3.320)	(3.073)	(4.948)	(4.876)	(3.981)	(4.964)
		5,000	452.647	460.142	417.141	483.318	398.361	513.636
			(7.767)	(7.230)	(4.835)	(7.448)	(5.677)	(5.911)

Appendix

A Some Definitions in Computational Complexity

In this section, we introduce some basic definitions in the computational complexity theory. More detailed explanation and discussion can be found in iterature such as Arora and Barak (2009).

Definition A.1 (Class of P)**.** The complexity class P is defined as all problems that can be solved by a deterministic Turing machine using a polynomial time.

Definition A.2 (Class of NP)**.** The complexity class NP (Nondeterministic Polynomial time) is defined as all problems whose solutions can be verified by a deterministic Turing machine using a polynomial time.

Definition A.3 (Class of NP-hard)**.** A problem H is NP-hard if for any problem L in NP, there is a polynomial-time reduction from L to H, where a polynomial-time reduction is a method of

solving one problem by means of a hypothetical subroutine for solving a different problem, that uses polynomial time excluding the time within the subroutine.

Remark A.4. Roughly speaking, a polynomial-time reduction from L to H is that, given a problem L of size d, we can construct a problem H in a polynomial-time $\mathcal{O}(d^k)$ for some $k \in \mathbb{N}$, and solving the problem H will provide a solution to the original problem L. Consequently, finding an algorithm to solve any one of NP-hard problems within a polynomial-time would provide a universal algorithm to solve all the problems in NP.

Definition A.5 (Class of NP-Complete)**.** A problem H is NP-complete if it satisfies: (1) H is in the class of NP, and (2) For every problem L in NP, there exists a polynomial-time reduction from L to H.

B Proof of Theorem 2.2

Proof. To prove the claim, we construct a two way polynomial-time reduction (i) from a problem instance of knapsack problem to a problem instance of (2.3) and (ii) from a problem instance of (2.3) to a problem instance of knapsack problem.

(i) We first prove that given a knapsack problem instance (2.1) with input $(\mathbf{c}, \mathbf{b}, b_0)$, we can find a problem instance of the form (2.3), and by solving the new problem with a ϵ-optimal solution, we can recover the optimal solution of the knapsack problem. We consider the general case that we assume all the coefficients b_0, b_j's and c_j's are rational numbers. This assumption is general in the sense that computers can only take input as rational numbers. In addition, our analysis can be generalized to irrational case using numerical analysis techniques, see Trefethen and Bau III (1997) for example. To facilitate our discussion, denote by c the least common multiple of the denominators of b_0, all b_j's and c_j's. Since there are only finitely many feasible solutions to the knapsack problem, we have that there exists a positive gap δ between the optimal value and other feasible solutions' corresponding objective values. In addition, since we assume all the components are rational numbers, we have $\delta \geq 1/c$. Choosing $\epsilon = 1/2c$, by (A.1), within a polynomial time, we can construct $\mathbb{X}_j$, $\mathbb{Y}_j$, such that

$$\mathcal{L}(0; \mathbb{X}_j, \mathbb{Y}_j) = 0, \text{ and } \min_{\beta_j} \mathcal{L}(\beta_j; \mathbb{X}_j, \mathbb{Y}_j) = -c_j - \epsilon/d, \text{ for all } j.$$

Here we let $\mathcal{L}(0; \mathbb{X}_j, \mathbb{Y}_j) = 0$ instead of c_0 in (A.1) for ease of presentation, which does not lose generality. We essentially need to construct $\mathbb{X}_j$'s and $\mathbb{Y}_j$'s such that the difference between $\mathcal{L}(0, \mathbb{X}_j, \mathbb{Y}_j)$ and $\min_{\beta_j} \mathcal{L}(\beta_j; \mathbb{X}_j, \mathbb{Y}_j)$ is $-c_j - \epsilon/d$ for all j, and all $\mathcal{L}(0, \mathbb{X}_j, \mathbb{Y}_j)$ are identical. In addition, we let $\mathcal{R}_j(\beta_j) = b_j \|\beta_j\|_0$. This constructs an instance of problem (2.3), and the optimal solution to this instance of problem (2.3) lies in a compact set that $\beta_j^* \in [-r, r]$ for all j, where $r = \max_j \|\beta_j'\|_\infty$, and $\beta_j' = \operatorname{argmin}_{\beta_j} \mathcal{L}(\beta_j; \mathbb{X}_j, \mathbb{Y}_j)$. Also, since $\mathcal{L}$ is convex, there exists a constant $g > 0$, such that $|\mathcal{L}(\beta_j; \mathbb{X}_j, \mathbb{Y}_j) - \mathcal{L}(\widehat{\beta}_j; \mathbb{X}_j, \mathbb{Y}_j)| \leq g\|\beta_j - \widehat{\beta}_j\|_2$ for any $\beta_j, \widehat{\beta}_j \in [-r, r]$ for all j.

Denote by f_K^* and $-f_L^*$ the optimal objective values of the knapsack problem and the new problem. First, we show that $f_L^* \geq f_K^* + \epsilon$. Denote by $\mathbf{x}^*$ an optimal solution to the multiple-choice

knapsack problem. Looking at the solution

$$\widehat{\beta}_j = \begin{cases} \text{argmin}_{\beta_j} \mathcal{L}(\beta_j; \mathbb{X}_j, \mathbb{Y}_j), & \text{if } x_j^* = 1, \\ 0 & \text{otherwise,} \end{cases}$$

we have $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$ is a feasible solution to the instance of problem (2.3), and the corresponding objective value equals $-f_K^* - \epsilon$. Thus, we conclude $f_L^* \geq f_K^* + \epsilon$. Consequently, letting an ϵ-optimal solution's corresponding value be $-f_L'$, we have $f_L' \geq f_K^*$.

Next, we show that given an ϵ-optimal solution $\{\beta_j^*\}_{j\in[d]}$ to the instance of problem (2.3), we have that there exists an optimal solution $\mathbf{x}^*$ to the knapsack problem that $x_j^* = 1$ if $\beta_j^* \neq 0$, and $x_j^* = 0$ otherwise. Then, we can recover an optimal solution $\mathbf{x}^*$ to the knapsack problem by setting $x_j^* = 1$ if $\beta_j^* \neq 0$ and 0 otherwise.

For any feasible solution $\{\beta_j\}_{j\in[d]}$, suppose that there does not exist an optimal solution $\mathbf{x}^*$ to the multiple-choice knapsack problem such that $\mathcal{R}_j(\beta_j) > 0$ for $x_j^* = 1$ and $\mathcal{R}_j(\beta_j) = 0$ for $x_j^* = 0$. We have $\{\beta_j\}_{j\in[d]}$'s corresponding objective value is upper bounded by $\sum_{j:\beta_j\neq 0} c_j + \epsilon + \sum_{j:\beta_j=0} \mathcal{L}(0; \mathbb{X}_j, \mathbb{Y}_j)$, where the term $\sum_{j:\beta_j\neq 0} c_j$ is upper bounded by $f_K^* - \delta$, and the term $\sum_{j:\beta_j=0} \mathcal{L}(0; \mathbb{X}_j, \mathbb{Y}_j)$ is 0 by construction. By our choices of ϵ, the objective is no greater than f_K^*. This proves $\{\beta_j\}_{j\in[d]}$ cannot be an ϵ-optimal solution to the problem (2.3). Meanwhile, for any ϵ-optimal solution $\{\beta_j^*\}_{j\in[d]}$, we have its corresponding objective value is lower-bounded by $\sum_{j:\beta_j^*\neq 0} c_j + \epsilon + \sum_{j:\beta_j=0} \mathcal{L}(0; \mathbb{X}_j, \mathbb{Y}_j) \geq f_K^*$. Since the term $\sum_{j:\beta_j=0} \mathcal{L}(0; \mathbb{X}_j, \mathbb{Y}_j)$ is 0, we have the term $\sum_{j:\beta_j^*\neq 0} c_j + \epsilon$ is strictly larger than $f_K^* - \delta$, which implies that $\sum_{j:\beta_j^*\neq 0} c_j$ is strictly larger than $f_K^* - \delta$ since $\epsilon < \delta$ and $\sum_{j:\beta_j^*\neq 0} c_j$ can by difference at least δ. Meanwhile, we have $\sum_{j:\beta_j^*\neq 0} b_j \leq b_0$. This proves the feasibility of $\mathbf{x}^*$. Thus, we have $\mathbf{x}^*$ is feasible, and its corresponding objective value is strictly greater than $f_K^* - \delta$, which implies its corresponding objective can only be the optimal value f_K^*. We have that an ϵ-optimal solution $\{\beta^*\}_{j\in[d]}$ to the instance of problem (2.3) recovers an optimal solution to the knapsack problem by setting $x_j^* = 1$ if $\beta_j^* \neq 0$, and $x_j^* = 0$ otherwise.

(ii) For the other direction, we first introduce the multiple-choice knapsack problem for ease of presentation. Denote by $\mathbf{C} = (\mathbf{c}_1, ..., \mathbf{c}_d)^T \in \mathbb{R}^{d\times d_0}$ and $\mathbf{B} = (\mathbf{b}_1, ..., \mathbf{b}_d)^T \in \mathbb{R}^{d\times d_0}$, where $\mathbf{c}_j = (c_{j1}, ..., c_{jd_0})^T, \mathbf{b}_j = (b_{j1}, ..., b_{jd_0})^T$. Consider the multiple-choice knapsack problem with input $(\mathbf{C}, \mathbf{B}, b_0)$, where all the coefficients are positive rational numbers:

$$\max_{x_{jk}} \sum_{j=1}^{d}\sum_{k=1}^{d_0} c_{jk}x_{jk}, \text{ subject to } \sum_{j=1}^{d}\sum_{k=1}^{d_0} b_{jk}x_{jk} \leq b_0, \sum_{k=1}^{d_0} x_{jk} \leq 1, \ x_{jk} \in \{0, 1\},$$

for all $j \in [d]$ and all $k \in [d_0]$. It is well known that there exists a two-way polynomial-time reduction between the 0-1 knapsack problem (2.1) and the multiple-choice knapsack problem (Williamson and Shmoys, 2011). Thus, given a problem of the form (2.1), we only need to find a polynomial time reduction to a problem instance of the multiple-choice knapsack problem.

Without loss of generality, we assume all c_{jk}, b_{jk}'s and b_0 are rational numbers. Given an instance of problem (2.3) with input $(\{\mathbb{X}_j, \mathbb{Y}_j\}_{j\in[d]}, b_0)$, the solution belongs to a bounded region that $\boldsymbol{\beta}_j^* \in [-r, r]^{d_0}$ for all j. Since $\mathcal{L}$ is convex, we have that there exists a constant $g > 0$, such that $|\mathcal{L}(\boldsymbol{\beta}_j; \mathbb{X}_j, \mathbb{Y}_j) - \mathcal{L}(\widehat{\boldsymbol{\beta}}_j; \mathbb{X}_j, \mathbb{Y}_j)| \leq g\|\boldsymbol{\beta}_j - \widehat{\boldsymbol{\beta}}_j\|_2$ for any $\boldsymbol{\beta}_j, \widehat{\boldsymbol{\beta}}_j \in [-r, r]^{d_0}$ for all j. For each j,

we first discretize $[-r,r]^{d_0}$ into a set of d' points $\{\mathbf{p}_1^{(j)},...,\mathbf{p}_{d'}^{(j)}\}$ denoted by $\mathcal{B}_j$, where we discretize the set sufficiently finely that for any $\boldsymbol{\beta}_j \in \mathcal{B}_j$, there exists a $\mathbf{p}^{(j)} \in \mathcal{B}_j$ such that $\|\mathbf{p}^{(j)} - \boldsymbol{\beta}_j\|_\infty < \delta'$ and $\mathcal{R}(\mathbf{p}^{(j)}) \leq \mathcal{R}(\boldsymbol{\beta}_j)$, where $\delta' < (dg)^{-1}\epsilon$. By (A.2), this can be done within a polynomial time by assumption (A.2). We further compute corresponding objective and constraint values for all points $\mathbf{p}_k^{(j)} \in \mathcal{B}_j$. Specifically, for any $\mathbf{p}_k^{(j)} \in \mathcal{B}_j$, we compute $b'_{jk} = \mathcal{R}_j(\mathbf{p}_k^{(j)})$ and $c'_{jk} = -\mathcal{L}(\mathbf{p}_k^{(j)}; \mathbb{X}_j, \mathbb{Y}_j)$. Also, letting $b'_0 = b_0$, we have an instance of multiple-choice knapsack problem with input $(\mathbf{C}', \mathbf{B}', b'_0)$. It is not difficult to see that an optimal solution of the multiple-choice knapsack problem $\mathbf{x}^*$ gives a feasible solution of the original problem (2.3) by assigning $\widehat{\boldsymbol{\beta}}_j = \mathbf{p}_k^{(j)}$ if $x^*_{jk} = 1$. Next, consider an optimal solution $\{\boldsymbol{\beta}_j^*\}_{j=1}^d$ for the problem (2.3), we have, by our discretization, there exists some feasible point $\{\widehat{\mathbf{p}}^{(j)}\}_{j\in[d]}$ belongs to the discretized set, and $\|\widehat{\mathbf{p}}^{(j)} - \boldsymbol{\beta}_j^*\|_2 < \delta'$ for all j. Consequently, we have $\{\widehat{\mathbf{p}}^{(j)}\}$'s corresponding objective is lower bounded by $f_L^* - dg\delta' \geq f_L^* - \epsilon$. Thus, we have that $\{\widehat{\mathbf{p}}^{(j)}\}_{j\in[d]}$ is an ϵ-optimal solution to the problem (2.3).This finishes our proof. □

C Shapley-Folkman Lemma

In this section, we provide the mathematical details of Shapley-Folkman Lemma. This lemma studies the geometry of Minkowski sum of sets in vector space. The Minkowski sum of sets is defined as follows.

Definition C.1. The Minkowski sum of sets of vectors is formed by summing one vector of each set, i.e., the Minkowski sum of sets $\mathcal{A}_i$'s, for $i = 1, ..., I$ is

$$\mathcal{A}_1 \oplus \mathcal{A}_2, ..., \oplus \mathcal{A}_I = \Big\{ \sum_{i=1}^{I} a_i : a_i \in \mathcal{A}_i \text{ for all } i \Big\}.$$

Shapley-Folkman Lemma is formally stated as follows.

Theorem C.2 (Shapley-Folkman Lemma)**.** Let $\{\mathcal{A}_i\}_{i=1}^I$ be a collection of sets of $\mathbb{R}^k$. Let a belongs the convex hull of the Minkowski sum $\oplus_{i=1}^I \mathcal{A}_i$, i.e., $a \in \text{Conv}\big(\oplus_{i=1}^I \mathcal{A}_i\big)$. If $k < I$, we have, a belongs to the sum of convex hulls of k sets and the Minkowski sum of the rest sets, i.e.,

$$a \in \sum_{1\leq i\leq k} \text{Conv}(\mathcal{A}_i^a) + \mathcal{A}_{k+1}^a \oplus \mathcal{A}_{k+2}^a, ..., \oplus \mathcal{A}_I^a,$$

where $\{\mathcal{A}_i^a\}_{i=1}^I$ is some re-indexing of $\{\mathcal{A}_i\}_{i=1}^I$ depends on the point a.

This Lemma is used to prove the following theorem which bounds the distance between a Minkowski sum and its convex hull.

Theorem C.3. Let $\{\mathcal{A}_i\}_{i=1}^I$ be a collection of sets of $\mathbb{R}^k$. If $I \geq k$, for any point in the convex hull of a Minkowski sum, $\text{Conv}(\sum_{i=1}^I \mathcal{A}_i)$, its distance to the Minkowski sum $\oplus_{i=1}^I \mathcal{A}_i$ is upper bounded by the sum of the squares of the squares of the k largest circumradii of the sets $\mathcal{A}_i$'s, where the circumradii of a set is defined as the radii of the smallest sphere in $\mathbb{R}^k$ enclosing the set.

We point out the the above theorem is independent of the number I as long as $I > k$ holds. In the constraint of our problem (1.1), the dimension of each set is small, and the number of sets d is very large. The above theorem provides the intuition behind the decrease of the distance between the feasible set and its convex hull. This further results the decrease of duality gap, and the distance is upper bounded in Section 4.

D Lemmas for Proving Theorem 4.1

In this section, we provide lemmas used in the proof of Theorem 4.1.

Lemma D.1. Assume that there exists a feasible primal optimal solution to the problem (1.1). Then, the dual optimal solution $\big(\widehat{\lambda}, \{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}\big)$ also exists, where the dual optimal solution is defined as in (3.2).

Proof. We first prove the existence of the dual optimal Lagrangian multiplier $\widehat{\lambda}$. By the definition of $\sum_{j\in[d]} \mathcal{Q}_j(\lambda) - \lambda K$ in (3.1), we have $\sum_{j\in[d]} \mathcal{Q}_j(\lambda) - \lambda K$ is the supreme of a collection of sum of linear functions. This implies that $\sum_{j\in[d]} \mathcal{Q}_j(\lambda) - \lambda K$ is a concave function. Also, we observe that if $\lambda \to \infty$, then $\sum_{j\in[d]} \mathcal{Q}_j(\lambda) - \lambda K \to -\infty$ as shown in Lemma D.2. Thus, $\sum_{j\in[d]} \mathcal{Q}_j(\lambda) - \lambda K$ has compact level sets due to its concavity, i.e., for any $\alpha \in \mathbb{R}$, the set $\{\lambda | \sum_{j\in[d]} \mathcal{Q}_j(\lambda) - \lambda K \geq \alpha\}$ is compact. By Bolzano-Weiestrass Theorem, there exits at least one optimal Lagrangian multiplier $\widehat{\lambda}$ that maximizes $\sum_{j\in[d]} \mathcal{Q}_j(\lambda) - \lambda K$ over $\lambda \geq 0$.

Next, we prove the existence of $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$. Given the optimal Lagrangian multiplier $\widehat{\lambda}$, for each j, the dual solution $\widehat{\boldsymbol{\beta}}_j$ is

$$\widehat{\boldsymbol{\beta}}_j = \operatorname*{argmin}_{\boldsymbol{\beta}_j} \mathcal{L}_j(\boldsymbol{\beta}_j) + \widehat{\lambda}\|\boldsymbol{\beta}_j\|_0, \text{ or equivalently, } \widehat{\boldsymbol{\beta}}_j = \operatorname*{argmin}_{\widehat{\boldsymbol{\beta}}_j(k)} \mathcal{L}_j\big\{\boldsymbol{\beta}_j(k)\big\} + \widehat{\lambda}k,$$

where $\widehat{\boldsymbol{\beta}}_j(k_j) = \operatorname{argmin}_{\|\boldsymbol{\beta}_j\|_0 \leq k_j} \mathcal{L}_j(\boldsymbol{\beta}_j)$. Since the primal solution exists, all $\mathcal{L}_j$'s are bounded below. Thus each subproblem j can be solved by branch-and-bound for each $k \in [d_0]$, and the solution $\widehat{\boldsymbol{\beta}}_j$ is obtained by minimizing the objective above over k. Therefore there exists at least one dual optimal solution $\widehat{\boldsymbol{\beta}}_j$ for each j. □

Lemma D.2. Consider the Lagrangian dual $\sum_{j\in[d]} \mathcal{Q}_j(\lambda) - \lambda K$ of the problem (1.1). If a primal optimal solution $\{\widetilde{\boldsymbol{\beta}}_j\}_{j=1}^d$ exists for the primal problem, we have $\sum_{j\in[d]} \mathcal{Q}_j(\lambda) - \lambda K \to \infty$ as $\lambda \to \infty$.

Proof. We consider each $\mathcal{Q}_j(z)$ separately. For each j, denote $\widehat{\boldsymbol{\beta}}_j(k)$ as the optimal solution for the problem

$$\widetilde{\boldsymbol{\beta}}_j(k) = \operatorname*{argmin}_{\boldsymbol{\beta}_j} \mathcal{L}_j(\boldsymbol{\beta}_j), \text{ subject to } \|\boldsymbol{\beta}_j\|_0 = k.$$

Since the primal solution exits, it is not difficult to check that, if $\lambda > \mathcal{L}_j\big\{\boldsymbol{\beta}_j(k)\big\} - \mathcal{L}_j\big\{\widetilde{\boldsymbol{\beta}}_j(k+1)\big\}$ for all $k = 0, ..., d_j - 1$, we have $\mathbf{0} = \operatorname{argmin}_{\boldsymbol{\beta}_j} \mathcal{L}_j(\boldsymbol{\beta}_j) + \lambda\|\boldsymbol{\beta}_j\|_0$, which implies $\mathcal{Q}_j(\lambda) = \mathcal{L}_j(\mathbf{0})$. Thus, when λ is large, $\sum_{j\in[d]} \mathcal{Q}_j(\lambda) - \lambda K = \sum_{j\in[d]} \mathcal{L}_j(\mathbf{0}) - \lambda K$, which decreases linearly. Since $K > 0$, our claim follows as desired. □

Lemma D.3. For the dual problem (3.2), suppose that the optimal Lagrangian multiplier $\widehat{\lambda} > 0$. One of the following two cases holds:

(i) There exists a dual optimal solution-multiplier pair $(\widehat{\lambda}, \{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]})$, where $\sum_{j\in[d]} \|\widehat{\boldsymbol{\beta}}_j\|_0 = K$.

(ii) There exists at least two solutions achieve optimal dual objective, denoted as $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$ and $\{\widehat{\boldsymbol{\beta}}'_j\}_{j\in[d]}$, such that $\sum_{j=1}^d \|\widehat{\boldsymbol{\beta}}_j\|_0 < K$ and $\sum_{j=1}^d \|\widehat{\boldsymbol{\beta}}'_j\|_0 > K$.

Proof. We prove the lemma by contradiction. Assume to the contrary that either $\sum_{j\in[d]} \|\widehat{\boldsymbol{\beta}}_j\|_0 > K$ for all solutions $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$ achieve dual optimal objective (for the case $\sum_{j\in[d]} \|\widehat{\boldsymbol{\beta}}_j\|_0 < K$, the claim follows by similar arguments.). We have

$$\begin{aligned}\sum_{j=1}^d \mathcal{Q}_j(\widehat{\lambda}) - \widehat{\lambda}K &= \min_{\boldsymbol{\beta}_j} \Big\{ \sum_{j=1}^d \mathcal{L}_j(\boldsymbol{\beta}_j) + \widehat{\lambda}\big(\sum_{j=1}^d \|\boldsymbol{\beta}_j\|_0 - K\big)\Big\} \\ &= \min_{k_j\in\{1,\ldots,d_0\}} \Big\{ \sum_{j=1}^d \mathcal{L}_j\big\{\widehat{\boldsymbol{\beta}}_j(k_j)\big\} + \widehat{\lambda}\big(\sum_{j=1}^d k_j - K\big)\Big\},\end{aligned}$$

where $\widehat{\boldsymbol{\beta}}_j(k_j) = \operatorname{argmin}_{\|\boldsymbol{\beta}_j\|_0\le k_j} \mathcal{L}_j(\boldsymbol{\beta}_j)$.

By the assumption, it holds that $\sum_{j\in[d]} k_j$ is strictly greater than K. As shown in Lemma D.4, we have that for small $\epsilon > 0$ (or $\epsilon < 0$ if $\sum_{j\in[d]} k_j < K$), we obtain

$$\begin{aligned}\sum_{j=1}^d \mathcal{Q}_j(\widehat{\lambda}+\epsilon) - (\widehat{\lambda}+\epsilon)K &= \min_{k_j=1,\ldots,d} \Big[\sum_{j=1}^d \mathcal{L}_j\big\{\widehat{\boldsymbol{\beta}}_j(k_j)\big\} + (\widehat{\lambda}+\epsilon)\big(\sum_{j=1}^d k_j - K\big)\Big] \\ &> \min_{k_j=1,\ldots,d} \Big[\sum_{j=1}^d \mathcal{L}_j\big\{\widehat{\boldsymbol{\beta}}_j(k_j)\big\} + \widehat{\lambda}\big(\sum_{j=1}^d k_j - K\big)\Big] \\ &= \sum_{j=1}^d \mathcal{Q}_j(\widehat{\lambda}) - \widehat{\lambda}K.\end{aligned}$$

Meanwhile, since $\widehat{\lambda}$ is the optimal Lagrangian multiplier by assumption, it maximizes the function $\sum_{j\in[d]} \mathcal{Q}_j(\lambda) - K$. The above result yields a contradiction. □

Lemma D.4. Suppose that the dual optimal Lagrangian multiplier $\widehat{\lambda} > 0$. Let $k_j = \|\widehat{\boldsymbol{\beta}}_j\|_0$ for some optimal dual solution $\widehat{\boldsymbol{\beta}}_j$'s, and let $\widehat{\boldsymbol{\beta}}_j(k_j) = \operatorname{argmin}_{\|\boldsymbol{\beta}_j\|_0\le k_j} \mathcal{L}_j(\boldsymbol{\beta}_j)$. It holds that if $0 < \epsilon < \min_{j\in[d]} \big\{\mathcal{L}_j\big(\widehat{\boldsymbol{\beta}}_j(k_j-1)\big) - \mathcal{L}_j\big(\widehat{\boldsymbol{\beta}}_j(k_j)\big)\big\}$,

$$\sum_{j=1}^d \mathcal{Q}_j(\widehat{\lambda}+\epsilon) - (\widehat{\lambda}+\epsilon)K = \min_{k_j=1,\ldots,d} \Big\{ \sum_{j=1}^d \mathcal{L}_j\big\{\boldsymbol{\beta}_j(k_j)\big\} + (\widehat{\lambda}+\epsilon)\big(\sum_{j=1}^d k_j - K\big)\Big\}.$$

Proof. By the definition of $\mathcal{Q}_j(\cdot)$ in (3.1), when we perturb the optimal Lagrangian multiplier $\widehat{\lambda}$ to $\widehat{\lambda}+\epsilon$, the corresponding dual solutions become

$$\widehat{\boldsymbol{\beta}}'_j = \operatorname*{argmin}_{\boldsymbol{\beta}_j} \mathcal{L}_j(\boldsymbol{\beta}_j) + (\widehat{\lambda}+\epsilon)\|\boldsymbol{\beta}_j\|_0, \text{ for all } j = 1, ..., d.$$

We have, if

$$\mathcal{L}_j\{\widehat{\boldsymbol{\beta}}_j(k_j)\} + (\widehat{\lambda}+\epsilon)k_j < \mathcal{L}_j\{\widehat{\boldsymbol{\beta}}_j(k_j-1)\} + (\widehat{\lambda}+\epsilon)(k_j-1) \text{ for all } j,$$

letting $\widehat{\boldsymbol{\beta}}'_j = \widehat{\boldsymbol{\beta}}_j$ for all j provides dual optimal solutions. Meanwhile, by our assumption of ϵ, it is not difficult to check that the above inequality holds for all j. Thus, our claim holds as desired. □

Lemma D.5. Let $\{\boldsymbol{\beta}_j^{(1)}\}_{j\in[d]},...,\{\boldsymbol{\beta}_j^{(m)}\}_{j\in[d]}$ be the sequence of dual solutions ranked by their corresponding primal objective values as defined in (4.5). We have, for any $m' \in \{1,...,m-1\}$, it holds that the difference of primal objective values of two consecutive solutions is bounded by C_g as defined in (4.3), i.e.,

$$\sum_{j=1}^{d} \mathcal{L}_j\big(\boldsymbol{\beta}_j^{(m'+1)}\big) - \sum_{j=1}^{d} \mathcal{L}_j\big(\boldsymbol{\beta}_j^{(m')}\big) \le C_g.$$

Proof. By the dual optimality of the solutions and the separable structure of the dual objective, given the optimal Lagrangian multiplier $\widehat{\lambda} > 0$, we have,

$$\boldsymbol{\beta}_j^{(m'')} \in \operatorname*{argmin}_{\boldsymbol{\beta}_j} \mathcal{L}_j(\boldsymbol{\beta}_j) + \widehat{\lambda}\|\boldsymbol{\beta}_j\|_0, \text{ for all } j = 1,...,d \text{ and } m'' = 1,...,m.$$

In addition, due to the ℓ_0-penalization term, we have

$$\boldsymbol{\beta}_j^{(m'')} \in \{\widehat{\boldsymbol{\beta}}_j(k)\}_{k\in[d_j]}, \text{ for all } j = 1,...,d, \text{ and } m'' = 1,...,m,$$

where $\widehat{\boldsymbol{\beta}}_j(k) = \operatorname{argmin}_{\|\boldsymbol{\beta}_j\|_0\le k} \mathcal{L}_j(\boldsymbol{\beta}_j)$.

Thus, for each optimal solution $\{\boldsymbol{\beta}_j^{(m'')}\}_{j\in[d]}$, it is a combination of the optimal solutions of the ℓ_0-penalized subproblems. By the dual decomposition structure as seen in (3.2), we have, for any solution achieves dual optimal objective $\{\widehat{\boldsymbol{\beta}}_j^{(m'')}\}$, there exists solution achieves dual optimal objective $\{\widehat{\boldsymbol{\beta}}_j^{(m')}\}$, such that these two solutions differ by at most one $\boldsymbol{\beta}_{j'}$ for some j', i.e., for any $m' \in \{1,...,m-1\}$, there exists one $j' \in \{1,...,d\}$, such that

$$\boldsymbol{\beta}_{j'}^{(m')} \neq \boldsymbol{\beta}_{j'}^{(m'')}, \text{ and } \boldsymbol{\beta}_j^{(m')} = \boldsymbol{\beta}_j^{(m'')} \text{ for all } j \in \{1,...,j'-1,j'+1,...,d\}. \tag{D.1}$$

Consequently, we have

$$\begin{aligned}
\Big|\sum_{j=1}^{d} \mathcal{L}_j\big(\boldsymbol{\beta}_j^{(m')}\big) - \sum_{j=1}^{d} \mathcal{L}_j\big(\boldsymbol{\beta}_j^{(m'')}\big)\Big| &= \Big|\mathcal{L}_j\big(\boldsymbol{\beta}_j^{(m')}\big) - \mathcal{L}_j\big(\boldsymbol{\beta}_j^{(m')}\big)\Big| \\
&\le \big|\mathcal{L}_j(\mathbf{0}) - \min_{\boldsymbol{\beta}_{j'}\in\mathcal{C}_{j'}} \mathcal{L}_{j'}\big(\boldsymbol{\beta}_{j'}\big)\big| \\
&\le C_g,
\end{aligned}$$

where the first inequality holds by $\mathcal{L}_j(\mathbf{0}) \ge \mathcal{L}_j\big(\widehat{\boldsymbol{\beta}}_j(1)\big) \ge,...,\ge \mathcal{L}_j\big(\widehat{\boldsymbol{\beta}}_j(d_j)\big) = \min_{\boldsymbol{\beta}_j\in\mathcal{C}_j} \mathcal{L}_j\big(\boldsymbol{\beta}_j\big)$ for any j, and the second inequality holds by the definition of C_g in (4.3).

Since the sequence $\{\boldsymbol{\beta}_j^{(1)}\}_{j\in[d]}$, $\{\boldsymbol{\beta}_j^{(2)}\}_{j\in[d]}$,..., $\{\boldsymbol{\beta}_j^{(m)}\}_{j\in[d]}$ is ordered by their corresponding objective values, It follows immediately from the above inequality that the difference between any two consecutive objective values is bounded by C_g also, i.e.,

$$\Big|\sum_{j=1}^{d}\mathcal{L}_j\big(\boldsymbol{\beta}_j^{(m'+1)}\big)-\sum_{j=1}^{d}\mathcal{L}_j\big(\boldsymbol{\beta}_j^{(m')}\big)\Big|\le C_g,\ \text{for any } m'=1,...,m-1,$$

and our claim holds as desired. □

E Computational Complexity

We briefly discuss the computational complexities of the dynamic programming approach and the SPICA algorithm in this section. We denote by d_m the average degree per-vertex, so we have

$$d^{-1}K=\Theta(d_m).$$

We are interested in the computational complexity for solving the problem (1.1). The dual method proposed in this paper involves iteratively calling a solver to solve the subproblems

$$\max_{\beta_j}\mathcal{L}_j(\beta_j)+\lambda\|\beta_j\|_0,$$

for every $j=1,\dots,d$. Each subproblem is a combinatorial problem with dimension d_0. We assume that each call to this subproblem solver incurs an identical time complexity, which is a function of d_0, and we denote by $\mathcal{T}(d_0)$.

In the SPICA algorithm, we use the golden section method for maximizing the dual function. To achieve an $\epsilon-$optimal numerical solution, the golden section method requires $\mathcal{O}\left(\log\frac{1}{\epsilon}\right)$ iterations. Each iteration involves calling the subproblem solver for d times. The total computational complexity is

$$\mathcal{O}\left(d\cdot\log\frac{1}{\epsilon}\cdot\mathcal{T}(d_0)\right).$$

Suppose the optimal statistical error is of the order $\mathcal{O}(\frac{K\log d}{dn})$. Consider the regime where the duality gap is no larger than the statistical error, i.e., $\frac{1}{d}=\mathcal{O}(\frac{K\log d}{dn})$, we need $n=\mathcal{O}(K\log d)=\mathcal{O}(d_m d\log d)$. We also require that the optimization error ϵ to be bounded by the statistical error, i.e., $\epsilon=\Theta(\frac{K\log d}{dn})$, which means that the computational complexity becomes

$$\begin{aligned}
d\cdot\log\frac{1}{\epsilon}\cdot\mathcal{T}(d_0)&=d\cdot\log\frac{dn}{K\log d}\cdot\mathcal{T}(d_0)\\
&=d\cdot(\log n-\log d_m-\log\log d)\cdot\mathcal{T}(d_m)\\
&=d\cdot(\log(d_m d\log d)-\log d_m-\log\log d)\cdot\mathcal{T}(d_m)\\
&=\mathcal{O}\Big(d\cdot\log d\cdot\mathcal{T}(d_m)\Big).
\end{aligned}$$

This is the time complexity to obtain an overall error rate $\mathcal{O}\left(\frac{K\log d}{dn}\right)$.

For comparison, let us consider the scheme in which we enumerate all possible subproblem solutions and solves the problem using dynamic programming as discussed in Section 2. This approach provides an exact solution, and its computational complexity is dd_0K. However, enumerating all possible subproblem solutions is costly. For each node, there are d_0 subproblems. The total computational complexity is

$$\mathcal{O}\Big(dd_0 \cdot \mathcal{T}(d_0) + dd_0K\Big).$$

Now we compare the time complexity for these two methods if they both achieve the optimal statistical rates. Suppose that the sample size n is fixed. We consider the relation between the graph's property and the algorithms' computational complexity. Clearly, the SPICA algorithm is more time efficient for graphs with more potential local neighbors such that $d_0 \gg \log(d)$, and the dynamic programming approach is more efficient if $d_0 \ll \log(d)$. Meanwhile, suppose that d_0 is fixed. The SPICA algorithm is faster if $\log n - \log d_m - \log\log d \ll d_m$, i.e.,

$$n \leq \mathcal{O}\Big(d_m \log d \cdot \min\{e^{d_m}, d\}\Big),$$

and the dynamic programming approach is faster otherwise. To summarize, there always exists an algorithm that achieves an statistical error

$$\mathcal{O}\Big(\frac{K\log d}{dn}\Big) = \mathcal{O}\Big(\frac{d_m \log d}{n}\Big)$$

within computational complexity

$$\mathcal{O}\Big(d \cdot \min\{\log d, d_m\} \cdot \mathcal{T}(d_0)\Big).$$

F Proof in Section 4.2

F.1 Proof of Corollary 4.3

Proof. The proof is a consequence of the following two lemmas. The first lemma quantifies the risk of the estimator, which involves the duality gap. The second lemma quantifies the duality gap C_g incurred by the SPICA algorithm.

Lemma F.1. Suppose we have n independent samples of $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma}) \in \mathbb{R}^d$, and the prior information that each X_j can only connect to a set of nodes $\mathcal{N}_j \subset \{1, ..., d\}$, i.e., $\boldsymbol{\Theta}_{jk} = 0$ if $k \notin \mathcal{N}_j$. Let $\{\widehat{\boldsymbol{\beta}}_j\}_{j\in[d]}$ be the estimator obtained by the SPICA algorithm. Assume $2K \leq dd_0$, $s \leq K$, $\sum_{j\in[d]} \|\boldsymbol{\beta}_j^*\|_0 = s$, where $\boldsymbol{\beta}_j^*$'s are defined in (4.7). We further assume $\mathrm{diag}(\boldsymbol{\Sigma}) \leq \sigma^2$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, we have

$$\begin{aligned} &\frac{1}{n}\sum_{j=1}^{d} \|\mathbb{X}_{\mathcal{N}_j}\widehat{\boldsymbol{\beta}}_j - \mathbb{X}_{\mathcal{N}_j}\boldsymbol{\beta}_j^*\|_2^2 \\ &\leq 64\frac{\sigma^2}{n}\log\Big\{\sum_{j=1}^{2K}\binom{dd_0}{j}\Big\} + \frac{128\sigma^2 K}{n}\log 6 + \frac{64\sigma^2}{n}\log(\sigma^{-1}) + 2C_g, \end{aligned} \tag{F.1}$$

where the constant C_g is the duality gap incurred by the SPICA algorithm.

Proof. By Theorem 4.1, we have

$$\frac{1}{n}\sum_{j=1}^{d}\|\mathbb{X}_{\mathcal{N}_j}\widehat{\boldsymbol{\beta}}_j-\mathbb{X}_j\|_2^2\leq\frac{1}{n}\sum_{j=1}^{d}\|\mathbb{X}_{\mathcal{N}_j}\boldsymbol{\beta}_j^*-\mathbb{X}_j\|_2^2+C_g,$$

where C_g is a constant does not scale with d.

Let

$$\begin{gathered}\widehat{\mathbb{X}}=\begin{pmatrix}\mathbb{X}_{\mathcal{N}_1}&\mathbf{0}&\cdots&\mathbf{0}\\\mathbf{0}&\mathbb{X}_{\mathcal{N}_2}&\cdots&\mathbf{0}\\\vdots&\vdots&\ddots&\vdots\\\mathbf{0}&\mathbf{0}&\cdots&\mathbb{X}_{\mathcal{N}_d}\end{pmatrix}\in\mathbb{R}^{nd\times d_0d},\\\mathbf{y}=\begin{pmatrix}\mathbb{X}_1\\\mathbb{X}_2\\\vdots\\\mathbb{X}_d\end{pmatrix}\in\mathbb{R}^{nd},\text{ and }\widehat{\boldsymbol{\beta}}=\begin{pmatrix}\widehat{\boldsymbol{\beta}}_1\\\widehat{\boldsymbol{\beta}}_2\\\vdots\\\widehat{\boldsymbol{\beta}}_d\end{pmatrix},\boldsymbol{\beta}^*=\begin{pmatrix}\boldsymbol{\beta}_1^*\\\boldsymbol{\beta}_2^*\\\vdots\\\boldsymbol{\beta}_d^*\end{pmatrix}\in\mathbb{R}^{d_0d}.\end{gathered}\tag{F.2}$$

Let $\mathbf{y}-\widehat{\mathbb{X}}\boldsymbol{\beta}^*=(\boldsymbol{\epsilon}_1^T,...,\boldsymbol{\epsilon}_d^T)^T=\boldsymbol{\epsilon}^T\in\mathbb{R}^{nd}$. It holds that each component of $\boldsymbol{\epsilon}$ follows a Gaussian distribution with marginal variance at most σ^2. We have

$$\|\mathbf{y}-\widehat{\mathbb{X}}\widehat{\boldsymbol{\beta}}\|_2^2\leq\|\mathbf{y}-\widehat{\mathbb{X}}\boldsymbol{\beta}^*\|_2^2+nC_g=\|\boldsymbol{\epsilon}\|_2^2+nC_g.$$

Meanwhile, it holds that

$$\|\mathbf{y}-\widehat{\mathbb{X}}\widehat{\boldsymbol{\beta}}\|_2^2=\|\widehat{\mathbb{X}}\boldsymbol{\beta}^*+\boldsymbol{\epsilon}-\widehat{\mathbb{X}}\widehat{\boldsymbol{\beta}}\|_2^2=\|\widehat{\mathbb{X}}\widehat{\boldsymbol{\beta}}-\widehat{\mathbb{X}}\boldsymbol{\beta}^*\|_2^2-2\boldsymbol{\epsilon}^T\widehat{\mathbb{X}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)+\|\boldsymbol{\epsilon}\|_2^2.$$

Combining the above two relations, we have

$$\|\widehat{\mathbb{X}}\widehat{\boldsymbol{\beta}}-\widehat{\mathbb{X}}\boldsymbol{\beta}^*\|_2^2\leq2\boldsymbol{\epsilon}^T\widehat{\mathbb{X}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)+nC_g=\|\widehat{\mathbb{X}}\widehat{\boldsymbol{\beta}}-\widehat{\mathbb{X}}\boldsymbol{\beta}^*\|_2\frac{2\boldsymbol{\epsilon}^T\widehat{\mathbb{X}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)}{\|\widehat{\mathbb{X}}\widehat{\boldsymbol{\beta}}-\widehat{\mathbb{X}}\boldsymbol{\beta}^*\|_2}+nC_g.\tag{F.3}$$

By the assumptions that $s\leq K$ and $2K\leq dd_0$, letting $\mathcal{B}_0(2K)=\{\mathbf{v}\in\mathbb{R}^{dd_0}:\|\mathbf{v}\|_0\leq2K\}$, we have that $\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\in\mathcal{B}_0(2K)$. Denote by $\widehat{\mathcal{K}}=\text{supp}(\boldsymbol{\beta}^*-\widehat{\boldsymbol{\beta}})$, and let the $nd\times|\widehat{\mathcal{K}}|$ submatrix of $\widehat{\mathbb{X}}$ be $\widehat{\mathbb{X}}_{\widehat{\mathcal{K}}}$, where $\widehat{\mathcal{K}}\subset\{1,2,....,d_0d\}$. Assume that $\widehat{\mathbb{X}}_{\widehat{\mathcal{K}}}$ is of rank r. Let $\Psi_r=[\psi_1,...,\psi_r]=\mathbb{R}^{nd\times r}$ be an orthonormal basis for the column space of $\widehat{\mathbb{X}}_{\widehat{\mathcal{K}}}$. We have that there exists a vector $\boldsymbol{\nu}\in\mathbb{R}^r$,

$$\widehat{\mathbb{X}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)=\widehat{\mathbb{X}}_{\widehat{\mathcal{K}}}\big\{\widehat{\boldsymbol{\beta}}(\widehat{\mathcal{K}})-\boldsymbol{\beta}^*(\widehat{\mathcal{K}})\big\}=\Psi_r\boldsymbol{\nu}.$$

Plugging the above equation into (F.3), we have

$$\frac{\boldsymbol{\epsilon}^T\widehat{\mathbb{X}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)}{\|\widehat{\mathbb{X}}\widehat{\boldsymbol{\beta}}-\widehat{\mathbb{X}}\boldsymbol{\beta}^*\|_2}=\frac{\boldsymbol{\epsilon}^T\Psi_r\boldsymbol{\nu}}{\|\boldsymbol{\nu}\|_2}\leq\max_{r\leq2K}\sup_{\mathbf{v}\in\mathbb{S}^{r-1}}[\boldsymbol{\epsilon}^T\Psi_r]\mathbf{v},$$

where $\mathbb{S}^{r-1}=\{\mathbf{z}\in\mathbb{R}^r:\|\mathbf{z}\|_2\leq1\}$. Then, by (F.3) and some algebraic manipulation, we have

$$\|\widehat{\mathbb{X}}\widehat{\boldsymbol{\beta}}-\widehat{\mathbb{X}}\boldsymbol{\beta}^*\|_2^2\leq8\max_{r\leq2K}\sup_{\mathbf{v}\in\mathbb{S}^{r-1}}(\overline{\boldsymbol{\epsilon}}_r^T\mathbf{v})^2+2nC_g,$$

where $\widetilde{\boldsymbol{\epsilon}}_r = \Psi_r^T \boldsymbol{\epsilon}$ is a Gaussian random vector with marginal variance at most σ^2.

Taking a union bound, we have that for any $t > 0$,

$$\mathbb{P}\big\{ \max_{r \le 2K} \sup_{\mathbf{v} \in \mathbb{S}^{r-1}} (\widetilde{\boldsymbol{\epsilon}}_r^T \mathbf{v})^2 > t \big\} \le \sum_{r \le 2K} \mathbb{P}\big\{ \sup_{\mathbf{v} \in \mathbb{S}^{r-1}} (\widetilde{\boldsymbol{\epsilon}}_r^T \mathbf{v})^2 > t \big\}.$$

By Lemma G.1, we have

$$\mathbb{P}\big\{ \max_{r \le 2K} \sup_{\mathbf{v} \in \mathbb{S}^{r-1}} (\widetilde{\boldsymbol{\epsilon}}_r^T \mathbf{v})^2 > t \big\} \le 6^r \exp\big\{ -t/(8\sigma^2) \big\} \le 6^{2K} \exp\big\{ -t/(8\sigma^2) \big\}.$$

Combining the above three inequalities, we have, for any $t > 0$,

$$\mathbb{P}\big(\|\widehat{\mathbb{X}}\widehat{\boldsymbol{\beta}} - \widehat{\mathbb{X}}\boldsymbol{\beta}^*\|_2^2 > 8t + 2nC_g \big) \le \sum_{j=1}^{2K} \binom{dd_0}{j} 6^{2K} e^{-\frac{t}{8\sigma^2}}.$$

To ensure that the RHS of the above inequality is bounded up by δ, we need to ensure that

$$t \ge 8\sigma^2 \log\Big(\sum_{j=1}^{2K} \binom{dd_0}{j} \Big) + 16K\sigma^2 \log(6) + 8\sigma^2 \log(1/\delta),$$

and the claim (F.1) holds as desired. □

The following lemma provides a bound for the duality gap C_g.

Lemma F.2. Suppose we have n independent samples of $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma}) \in \mathbb{R}^d$. Let $\mathcal{L}_j(\boldsymbol{\beta}_j) = \|\mathbb{X}_j - \mathbb{X}_{\mathcal{N}_j}\boldsymbol{\beta}_j\|_2^2$ be the least square loss. We have,

$$\max_j \big\{ \mathcal{L}_j(\mathbf{0}) - \min_{\boldsymbol{\beta}_j} \mathcal{L}_j(\boldsymbol{\beta}_j) \big\} \le n\sigma^2 + C \cdot n \log d,$$

with probability at least $1 - \mathcal{O}(d^{-1})$, where C is a constant.

Proof. For each j, since the loss function $\mathcal{L}_j(\cdot)$ is nonnegative, we have,

$$\mathcal{L}_j(\mathbf{0}) - \min_{\boldsymbol{\beta}_j} \mathcal{L}_j(\boldsymbol{\beta}_j) \le \mathcal{L}_j(\mathbf{0}) = \sum_{i=1}^n x_{ij}^2.$$

Since each X_j follows a normal distribution with variance at most σ^2, we have $\nu^{-1} \sum_{i \in [n]} x_{ij}^2$ follows a chi-squared distribution with n degrees of freedom, where $\nu < \sigma^2$ is a positive constant. Denote by $Z_j = \nu^{-1} \sum_{i \in [n]} x_{ij}^2$. By Vempala (2005), we have,

$$\mathbb{P}\big(\big| n^{-1} Z_j - 1 \big| \ge t \big) \le 2 \exp\big\{ -4^{-1} n (t^2 - t^3) \big\}, \text{ for any } t > 0.$$

Using Bonferroni's method, we have

$$\mathbb{P}\big(\big| n^{-1} \max_{j \le d} Z_j - 1 \big| \ge t \big) \le 2d \exp\big\{ -4^{-1} n (t^2 - t^3) \big\}, \text{ for any } t > 0.$$

Consequently, our claim follows by performing some algebraic manipulations. □

Combining the above two lemmas, and plugging G.1 into (F.1), our claim follows as desired. □

F.2 Proof of Corollary 4.4

Proof. We first prove the global minimum of the total cardinality approach recovers the support with high probability. By (D.1), the estimator obtained by the SPICA is an optimal solution under the constraint $\sum_{j\in[d]}\|\boldsymbol{\beta}_j\|_0 = K-k$, where $0\le k\le d_n$. When $K=s$, it is not difficult to generalize the proof to prove that this estimator recovers $s-s_1$ components of true support.

Adopt the notations in (F.2). For any subset $\mathcal{T}\subset\{1,...,d_0d\}$, let

$$f(\mathcal{T}) = \min_{\boldsymbol{\beta}\in\mathbb{R}^{|\mathcal{T}|}} \|\mathbf{y}-\widehat{\mathbb{X}}_{\mathcal{T}}\boldsymbol{\beta}_{\mathcal{T}}\|^2.$$

For a fixed set $\mathcal{S}$, let $\Delta(\mathcal{K}) = f(\mathcal{K}) - f(\mathcal{S})$ and $\mathcal{G} = \{\mathcal{K}\subset\{1,...,d_0d\} : |\mathcal{K}| = s, \mathcal{K}\neq\mathcal{S}\}$. Then

$$\mathbb{P}(\mathcal{K}\neq\mathcal{S}) = \mathbb{P}\big(\cup_{\mathcal{K}\in\mathcal{G}}\{\Delta(\mathcal{K})<0\}\big) \le \sum_{\mathcal{K}\in\mathcal{G}}\mathbb{P}\big(\Delta(\mathcal{K})<0\big).$$

Denote the least square estimator restricted to the support $\mathcal{K}$ by

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}_{\mathcal{K}} &= (\widehat{\mathbb{X}}_{\mathcal{K}}^T\widehat{\mathbb{X}}_{\mathcal{K}})^{-1}\widehat{\mathbb{X}}_{\mathcal{K}}^T\mathbf{y} = (\widehat{\mathbb{X}}_{\mathcal{K}}^T\widehat{\mathbb{X}}_{\mathcal{K}})^{-1}\widehat{\mathbb{X}}_{\mathcal{K}}^T(\widehat{\mathbb{X}}_{\mathcal{K}}\boldsymbol{\beta}^*_{\mathcal{K}} + \widehat{\mathbb{X}}_{\mathcal{D}}\boldsymbol{\beta}^*_{\mathcal{D}} + \boldsymbol{\epsilon})\\ &= \boldsymbol{\beta}^*_{\mathcal{K}} + (\widehat{\mathbb{X}}_{\mathcal{K}}^T\widehat{\mathbb{X}}_{\mathcal{K}})^{-1}\widehat{\mathbb{X}}_{\mathcal{K}}(\widehat{\mathbb{X}}_{\mathcal{D}}\boldsymbol{\beta}^*_{\mathcal{D}} + \boldsymbol{\epsilon}).\end{aligned}$$

Let the difference between $\mathcal{S}$ and $\mathcal{K}$ be $\mathcal{D} = \mathcal{S}\backslash\mathcal{K}$. We have,

$$\begin{aligned}&\|\mathbf{y}-\widehat{\mathbb{X}}_{\mathcal{K}}\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}_{\mathcal{K}}\|_2^2\\ =&\|\boldsymbol{\epsilon} + \widehat{\mathbb{X}}_{\mathcal{K}}\boldsymbol{\beta}^*_{\mathcal{K}} + \widehat{\mathbb{X}}_{\mathcal{D}}\boldsymbol{\beta}^*_{\mathcal{D}} - \widehat{\mathbb{X}}_{\mathcal{K}}\boldsymbol{\beta}^*_{\mathcal{K}} - \widehat{\mathbb{X}}_{\mathcal{K}}(\widehat{\mathbb{X}}_{\mathcal{K}}^T\widehat{\mathbb{X}}_{\mathcal{K}})^{-1}\widehat{\mathbb{X}}_{\mathcal{K}}^T(\widehat{\mathbb{X}}_{\mathcal{D}}\boldsymbol{\beta}^*_{\mathcal{D}} + \boldsymbol{\epsilon})\|_2^2\\ =&\|\big\{\mathbf{I}_{nd} - \widehat{\mathbb{X}}_{\mathcal{K}}(\widehat{\mathbb{X}}_{\mathcal{K}}^T\widehat{\mathbb{X}}_{\mathcal{K}})^{-1}\widehat{\mathbb{X}}_{\mathcal{K}}^T\big\}\boldsymbol{\epsilon} + \big\{\mathbf{I}_{nd} - \widehat{\mathbb{X}}_{\mathcal{K}}(\widehat{\mathbb{X}}_{\mathcal{K}}^T\widehat{\mathbb{X}}_{\mathcal{K}})^{-1}\widehat{\mathbb{X}}_{\mathcal{K}}^T\big\}\widehat{\mathbb{X}}_{\mathcal{D}}\boldsymbol{\beta}^*_{\mathcal{D}}\|_2^2\\ =&\boldsymbol{\epsilon}^T\mathbf{P}^{\perp}_{\mathcal{K}}\boldsymbol{\epsilon} + 2\boldsymbol{\epsilon}^T\mathbf{P}^{\perp}_{\mathcal{K}}\widehat{\mathbb{X}}_{\mathcal{D}}\boldsymbol{\beta}^*_{\mathcal{D}} + \|\mathbf{P}^{\perp}_{\mathcal{K}}\widehat{\mathbb{X}}_{\mathcal{D}}\boldsymbol{\beta}^*_{\mathcal{D}}\|_2^2,\end{aligned}$$

where $\mathbf{P}^{\perp}_{\mathcal{K}} = \mathbf{I}_{nd} - \mathbf{P}_{\mathcal{K}}$, and $\mathbf{P}_{\mathcal{K}} = \widehat{\mathbb{X}}_{\mathcal{K}}(\widehat{\mathbb{X}}_{\mathcal{K}}^T\widehat{\mathbb{X}}_{\mathcal{K}})^{-1}\widehat{\mathbb{X}}_{\mathcal{K}}^T$ is a projection matrix, i.e., $\mathbf{P}^2_{\mathcal{K}} = \mathbf{P}_{\mathcal{K}}$, which implies that $\mathbf{P}^{\perp}_{\mathcal{K}}$ is also a projection matrix.

By the same argument, we have

$$f(\mathcal{S}) = \boldsymbol{\epsilon}^T\mathbf{P}^{\perp}_{\mathcal{S}}\boldsymbol{\epsilon}.$$

We have, when $\Delta(\mathcal{K})<0$,

$$\boldsymbol{\epsilon}^T\mathbf{P}^{\perp}_{\mathcal{K}}\boldsymbol{\epsilon} + 2\boldsymbol{\epsilon}^T\mathbf{P}^{\perp}_{\mathcal{K}}\widehat{\mathbb{X}}_{\mathcal{D}}\boldsymbol{\beta}^*_{\mathcal{D}} + \|\mathbf{P}^{\perp}_{\mathcal{K}}\widehat{\mathbb{X}}_{\mathcal{D}}\boldsymbol{\beta}^*_{\mathcal{D}}\|_2^2 < \boldsymbol{\epsilon}^T\mathbf{P}^{\perp}_{\mathcal{S}}\boldsymbol{\epsilon}.$$

Let $\mathcal{S}_j = \mathrm{supp}(\boldsymbol{\beta}^*_j)$ and $\mathcal{K}_j = \mathrm{supp}(\widehat{\boldsymbol{\beta}}^K_j)$, and let s' be the number of $\mathcal{S}_j\neq\mathcal{K}_j$. By Lemma G.3, we have

$$\Delta(\mathcal{K}) \ge \widehat{\Delta}(\mathcal{K}) = 2\boldsymbol{\epsilon}\mathbf{P}^{\perp}_{\mathcal{K}}\widehat{\mathbb{X}}_{\mathcal{D}}\boldsymbol{\beta}^*_{\mathcal{D}} + \|\mathbf{P}^{\perp}_{\mathcal{K}}\widehat{\mathbb{X}}_{\mathcal{D}}\boldsymbol{\beta}^*_{\mathcal{D}}\|_2^2 - \sigma^2R,$$

where R follows a chi-square distribution with $s'd_0$ degrees of freedom.

We have, when $\widehat{\Delta}(\mathcal{K})<0$ holds,

$$-2\boldsymbol{\epsilon}^T\mathbf{P}^{\perp}_{\mathcal{K}}\widehat{\mathbb{X}}_{\mathcal{D}}\boldsymbol{\beta}^*_{\mathcal{D}} \ge \|\mathbf{P}^{\perp}_{\mathcal{K}}\widehat{\mathbb{X}}_{\mathcal{D}}\boldsymbol{\beta}^*_{\mathcal{D}}\|_2^2 - \sigma^2R.$$

Using the techniques in the proof of Theorem 4.3, we get,

$$\|\mathbf{P}_{\mathcal{K}}^{\perp}\widehat{\mathbb{X}}_{\mathcal{D}}\boldsymbol{\beta}_{\mathcal{D}}^{*}\|_2^2 \leq 8 \sup_{\mathbf{u}\in\mathbb{R}^{|\mathcal{D}|}} (\widetilde{\boldsymbol{\epsilon}}\mathbf{u})^2 + 2\sigma^2 R.$$

With probability at least $1-2\delta$, it follows that, by Lemma G.1 and Lemma 10 of Kolar and Liu (2013),

$$\|\mathbf{P}_{\mathcal{K}}^{\perp}\widehat{\mathbb{X}}_{\mathcal{D}}\boldsymbol{\beta}_{\mathcal{D}}^{*}\|_2^2 \leq 64\sigma^2|\mathcal{D}|\log 6 + 64\sigma^2\log\delta^{-1} + 2\big(s'd_0\sigma^2 + C\sqrt{s'd_0\log\delta^{-1}}\big),$$

for some constant C.

By Lemma G.2 and the sparse eigenvalue assumption, $\Lambda_{\max}(\widehat{\mathbb{X}}_{\mathcal{K}}^T\widehat{\mathbb{X}}_{\mathcal{K}}) \geq \Lambda_{\max}(\widehat{\mathbb{X}}_{\mathcal{D}}^T\mathbf{P}_{\mathcal{K}}^{\perp}\widehat{\mathbb{X}}_{\mathcal{D}}) \geq \Lambda_{\min}(\widehat{\mathbb{X}}_{\mathcal{D}}^T\mathbf{P}_{\mathcal{K}}^{\perp}\widehat{\mathbb{X}}_{\mathcal{D}}) \geq \Lambda_{\min}(\widehat{\mathbb{X}}_{\mathcal{K}}^T\widehat{\mathbb{X}}_{\mathcal{K}}) \geq \rho$ with probability at least $1-\mathcal{O}(d^{-1})$. Let s_1 be the number of different support of $\mathcal{S}$ and $\mathcal{K}$, i.e., $s_1 = |\mathcal{D}|$. Taking $\delta = \big\{\binom{dd_0-s}{s_1}\binom{s}{s-s_1}d\big\}^{-1}$, as $s > s_1$, it is not difficult to see that the dominating term is $64\sigma^2\log\delta^{-1}$, and by Lemma G.1,

$$\begin{aligned}\log\delta^{-1} &= \log\left\{\binom{dd_0-s}{s_1}\right\} + \log\left\{\binom{s}{s_1}\right\} + \log d \\ &\leq C_1 s_1\log dd_0 + C_2 s_1\log s + \log d,\end{aligned}$$

where the dominating term is $s_1\log dd_0$.

Consequently, we have that, with probability at least $1-\mathcal{O}\Big(\big\{\binom{dd_0-s}{s_1}\binom{s}{s-s_1}d\big\}^{-1}\Big)$,

$$s_1\|\boldsymbol{\beta}_{\mathcal{D}}^{*}\|_{\min}^2 \leq \frac{64d_0s_1\sigma^2\log d}{nC},$$

for some constant C, which violates our assumption (4.9). This implies that $\Delta(\mathcal{K}) < 0$ holds with probability at most $\mathcal{O}\Big(\big\{\binom{dd_0-s}{s_1}\binom{s}{s-s_1}d\big\}^{-1}\Big)$. Taking a union bound, we have

$$\mathbb{P}(\mathcal{K}\neq\mathcal{S}) = \mathcal{O}(d^{-1}),$$

as desired. □

F.3 Proof of Corollary 4.6

Proof. Recall that the logistic loss of the j-th variable is

$$\mathcal{L}_j(\boldsymbol{\beta}_j) = -\frac{1}{n}\sum_{i=1}^{n}\log\big\{1+\exp(-x_{ij}\mathbf{x}_{i,\mathcal{N}_j}^T\boldsymbol{\beta}_j)\big\},$$

and its gradient at $\boldsymbol{\beta}_j^*$ is

$$\nabla\mathcal{L}_j(\boldsymbol{\beta}_j^*) = \frac{1}{n}\sum_{i=1}^{n}x_{ij}\mathbf{x}_{i,N_j}\big\{1+\exp(x_{ij}\mathbf{x}_{i,N_j}\boldsymbol{\beta}^*)\big\}.$$

Taking expectation, by tower rule, we have

$$\begin{aligned}\mathbb{E}[\nabla\mathcal{L}_j(\boldsymbol{\beta}_j^*)] = \mathbb{E}\Big[\mathbb{E}\big[&\frac{\boldsymbol{X}_{\mathcal{N}_j}}{1+\exp(\boldsymbol{X}_{\mathcal{N}_j}\boldsymbol{\beta}_j^*)}\mathbb{P}(X_j=1|\boldsymbol{X}_{\mathcal{N}_j}) \\ &- \frac{\boldsymbol{X}_{\mathcal{N}_j}}{1+\exp(-\boldsymbol{X}_{\mathcal{N}_j}\boldsymbol{\beta}_j^*)}\mathbb{P}(X_j=-1|\boldsymbol{X}_{\mathcal{N}_j})\big|\boldsymbol{X}_{\mathcal{N}_j}\big]\Big] \\ = \mathbf{0}.\end{aligned}$$

Note that each component of $\nabla \mathcal{L}_j(\boldsymbol{\beta}^*) \in \mathbb{R}^{d_0}$ is bounded in $[-1, 1]$. By Hoeffding's inequality, we have that, for any $k \in \mathcal{N}_j$, and any $t > 0$, $\nabla_k \mathcal{L}_j(\boldsymbol{\beta}_j^*)$ satisfies

$$\mathbb{P}(|\nabla_k \mathcal{L}_j(\boldsymbol{\beta}_j^*)| \geq t) \leq 2\exp(-nt^2/2).$$

Using Bonferroni's method, we have, with probability at least $1 - \mathcal{O}(d^{-1})$, for some constant $C > 0$

$$\max_j \|\nabla \mathcal{L}_j(\boldsymbol{\beta}_j^*)\|_\infty \leq C \cdot \sqrt{\frac{\log d}{n}}. \tag{F.4}$$

Next, we look at the Hessian matrix of $\mathcal{L}_j(\boldsymbol{\beta}_j)$,

$$\nabla^2 \mathcal{L}_j(\boldsymbol{\beta}_j) = \frac{1}{n}\sum_{i=1}^n \frac{\exp(x_{ij}\mathbf{x}_{i,\mathcal{N}_j}^T \boldsymbol{\beta}_j) x_{ij}^2 \mathbf{x}_{i,\mathcal{N}_j}^T \mathbf{x}_{i,\mathcal{N}_j}}{\{1 + \exp(x_{ij}\mathbf{x}_{i,\mathcal{N}_j}^T \boldsymbol{\beta}_j)\}^2}.$$

Also, the Hessian of $\mathcal{L}(\boldsymbol{\beta})$ is a block diagonal matrix of $\nabla^2 \mathcal{L}_1(\boldsymbol{\beta}_1), ..., \nabla^2 \mathcal{L}_d(\boldsymbol{\beta}_d)$. By assumptions (B.1) and (B.2), it has been shown by Lemma 5 of Ravikumar et al. (2010) that $\sum_{j=1}^d \mathcal{L}_j(\boldsymbol{\beta}_j)$ is ρ-strongly convex with respect to the support $\mathcal{S} \cup \mathcal{K}$ with probability at least $1 - \mathcal{O}(d^{-1})$.

Let $\mathcal{L}^{(\mathcal{K})} : \mathbb{R}^{s+K} \to \mathbb{R}$ be the empirical loss function restricted to the support $\mathcal{K} \cup \mathcal{S}$, where $\mathcal{K}$ is the support of $\{\widehat{\boldsymbol{\beta}}_j\}_{j=1}^d$. By the definition of ρ-strongly convexity, we have, with probability at least $1 - O(d^{-1})$,

$$\sum_{j=1}^d \mathcal{L}_j^{(\mathcal{K})}(\widehat{\boldsymbol{\beta}}_j) \geq \sum_{j=1}^d \mathcal{L}_j^{(\mathcal{K})}(\boldsymbol{\beta}_j^*) + \sum_{j=1}^d \nabla \mathcal{L}^{(\mathcal{K})}(\boldsymbol{\beta}_j^*)^T(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*) + \frac{\rho}{2}\sum_{j=1}^d \|\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\|_2^2.$$

Next, by the blessing of massiveness result, we have, since the logistic loss is bounded under the

$$\sum_{j=1}^d \mathcal{L}_j(\widehat{\boldsymbol{\beta}}) \leq \sum_{j=1}^d \mathcal{L}_j(\boldsymbol{\beta}^*) + C_g$$

where $C_g = \max_j \left\{\mathcal{L}_j(\mathbf{0}) - \min_{\boldsymbol{\beta}_j} \mathcal{L}_j(\boldsymbol{\beta}_j)\right\}$. Under the assumption (B.1), we have the logistic loss is bounded, which implies that C_g is a bounded constant. Thus, together with (F.4), we have, with probability at least $1 - \mathcal{O}(d^{-1})$,

$$\sum_{j=1}^d \|\boldsymbol{\beta}_j^* - \widehat{\boldsymbol{\beta}}_j\|_2^2 \leq C\Big(\frac{(s+K)\log d}{n\rho^2} + \frac{C_g}{\alpha}\Big),$$

for some constant $C > 0$, as desired. □

G Technical Lemmas

Lemma G.1. Let $\boldsymbol{X} \in \mathbb{R}^d$ be a sub-Gaussian random vector with mean 0 and variance proxy σ^2. Then,

$$\mathbb{P}(\max_{\boldsymbol{\theta} \in \mathbb{S}^{d-1}} \boldsymbol{\theta}^T \boldsymbol{X} > t) \leq 6^d \exp(-t^2/8\sigma^2).$$

Proof. Let $\mathcal{N}$ be a 1/2-net of $\mathbb{S}^{d-1}$ with respect to the ℓ_2-norm that satisfies $|\mathcal{N}| \leq 6^d$, where such a net exists by Boucheron et al. (2013). For any $\boldsymbol{\theta} \in \mathbb{S}^{d-1}$, there exists $\mathbf{z}_1 \in \mathcal{N}$ and $\|\mathbf{z}_2\|_2 \leq 1/2$ such that $\boldsymbol{\theta} = \mathbf{z}_1 + \mathbf{z}_2$. We have

$$\max_{\boldsymbol{\theta} \in \mathbb{S}^{d-1}} \boldsymbol{\theta}^T \boldsymbol{X} \leq \max_{\mathbf{z}_1 \in \mathcal{N}} \mathbf{z}_1^T \boldsymbol{X} + \max_{\mathbf{z}_2 : \|\mathbf{z}_2\|_2 \leq 1/2} \mathbf{z}_2^T \boldsymbol{X} = \max_{\mathbf{z}_1 \in \mathcal{N}} \mathbf{z}_1^T \boldsymbol{X} + \frac{1}{2} \max_{\mathbf{z}_2 \in \mathbb{S}^{d-1}} \mathbf{z}_2^T \boldsymbol{X}.$$

Therefore, we have

$$\max_{\boldsymbol{\theta} \in \mathbb{S}^{d-1}} \boldsymbol{\theta}^T \boldsymbol{X} \leq 2 \max_{\mathbf{z} \in \mathcal{N}} \mathbf{z}^T \boldsymbol{X}.$$

Consequently, it holds that

$$\mathbb{P}(\max_{\boldsymbol{\theta} \in \mathbb{S}^{d-1}} \boldsymbol{\theta}^T \boldsymbol{X} > t) \leq \mathbb{P}(2 \max_{\mathbf{z} \in \mathcal{N}} \mathbf{z}^T \boldsymbol{X} > t) \leq |\mathcal{N}| e^{-t^2/8\sigma^2} \leq 6^d e^{-t^2/8\sigma^2}?$$

where the second inequality follows by taking a union bound and the fact that any sub-Gaussian random variable X with variance proxy σ^2, we have $\mathbb{P}(X > t) \leq \exp(-t/8\sigma^2)$. □

Lemma G.2. Given any $\mathbb{Z}_1 \in \mathbb{R}^{n \times d_1}$ and $\mathbb{Z}_2 \in \mathbb{R}^{n \times d_2}$, let $\mathbb{Z} = (\mathbb{Z}_1, \mathbb{Z}_2) \in \mathbb{R}^{n \times (d_1 + d_2)}$. We have

$$\Lambda_{\max}(\mathbb{Z}^T \mathbb{Z}) \geq \Lambda_{\max}(\mathbb{Z}_1^T \mathbf{P}_2 \mathbb{Z}_1) \geq \Lambda_{\min}(\mathbb{Z}_1^T \mathbf{P}_2^{\perp} \mathbb{Z}_1) \geq \Lambda_{\min}(\mathbb{Z}^T \mathbb{Z}),$$

where $\mathbf{P}_2 = (\mathbf{I}_n - \mathbb{Z}_2(\mathbb{Z}_2^T \mathbb{Z}_2)^{-1} \mathbb{Z}_2^T)$.

Proof. Observe that

$$\mathbb{Z}^T \mathbb{Z} = \begin{pmatrix} \mathbb{Z}_1^T \mathbb{Z}_1 & \mathbb{Z}_1^T \mathbb{Z}_2 \\ \mathbb{Z}_2^T \mathbb{Z}_1 & \mathbb{Z}_2^T \mathbb{Z}_2 \end{pmatrix},$$

and its the inverse of $\mathbb{Z}^T \mathbb{Z}$ is

$$\begin{aligned} &\begin{pmatrix} \mathbb{Z}_1^T \mathbb{Z}_1 & \mathbb{Z}_1^T \mathbb{Z}_2 \\ \mathbb{Z}_2^T \mathbb{Z}_1 & \mathbb{Z}_2^T \mathbb{Z}_2 \end{pmatrix}^{-1} \\ &= \begin{pmatrix} (\mathbb{Z}_{11} - \mathbb{Z}_{12} \mathbb{Z}_{22}^{-1} \mathbb{Z}_{21})^{-1} & -\mathbb{Z}_{11}^{-1} \mathbb{Z}_{12} (\mathbb{Z}_{22} - \mathbb{Z}_{21} \mathbb{Z}_{11}^{-1} \mathbb{Z}_{12})^{-1} \\ -\mathbb{Z}_{22}^{-1} \mathbb{Z}_{21} (\mathbb{Z}_{11} - \mathbb{Z}_{12} \mathbb{Z}_{22}^{-1} \mathbb{Z}_{21})^{-1} & (\mathbb{Z}_{22} - \mathbb{Z}_{21} \mathbb{Z}_{11}^{-1} \mathbb{Z}_{12})^{-1} \end{pmatrix}, \end{aligned}$$

where $\mathbb{Z}_{jk} = \mathbb{Z}_j^T \mathbb{Z}_k$ for $j, k = 1, 2$.

It is seen that $(\mathbb{Z}_1^T \mathbb{Z}_1 - \mathbb{Z}_1^T \mathbb{Z}_2 (\mathbb{Z}_2^T \mathbb{Z}_2)^{-1} \mathbb{Z}_2^T \mathbb{Z}_1)^{-1}$ is a submatrix of $(\mathbb{Z}^T \mathbb{Z})^{-1}$. Our claim follows immediately that the eigenvalues of $(\mathbb{Z}_1^T \mathbb{Z}_1 - \mathbb{Z}_1^T \mathbb{Z}_2 (\mathbb{Z}_2^T \mathbb{Z}_2)^{-1} \mathbb{Z}_2^T \mathbb{Z}_1)^{-1}$ is in the range $[\Lambda_{\min}(\mathbb{Z}^T \mathbb{Z}), \Lambda_{\max}(\mathbb{Z}^T \mathbb{Z})]$. □

Lemma G.3. Suppose each $\boldsymbol{\epsilon}_j = (\epsilon_{j1}, ..., \boldsymbol{\epsilon}_{jd})^T \in \mathbb{R}^d$ where $\epsilon_{ii'}$'s are i.i.d normally distributed with mean 0 and variance σ_j^2 for $j = 1, ..., d$. Let $\sigma^2 = \max_{j=1,...,d} \sigma_j^2$. Using the notations used in Theorem 4.4, let $\mathbf{P}_{\mathcal{K}} = \mathbb{X}_{\mathcal{K}} (\mathbb{X}_{\mathcal{K}}^T \mathbb{X}_{\mathcal{K}})^{-1} \mathbb{X}_{\mathcal{K}}$ and $\mathbf{P}_{\mathcal{S}} = \mathbb{X}_{\mathcal{S}} (\mathbb{X}_{\mathcal{S}}^T \mathbb{X}_{\mathcal{S}})^{-1} \mathbb{X}_{\mathcal{S}}$. Let $\mathcal{K}_j = \text{supp}(\widehat{\boldsymbol{\beta}}_j)$ and $\mathcal{S}_j = \text{supp}(\boldsymbol{\beta}_j^*)$ for $j = 1, ..., d$, and let s' be the number of $\mathcal{K}_j \neq \mathcal{S}_j$. We have,

$$\boldsymbol{\epsilon}^T \mathbf{P}_{\mathcal{S}} \boldsymbol{\epsilon} - \boldsymbol{\epsilon}^T \mathbf{P}_{\mathcal{K}} \boldsymbol{\epsilon} \geq -\sigma^2 Z,$$

where Z is a random variable follows a chi-square distribution with $d_0 s'$ degrees of freedom.

Proof. Let $\widehat{\mathcal{D}} = \{j_1', j_2', ..., j_{s'}'\}$ be the indices that $\mathcal{S}_{j_k'} \neq \mathcal{K}_{j_k'}$ for $k = 1, ..., s'$. It is readily seen that

$$\begin{aligned}
&\boldsymbol{\epsilon}^T \mathbf{P}_{\mathcal{S}} \boldsymbol{\epsilon} - \boldsymbol{\epsilon}^T \mathbf{P}_{\mathcal{K}} \boldsymbol{\epsilon}^T \\
&\qquad = \sum_{j=1}^{d} \boldsymbol{\epsilon}_j^T \mathbb{X}_{\mathcal{S}_j} (\mathbb{X}_{\mathcal{S}_j}^T \mathbb{X}_{\mathcal{S}_j})^{-1} \mathbb{X}_{\mathcal{S}_j} \boldsymbol{\epsilon}_j - \boldsymbol{\epsilon}_j^T \mathbb{X}_{\mathcal{K}_j} (\mathbb{X}_{\mathcal{K}_j}^T \mathbb{X}_{\mathcal{K}_j})^{-1} \mathbb{X}_{\mathcal{K}_j} \boldsymbol{\epsilon}_j \\
&\qquad = \sum_{j \in \{1,...,d\} \setminus \widehat{\mathcal{D}}} \Big(\boldsymbol{\epsilon}_j^T \mathbb{X}_{\mathcal{S}_j} (\mathbb{X}_{\mathcal{S}_j}^T \mathbb{X}_{\mathcal{S}_j})^{-1} \mathbb{X}_{\mathcal{S}_j} \boldsymbol{\epsilon}_j - \boldsymbol{\epsilon}_j^T \mathbb{X}_{\mathcal{K}_j} (\mathbb{X}_{\mathcal{K}_j}^T \mathbb{X}_{\mathcal{K}_j})^{-1} \mathbb{X}_{\mathcal{K}_j} \boldsymbol{\epsilon}_j \Big) \\
&\qquad\quad + \sum_{j \in \widehat{\mathcal{D}}} \Big(\boldsymbol{\epsilon}_j^T \mathbb{X}_{\mathcal{S}_j} (\mathbb{X}_{\mathcal{S}_j}^T \mathbb{X}_{\mathcal{S}_j})^{-1} \mathbb{X}_{\mathcal{S}_j} \boldsymbol{\epsilon}_j - \boldsymbol{\epsilon}_j^T \mathbb{X}_{\mathcal{K}_j} (\mathbb{X}_{\mathcal{K}_j}^T \mathbb{X}_{\mathcal{K}_j})^{-1} \mathbb{X}_{\mathcal{K}_j} \boldsymbol{\epsilon}_j \Big) \\
&\qquad \geq - \sum_{j \in \widehat{\mathcal{D}}}^{d} \boldsymbol{\epsilon}_j^T \mathbb{X}_{\mathcal{K}_j} (\mathbb{X}_{\mathcal{K}_j}^T \mathbb{X}_{\mathcal{K}_j})^{-1} \mathbb{X}_{\mathcal{K}_j} \boldsymbol{\epsilon}_j
\end{aligned}$$

Since $\boldsymbol{\epsilon}_j^T \mathbb{X}_{\mathcal{K}_j} (\mathbb{X}_{\mathcal{K}_j}^T \mathbb{X}_{\mathcal{K}_j})^{-1} \mathbb{X}_{\mathcal{K}_j} \boldsymbol{\epsilon}_j$ is a projection matrix, its eigenvalues are 1's and 0's. In addition, the matrix's rank is $|\mathcal{K}_j| \leq d_0$ as $n > d_0$. We have

$$- \sum_{j \in \widehat{\mathcal{D}}} \boldsymbol{\epsilon}_j^T \mathbb{X}_{\mathcal{K}_j} (\mathbb{X}_{\mathcal{K}_j}^T \mathbb{X}_{\mathcal{K}_j})^{-1} \mathbb{X}_{\mathcal{K}_j} \boldsymbol{\epsilon}_j = - \sum_{j \in \widehat{\mathcal{D}}} \sigma_j^2 Z_j \geq -\sigma^2 Z,$$

where $Z_j \sim \chi^2_{|\mathcal{K}_j|}$ and $Z \sim \chi^2_{s' d_0}$, and the claim holds as desired. □

Lemma G.4. For any integer k, such that $k \in [1, d]$, we have

$$\sum_{j=0}^{k} \binom{d}{j} \leq \Big(\frac{ed}{k} \Big)^k \tag{G.1}$$

Proof. Since the function $f(x) = (x^{-1} ed)^x$ is increasing when $x \geq 1$, if $k \geq d/2$, we have

$$\sum_{j=0}^{k} \binom{d}{j} \leq 2^d \leq (2e)^{d/2} = f(d/2) \leq f(k).$$

When $k < d/2$ and let $Z \sim \mathrm{Bin}(d, 0.5)$. We have

$$\mathbb{P}(Z \leq k) = \sum_{j=0}^{k} \binom{d}{j} 2^{-d}.$$

Thus, we have

$$\sum_{j=0}^{k} \binom{d}{j} = 2^d \mathbb{P}(Z - \mathbb{E}(Z) \leq k - d/2).$$

Taking a Chernoff bound, we have that for any $k' > 0$

$$\begin{aligned}
2^n \mathbb{P}(Z - \mathbb{E}(Z) \leq k - d/2) &\leq \exp\Big\{ d\phi(k') + k'(k - d/2) + d \log 2 \Big\} \\
&= \exp\big\{ kk' + d \log(1 + e^{-k'}) \big\},
\end{aligned} \tag{G.2}$$

where we let $U \sim \mathrm{Ber}(0.5)$ and

$$\phi(k') = \log \mathbb{E}\Big(e^{k'(1/2-U)}\Big) = \frac{k'}{2} + \log(1+e^{-k'}) - \log 2.$$

Next, we bound the term $kk' + d\log(1+e^{-k'}$. Taking $k^* = \log(\frac{d+k}{k})$ and $z = k/d < 1/2$, we have

$$k^*k + d\log(1+e^{-k^*}) = d\Big\{z\log\big(\frac{1+z}{z}\big) + \log\big(1+\frac{z}{1+z}\big)\Big\} \le d\big\{z - z\log z\big\},$$

where the inequality follows by Lemma G.5. Plugging this into (G.2), we have

$$\sum_{j=0}^{k}\binom{d}{j} \le \exp\big\{k + k\log(d/k)\big\} = \Big(\frac{en}{k}\Big)^k,$$

which finishes the proof. □

Lemma G.5. The function

$$\psi(z) = z\log\big(\frac{1+z}{z}\big) + \log\big(1+\frac{z}{1+z}\big)$$

satisfies

$$\psi(z) \le \phi(z) = z - z\log(z), \text{ for any } z \in (0, 1/2].$$

Proof. It is not difficult to verify that the function $\psi(z) - z\log(z)$ is convex. Thus, we only need to prove that $\psi(z) - z\log(z) \le z$ for $z = 0$ and $1/2$. By L'Hospital's rule, we have

$$\lim_{z\to 0^+} z\log(1+z) + \log\big(1+\frac{z}{1+z}\big) = 0.$$

Next, by some computation, we have $\phi(1/2) - \log(1/2)/2 < 1/2$, and our claim follows as desired. □

References

Amini, A. A. and Wainwright, M. J. (2009). High-dimensional analysis of semidefinite relaxations for sparse principal components. *Ann. Statist.* **37** 2877–2921.

Arora, S. and Barak, B. (2009). *Computational Complexity: A Modern Approach*. Cambridge University Press.

Berthet, Q. and Rigollet, P. (2013a). Complexity theoretic lower bounds for sparse principal component detection. *J. Mach. Learn. Res.* **30** 1046–1066 (electronic).

Berthet, Q. and Rigollet, P. (2013b). Optimal detection of sparse principal components in high dimension. *Ann. Statist.* **41** 1780–1815.

Bertsekas, D. P. (1999). *Nonlinear Programming*. Athena Scientific.

Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford.

Bullmore, E. and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience* **10** 186–198.

Bullmore, E. T. and Bassett, D. S. (2011). Brain graphs: graphical models of the human brain connectome. *Annual Review of Clinical Psychology* **7** 113–140.

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

Cao, L. and Fei-Fei, L. (2007). Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE.

d'Aspremont, A., El Ghaoui, L., Jordan, M. I. and Lanckriet, G. R. (2007). A direct formulation for sparse PCA using semidefinite programming. *SIAM Review* **49** 434–448.

Howard, A., Matarić, M. J. and Sukhatme, G. S. (2002). Mobile sensor network deployment using potential fields: A distributed, scalable solution to the area coverage problem. In *Distributed Autonomous Robotic Systems 5*. Springer, 299–308.

Kolar, M. and Liu, H. (2013). Optimal feature selection in high-dimensional discriminant analysis. *arXiv preprint arXiv:1306.6557* .

Krauthgamer, R., Nadler, B. and Vilenchik, D. (2013). Do semidefinite relaxations really solve sparse PCA? *arXiv preprint arXiv:1306.3690* .

Langendoen, K., Baggio, A. and Visser, O. (2006). Murphy loves potatoes: Experiences from a pilot sensor network deployment in precision agriculture. In — *Proceedings 20th IEEE International Parallel & Distributed Processing Symposium*. IEEE.

Lee, S. H., Lee, S., Song, H. and Lee, H. S. (2009). Wireless sensor network design for tactical military applications: remote large-scale environments. In *Military Communications Conference, 2009. MILCOM 2009. IEEE*. IEEE.

Lei, J. and Vu, V. Q. (2014). Sparsistency and agnostic inference in sparse PCA. *Ann. Statist., in press* .

Liu, H. and Wang, L. (2012). Tiger: A tuning-insensitive approach for optimally estimating gaussian graphical models. *arXiv preprint arXiv:1209.2437* .

Magazine, M. J. and Chern, M.-S. (1984). A note on approximation schemes for multidimensional knapsack problems. *Math. Oper. Res.* **9** 244–247.

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

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 417–473.

Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37** 246–270.

Oymak, S., Jalali, A., Fazel, M. and Hassibi, B. (2013). Noisy estimation of simultaneously structured models: Limitations of convex relaxation. In *IEEE CDC*. IEEE.

Pisinger, D. (1995). A minimal algorithm for the multiple-choice knapsack problem. *European Journal of Operational Research* **83** 394–410.

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

Starr, R. M. (1969). Quasi-equilibria in markets with non-convex preferences. *Econometrica* 25–38.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58** 267–288.

Trefethen, L. N. and Bau III, D. (1997). *Numerical Linear Algebra*. 50, SIAM.

Tropp, J. et al. (2004). Greed is good: Algorithmic results for sparse approximation. *Information Theory, IEEE Transactions on* **50** 2231–2242.

Vempala, S. S. (2005). *The Random Projection Method*, vol. 65. American Mathematical Society.

Vu, V. Q., Cho, J., Lei, J. and Rohe, K. (2013). Fantope projection and selection: A near-optimal convex relaxation of sparse PCA. In *NIPS*.

Williamson, D. P. and Shmoys, D. B. (2011). *The Design of Approximation Algorithms*. Cambridge University Press.

Xue, L., Zou, H. and Cai, T. (2012). Nonconcave penalized composite conditional likelihood estimation of sparse ising models. *Ann. Statist.* **40** 1403–1429.

Yick, J., Mukherjee, B. and Ghosal, D. (2008). Wireless sensor network survey. *Computer Networks* **52** 2292–2330.

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

Zhang, T. (2009). On the consistency of feature selection using greedy least squares regression. *J. Mach. Learn. Res.* **10** 555–568.

Zhang, Y., Wainwright, M. J. and Jordan, M. I. (2014). Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. *arXiv preprint arXiv:1402.1918* .

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563.