Inconsistency of Cross-Validation for Structure Learning in Gaussian Graphical Models

Zhao Lyu
University of Chicago

Wai Ming Tai
Nanyang Technological University

Mladen Kolar
University of Southern California

Bryon Aragam
University of Chicago

Abstract

Despite numerous years of research into the merits and trade-offs of various model selection criteria, obtaining robust results that elucidate the behavior of cross-validation remains a challenging endeavor. In this paper, we highlight the inherent limitations of cross-validation when employed to discern the structure of a Gaussian graphical model. We provide finite-sample bounds on the probability that the Lasso estimator for the neighborhood of a node within a Gaussian graphical model, optimized using a prediction oracle, misidentifies the neighborhood. Our results pertain to both undirected and directed acyclic graphs, encompassing general, sparse covariance structures. To support our theoretical findings, we conduct an empirical investigation of this inconsistency by contrasting our outcomes with other commonly used information criteria through an extensive simulation study. Given that many algorithms designed to learn the structure of graphical models require hyperparameter selection, the precise calibration of this hyperparameter is paramount for accurately estimating the inherent structure. Consequently, our observations shed light on this widely recognized practical challenge.

1 INTRODUCTION

Parameter tuning, also known as hyperparameter selection or model selection, is an unavoidable aspect of modern machine learning. For predictive tasks such as classification with deep neural networks, cross-validation is the gold standard for evaluating the performance of a model and tuning hyperparameters, and comes with its own set of practical challenges (Lei, 2020; Wilson et al., 2020; Bates et al., 2023). However, other applications, such as structure learning in graphical models, are not purely predictive in nature, and alternative criteria such as the Bayesian information criterion (BIC, Schwarz, 1978) and the Akaike information criterion (AIC, Akaike, 1974) are often used instead. Structure learning goes one step beyond prediction. Due to practical applications in causal inference, fairness, interpretability, and domain generalization, structure learning of undirected graphs (Cai et al., 2011; Friedman et al., 2008; Meinshausen and Bühlmann, 2006; Yuan and Lin, 2007) and directed graphs (Schmidt et al., 2007; Xiang and Kim, 2013; Fu and Zhou, 2013; Aragam and Zhou, 2015) has received renewed attention. Existing work suggests that predictive criteria, such as cross-validation, are not suitable for learning structure (learning the edge structure or the pattern of non-zero elements of the precision matrix) of a graphical model (Meinshausen and Bühlmann, 2006; Friedman and Yakhini, 1996). In particular, Meinshausen and Bühlmann (2006) (Proposition 1) proved that tuning the Lasso hyperparameter via a prediction oracle *provably* returns the wrong structure in the infinite-sample limit. This seminal result for Lasso-based graphical model estimators provides formal justification to the well-known folklore that choosing a model using predictive criteria may lead to undesirable overfitting. In practice, this leads to the question of which criterion to use for structure learning.

In this paper, we study the properties of cross-validation (CV) for structure learning of a graphical model using the Lasso to estimate the neighborhood of each node, extending the results of Meinshausen and Bühlmann (2006) to more general settings. We show

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

that in a precise sense cross-validation is inherently fallible and provide finite-sample bounds on the probability of structure inconsistency. As a motivating example, we will consider the special case of undirected Gaussian graphical models; however, our main result applies to neighborhood selection in general linear Gaussian models, and hence can be applied to directed acyclic graphs as well (Section 5). Thus, the main message of this paper can be summarized as follows:

> *For many structure learning applications, CV is provably inconsistent and alternative criteria should be used instead.*

While our particular results are specific to the Lasso, we note that this phenomenon is not specific to the Lasso: See Section 3 for a discussion of similar results for subset selection and bridge estimators.

Which alternative criteria should be used? This is an intriguing question, with numerous consistency results in the literature (Haughton, 1988; Foygel and Drton, 2010; Chen and Chen, 2008; Kim et al., 2012). To probe this question empirically, we conduct an extensive empirical study to compare the performance of different criteria. Our results indicate that extended BIC (Foygel and Drton, 2010) performs well, especially in high-dimensions (see Section 7).

While certain information criteria (IC) may be computationally challenging to evaluate—often necessitating the optimization of a nonconvex likelihood—CV is frequently proposed as a default substitute. However, our findings caution against such an approach. Despite the potential complexities in computing IC, resorting to CV as an alternative will lead to incorrect results, at least as far as structure learning is concerned.

Contributions We make the following contributions:

1. We provide finite-sample bounds on the probability that a Lasso estimate tuned with a prediction oracle will provably recover the *wrong* neighborhood in a linear Gaussian model (Theorem 1, Section 4);
2. We prove that CV indeed approximates this prediction oracle, which implies inconsistency of CV (Theorem 3, Section 4);
3. We apply these results to demonstrate inconsistency in structure learning for both undirected and directed acyclic graphical models (Corollaries 4 and 6, Section 5);
4. We provide an extensive simulation study comparing the virtues and tradeoffs of different parameter tuning strategies and algorithms (Section 7). These experiments confirm that the issues with CV are not restricted to the particular setting of our theoretical results and extend to other algorithms and non-Gaussian data.

2 GAUSSIAN GRAPHICAL MODELS AND STRUCTURE LEARNING

We will use the classical undirected Gaussian graphical model as a motivating example, but note that our results apply more generally (see Section 5). This preliminary section is intended to provide background and context for the structure learning problem; formal setup and details of our particular theoretical result can be found in Section 4.

Gaussian graphical models (GGMs) are widely used to represent and model statistical relations between variables. GGMs encode conditional independences with undirected graphs, which can also be read from the zero pattern in precision matrices (Lauritzen, 1996). GGMs have a wide range of applications in natural language processing (Manning and Schutze, 1999), computer vision (Cross and Jain, 1983), and computational biology (Menéndez et al., 2010; Varoquaux et al., 2010). We begin by recalling the definition of a GGM and then discuss the various learning tasks that one might consider in this model.

Let $X = \left[X_1, X_2, \ldots, X_p\right]^\top \sim \mathcal{N}(0, \Sigma)$ be p-dimensional joint Gaussian random vector with $\Sigma \succ 0$. The conditional independence relationships between each random variable can be represented by a graph $G = (V, E)$ on p nodes $V = X$ with edge set E. In particular, X_i and X_j are conditionally independent given all remaining variables $X_{\setminus\{i,j\}}$ if and only if $\Sigma^{-1}_{ij} = 0$, which corresponds to a missing edge between i and j (see, e.g., Lauritzen, 1996, for details). Thus, estimating the zero pattern of Σ^{-1} is equivalent to recovering E, a problem known as *structure learning*. Through its connection with neighborhood selection (Section 4), structure learning generalizes the variable selection and support recovery problems in classical regression models. To avoid complicating the presentation, we will not distinguish between these problems in the discussion.

It is worth comparing the different learning tasks in a GGM. *Prediction* refers to predicting the value of a particular node X_i given the values of the remaining nodes and is equivalent to linear regression. *Parameter estimation* refers to learning the precision matrix Σ^{-1} in some norm such as ℓ_2 or Frobenius. Both of these tasks are quite different from structure learning: One can predict X_i and/or estimate Σ^{-1} while at the same time getting the zero pattern of Σ^{-1} completely wrong

in any finite sample, and in general this is what happens in practice.

The distinction between prediction/estimation and structure learning is crucial when tuning hyperparameters, since the optimal choice of hyperparameter *depends on* what the learning goal is. This is well-known in the literature: For predictive tasks, CV/AIC are efficient (i.e. for prediction), but for structure learning, BIC is consistent (see, e.g., Arlot and Celisse, 2010, for detailed review of such results).[1] Notably, these results do not imply that CV is in fact *inconsistent* for structure learning, which is a stronger negative result for CV.

3 RELATED WORK

The related problems of hyperparameter tuning and model selection have been extensively studied in the machine learning and statistics literature, and we invite the reader to consult one of the many monographs on the subject for a detailed overview (Grünwald, 2007; Claeskens et al., 2008; Arlot and Celisse, 2010).

The study of model selection procedures such as CV, BIC, AIC, etc. dates back several decades (Mallows, 1973; Akaike, 1974; Stone, 1974; Geisser, 1975; Wahba and Wold, 1975; Stone, 1977; Schwarz, 1978; Efron, 1983; Picard and Cook, 1984; Herzberg and Tsukanov, 1986). It is well-known that BIC is consistent for structure learning in finite-dimensional models (Schwarz, 1978; Haughton, 1988) as well as high-dimensional models (Chen and Chen, 2008; Foygel and Drton, 2010; Kim et al., 2012). These results are central in the classical theory of structure learning for DAGs (Meek, 1997; Chickering and Meek, 2002; Chickering, 2003). We also mention the recent proposal to tune parameters in DAG models by Biza et al. (2020). At the same time, BIC can be inconsistent under misspecification Grünwald (2006, 2007); Grünwald and van Ommen (2017). On the other hand, AIC is known to select (possibly misspecified) models that are minimax optimal in estimation and/or prediction in a sense that can be made precise (see, e.g., Barron et al., 1999; Massart, 2007, and the references therein).

More relevant to the present work, Li (1987) and Shao (1993) studied the properties of CV and generalized CV (GCV). Li (1987) proved the *loss consistency* of CV, which is not the same as structure (or model selection) consistency, and more closely related to the minimax optimality results for AIC. To the best of our knowledge, the first proof of *structure (in)consistency* of CV appeared in Shao (1993) for a fixed design regression model using subset selection. Specifically, he showed that while leave-one-out CV (LOO) is inconsistent in selecting the true model, leave-k-out CV is consistent as long as $\frac{k}{n} \to 1$. Further work along these lines includes Zhang (1993); Shao (1997); Yang (2007). Meinshausen and Bühlmann (2006), Proposition 1, established the inconsistency of a prediction oracle for GGMs with a single edge; see also Meinshausen (2008). More recently, Chetverikov et al. (2021) recently showed that CV-tuned Lasso is minimax optimal for prediction and estimation, while for structure learning, Su et al. (2017) showed that false discoveries are asymptotically unavoidable. Later, Wang et al. (2020a) went beyond the Lasso and studied two-stage bridge estimators, and showed that the Lasso can be improved by two-stage approaches. Our results build upon these works in the Lasso setting and develops finite-sample bounds for arbitrary linear Gaussian models, with a focus on implications for graphical model structure learning.

Finally, while our paper is focused on provably *negative* consequences of CV, there is a long line of work proposing alternative tuning parameter selection methods. For example, tuning-parameter free approaches to structure learning abound (Wang et al., 2020b; Lederer and Müller, 2015; Yu and Bien, 2019; Belloni et al., 2011; Sun and Zhang, 2012; Chichignoud et al., 2016; Liu and Wang, 2017). Where structure learning is not the goal, the virtue of CV for predictive tasks is still a subject of intense study (see Wilson et al., 2020; Lei, 2020; Bates et al., 2023, and the references therein).

4 MAIN RESULTS

In this section, we present our setup and main theoretical result on the inconsistency of CV-tuned Lasso.

4.1 Neighborhood Selection

We first describe the formal setup for our main result. The neighborhood selection problem can be defined for a general linear model and does not require specific reference to a graphical model. In order to apply our main result to different types of graphs, we state our main result first in general, and then in Section 5 apply the general result to specific graphical models.

Let p be a positive integer and $[p] := \{1, 2, \ldots, p\}$. Let Σ be a p-by-p positive definite matrix and define Γ to be the $(p-1)$-by-$(p-1)$ submatrix such that $\Gamma_{i,j} = \Sigma_{i,j}$ for $i, j \in [p-1]$, v to be the $(p-1)$-dimensional vector such that $v_i = \Sigma_{i,p}$ for $i \in [p-1]$ and $a = \Sigma_{p,p}$, i.e.

$$\Sigma = \begin{bmatrix} \Gamma & v \\ v^\top & a \end{bmatrix}. \tag{1}$$

Let $X = \begin{bmatrix} X_1, X_2, \ldots, X_p \end{bmatrix}^\top \sim \mathcal{N}(0, \Sigma)$. The neighbor-

[1] More generally, structure learning is known as *model identification* or *model selection consistency*, i.e., selecting the correct model as opposed to an approximately correct one that obtains fast (e.g., minimax) rates of convergence.

hood selection problem seeks to learn the dependence of a target node in X on the rest of the observed variables. Without loss of generality, let the target node be X_p. Define

$$\theta^* := \arg\min_{\theta\in\mathbb{R}^{p-1}} \mathsf{E}_{X\sim\mathcal{N}(0,\Sigma)}(X_p - \sum_{j=1}^{p-1}\theta_j X_j)^2, \quad (2)$$
$$\mathsf{ne}^* := \{i \in [p-1] \mid \theta_i^* \neq 0\}.$$

It is easy to show that $\theta^* = \Gamma^{-1}v$. We assume that θ^* has s non-zero entries or equivalently $|\mathsf{ne}^*| = s$. The neighborhood selection problem attempts to recover ne^* from n i.i.d. observations of X, which we assemble into an $n \times p$ data matrix $\mathbf{X}$. For any matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ and $\lambda > 0$, the Lasso estimate with penalty parameter λ can be written as

$$\widehat{\theta}^{\lambda,\mathbf{X}} := \arg\min_{\theta\in\mathbb{R}^{p-1}} \frac{1}{2n}\|\mathbf{X}_p - \sum_{j=1}^{p-1}\theta_j\mathbf{X}_j\|_2^2 + \lambda\|\theta\|_1, \quad (3)$$

where $\mathbf{X}_i$ is the ith column of $\mathbf{X}$. The neighborhood estimated by the Lasso is defined by the non-zero entries of $\widehat{\theta}^{\lambda,\mathbf{X}}$, i.e.,

$$\widehat{\mathsf{ne}}^{\lambda,\mathbf{X}} := \left\{i \in [p-1] \mid \widehat{\theta}_i^{\lambda,\mathbf{X}} \neq 0\right\}.$$

See Meinshausen and Bühlmann (2006) for a more detailed review of the neighborhood selection problem.

Standard methods for selecting the penalty parameter λ include CV, BIC, and AIC, as discussed above. Here, we consider an oracle choice of penalty parameter, which reflects the limiting value of CV as $n \to \infty$ (see Theorem 3). For any matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$, the *oracle penalty* $\lambda_*^{\mathbf{X}}$ is defined as

$$\lambda_*^{\mathbf{X}} := \arg\min_{\lambda>0} \mathsf{E}_{Y\sim\mathcal{N}(0,\Sigma)}(Y_p - \sum_{j=1}^{p-1}\widehat{\theta}_j^{\lambda,\mathbf{X}}Y_j)^2. \quad (4)$$

Here, Y is a *new* sample from $\mathcal{N}(0,\Sigma)$, independent of the training data $\mathbf{X}$. This is known as the "oracle" penalty, as it involves the unknown data distribution. In practice, this value is approximated by CV (See Theorem 3 for a formal statement). For a fixed set of n samples $\mathbf{X}$, we shorten our notation to $\widehat{\theta}^{\lambda,\mathbf{X}}, \widehat{\mathsf{ne}}^{\lambda,\mathbf{X}}, \lambda_*^{\mathbf{X}}$ to $\widehat{\theta}^{\lambda}, \widehat{\mathsf{ne}}^{\lambda}, \lambda_*$ respectively if there is no ambiguity.

4.2 Inconsistency of Cross-validation

We would like to ask: When p and n are large, can we recover ne^* via the Lasso with the oracle penalty? Our first main result provides a finite-sample bound on the probability of exact recovery:

Theorem 1. *Let Σ be a p-by-p positive definite matrix such that $|\mathsf{ne}^*| = s$ for some positive integer s. Given a sample matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ where each row is an i.i.d. sample drawn from $\mathcal{N}(0,\Sigma)$, we have*

$$\mathsf{Pr}_{\mathbf{X}\sim\mathcal{N}(0,\Sigma)^n}(\widehat{\mathsf{ne}}^{\lambda_*} = \mathsf{ne}^*) < O\left(2^s \cdot \left(p^{-\Omega(\frac{1}{\kappa(\Sigma)})} + spe^{-\Omega(\frac{n}{s^2\kappa(\Sigma)^6})}\right)\right),$$

where $\kappa(\Sigma)$ is the condition number of Σ.

When the probability in Theorem 1 is strictly less than one, the prediction-oracle estimate is inconsistent. In particular, we have the following corollary, which answers the question in the negative in the sublinear sparsity regime:

Corollary 2. *Assume the setting in Theorem 1. Let $\delta \in [0,1)$ and $p > 1$. There exists a universal constant $C > 0$ such that if $s \leq C(\frac{1}{\kappa(\Sigma)}\log p - \log\frac{1}{\delta})$, then*

$$\mathsf{Pr}_{\mathbf{X}\sim\mathcal{N}(0,\Sigma)^n}(\widehat{\mathsf{ne}}^{\lambda_*} = \mathsf{ne}^*) < \delta \quad \text{as } n \to \infty.$$

Corollary 2 states that for sufficiently sparse graphs with $s = O(\log p)$, choosing λ via the prediction oracle is *provably inconsistent for structure learning*. With probability $1-\delta$, neighborhood selection will not recover the correct neighborhood in any non-trivial dimension p, even when p is fixed as $n \to \infty$. Corollary 2 is an immediate consequence of Theorem 1. Furthermore, this is just one possible example of inconsistency for certain choices of (n,p,s); clearly other configurations may lead to inconsistency as well.

In practice, the oracle penalty λ_* is unknown and is usually estimated by CV. For any positive integer K that divides n (for simplicity), let $I_1, I_2, \ldots, I_K$ be a partition of $[n]$ such that the size of each I_k is n/K for $k \in [K]$. For any matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$, the CV penalty $\lambda_{\mathrm{CV}}^{\mathbf{X}}$ is defined as

$$\lambda_{\mathrm{CV}}^{\mathbf{X}} := \arg\min_{\lambda>0} \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n/K}\|\mathbf{X}_p^{I_k} - \sum_{j=1}^{p-1}\widehat{\theta}_j^{\lambda,\mathbf{X}^{-I_k}}\mathbf{X}_j^{I_k}\|_2^2$$

where $\mathbf{X}^{I_k}$ (resp. $\mathbf{X}^{-I_k}$) is the (n/K)-by-p submatrix of $\mathbf{X}$ whose row indices are in I_k (resp. not in I_k) for $k \in [K]$. Although it is common to use $\lambda_{\mathrm{CV}}^{\mathbf{X}}$ to estimate $\lambda_*^{\mathbf{X}}$ in practice, we could not find a formal proof of this approximation in the literature. Therefore, we also prove the following theorem for completeness.

Theorem 3. *Let Σ be a p-by-p positive definite matrix. Suppose we are given a sample matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ where each row is an i.i.d sample drawn from $\mathcal{N}(0,\Sigma)$. Then, for every $\delta > 0$,*

$$\mathsf{Pr}_{\mathbf{X}\sim\mathcal{N}(0,\Sigma)^n}(|\lambda_{\mathrm{CV}}^{\mathbf{X}} - \lambda_*^{\mathbf{X}}| < \delta) \to 1 \quad \text{as } n \to \infty.$$

By combining Theorems 1 and 3 with the known properties of the solution path for the Lasso (Efron et al.,

2004), it is not hard to show that the CV-tuned neighborhoods are inconsistent, i.e.,

$$\mathsf{Pr}_{\mathbf{X}\sim\mathcal{N}(0,\Sigma)^n}(\widehat{\mathsf{ne}}^{\lambda_{\mathrm{CV}}^{\mathbf{X}},\mathbf{X}} = \mathsf{ne}^*) < \delta \quad \text{as } n \to \infty.$$

The number of folds K affects $\lambda_{\mathrm{CV}}^{\mathbf{X}}$ and further affects $\widehat{\mathsf{ne}}^{\lambda_{\mathrm{CV}}^{\mathbf{X}},\mathbf{X}}$. Briefly, if we only consider the dependence on n and K, the probability is roughly $1 - O(Ke^{-n^{<1}/K})$. Therefore, if $K = o(n)$ we have the probability $\to 1$ as $n \to \infty$. Details are postponed to proofs of Theorem 3.

5 APPLICATION TO GRAPHICAL MODELS

As stated above, our main result applies to neighborhood selection in a general linear Gaussian model with $X \sim \mathcal{N}(0, \Sigma)$. In this section, we apply this result to two important special cases: Undirected Gaussian graphical models and Gaussian DAG models.

5.1 Undirected Graphs

A popular approach to learning Gaussian graphical models is to directly apply neighborhood selection node-by-node, and use the neighborhood of each node to define a $p \times p$ graph (Meinshausen and Bühlmann, 2006). Let $\omega_j \in \mathbb{R}^p$ be the coefficient vector for the jth nodewise neighborhood regression problem, where $\omega_j = \begin{bmatrix}\omega_{1j}, \dots, \omega_{pj}\end{bmatrix}^\top$ for each j. Formally, ω_j solves (2) with p replaced by j (i.e. the target node is j), and we add a zero in the jth position. This defines a matrix $\Omega = [\omega_1 \,|\, \cdots \,|\, \omega_p] = \begin{bmatrix}\omega_{ij}\end{bmatrix} \in \mathbb{R}^{p\times p}$. The zero pattern of this matrix defines an undirected graph $G = (V, E)$, and is the same as the zero pattern of Σ^{-1}. We estimate Ω by $\widehat{\Omega}(\lambda) = [\widehat{\omega}_1(\lambda) \,|\, \cdots \,|\, \widehat{\omega}_p(\lambda)]$, where $\widehat{\Omega}(\lambda)$ is the solution to the following optimization problem:

$$\min_{\substack{\omega_1,\dots,\omega_p \\ \omega_j\in\mathbb{R}^{p-1}}} \frac{1}{2n}\sum_{j=1}^{p}\Big\{\|\mathbf{X}_j - \sum_{i\neq j}\omega_{ij}\mathbf{X}_i\|_2^2 + \lambda\|\omega_j\|_1\Big\}. \quad (5)$$

To estimate the structure G, we let $\widehat{G}(\lambda)$ be the undirected graph whose edges correspond to the nonzero entries in the solution $\widehat{\Omega}(\lambda)$ (see also Remark 5). It is easy to see that (5) is equivalent to solving p nodewise regression problems (3). This is also known as the *pseudo-likelihood* approach, since the objective is not a true (joint) likelihood. Nonetheless, it is well-known to provide a consistent estimate of the structure of G for certain choices of λ. Finally, let $\widehat{G}_{\mathrm{CV}} = \widehat{G}(\lambda_{\mathrm{CV}})$ be the estimate when CV is used to tune λ.

The following corollary is immediate from Theorems 1 and 3:

Corollary 4. *Suppose $\mathbf{X} \in \mathbb{R}^{n\times p}$ is a sample matrix where each row is an i.i.d. sample drawn from $\mathcal{N}(0, \Sigma)$ and let G be the undirected Gaussian graphical model associated with Σ^{-1}. Then, for any $\delta > 0$ satisfying the conditions in Corollary 2,*

$$\mathsf{Pr}_{\mathbf{X}\sim\mathcal{N}(0,\Sigma)^n}(\widehat{G}_{\mathrm{CV}} \neq G) < \delta \quad \textit{as } n \to \infty.$$

Thus, CV is inconsistent for learning the structured of an undirected Gaussian graphical model.

Remark 5. *Since Ω and $\widehat{\Omega}$ essentially capture partial regression coefficients, these matrices are not symmetric in general. Nonetheless, the support of Ω is always symmetric (see e.g. Sec 5.1.3 in Lauritzen, 1996), but $\widehat{\Omega}$ may not have a symmetric support on finite samples. Asymptotically, this does change anything, but on finite-samples we need to use either the AND or the OR rule to symmetrize $\widehat{\Omega}$, as discussed in Meinshausen and Bühlmann (2006).*

5.2 Directed Acyclic Graphs

Our results also apply to DAG models, which are popular for modeling causal relationships in ML. First, recall the general linear structural equation model (SEM):

$$\begin{aligned} X_j &= \sum_{i=1}^{p}\beta_{ij}X_i + \varepsilon_j, \quad \varepsilon_j \sim \mathcal{N}(0,\sigma_j^2), \\ E &= \{(i,j) : \beta_{ij} \neq 0\}. \end{aligned} \quad (6)$$

We collect the SEM coefficients β_{ij} into a $p \times p$ matrix $B = [\beta_1 \,|\, \cdots \,|\, \beta_p] = (\beta_{ij}) \in \mathbb{R}^{p\times p}$, with the same indexing conventions as Ω in Section 5.1. This defines a graph $G = (V, E)$ that be read off from the nonzero entries in B. When G is a DAG, (6) defines a Gaussian DAG model. We assume throughout that G is acyclic.

A common procedure to learn a DAG is to first learn a topological ordering of G, and then regress each node onto its predecessors in this ordering (e.g. Shojaie and Michailidis, 2010; Ghoshal and Honorio, 2017, 2018; Chen et al., 2019; Park and Kim, 2020). More precisely, given an ordering $\prec$ on the variables X_i, we define an SEM (6) by regressing each X_j onto the set $A_j = \{X_i : X_i \prec X_j\}$. The set of nonzero coefficients $\{i : \beta_{ij} \neq 0\}$ defines the parents of X_j in the ordering $\prec$.

Following this literature, let $\widehat{G}(\prec, \lambda)$ denote the estimate of G that results from using the order $\prec$ and ℓ_1-regularized least squares with $\lambda > 0$ to estimate each parent set from the candidate set A_j:

$$\min_{\substack{\beta_1,\dots,\beta_p \\ \beta_j\in\mathbb{R}^{|A_j|}}} \frac{1}{2n}\sum_{j=1}^{p}\Big\{\|\mathbf{X}_j - \sum_{i\in A_j}\beta_{ij}\mathbf{X}_i\|_2^2 + \lambda\|\beta_j\|_1\Big\}. \quad (7)$$

For each j, we are solving a neighborhood regression problem similar to (3), except instead of regressing the

jth node onto every other variable, we restrict attention to the candidate variables A_j induced by the ordering $\prec$. Finally, let $\widehat{G}_{\mathrm{CV}}(\prec) = \widehat{G}(\prec, \lambda_{\mathrm{CV}})$ be the resulting graph when CV is used to tune λ.

The following corollary is also immediate from Theorems 1 and 3:

Corollary 6. *Suppose we are given n i.i.d. samples from the model (6) with DAG G, and suppose further that we know the true ordering $\prec$ of G. Then, for any $\delta > 0$ satisfying the conditions in Corollary 2,*

$$\mathsf{Pr}_{\mathbf{X}\sim\mathcal{N}(0,\Sigma)^n}(\widehat{G}_{\mathrm{CV}}(\prec) \neq G) < \delta \quad \text{as } n \to \infty.$$

Thus, even if we know the true ordering, CV will return the wrong DAG. If we do not know the true ordering, this result says that $\widehat{G}_{\mathrm{CV}}(\prec)$ is an inconsistent estimate of the minimal I-map corresponding to $\prec$ (see Lauritzen, 1996 for definitions). Of course, assuming everything else is equal, structure learning with unknown ordering is at least as difficult as with a known ordering.

6 PROOF OVERVIEW

In this section, we outline the main idea of the proof of Theorem 1. Detailed proofs of both Theorem 1 and 3 are deferred to the supplementary materials.

We start with some observations. Without loss of generality, we assume that $\mathsf{ne}^* = [s]$. For any matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$, $i \in [p-1]$, and $\theta \in \mathbb{R}^{p-1}$, define

$$G_{\mathbf{X},i}(\theta) := \frac{1}{n}\langle \mathbf{X}_p - \sum_{j=1}^{p-1}\theta_j \mathbf{X}_j, \mathbf{X}_i\rangle. \tag{8}$$

We can write the classical KKT conditions for the Lasso as follows:

$$\begin{cases} G_{\mathbf{X},i}(\theta) = \mathsf{sign}(\theta_i)\lambda, & \text{for } \theta_i \neq 0 \\ |G_{\mathbf{X},i}(\theta)| \leq \lambda, & \text{for } \theta_i = 0. \end{cases} \tag{9}$$

Then θ satisfies (9) if and only if $\theta = \widehat{\theta}^{\lambda}$ is a Lasso solution for λ. Moreover, if we define an ellipsoid $\mathcal{E}$ by

$$\mathcal{E} := \{(\theta-\theta^*)^\top \Gamma(\theta-\theta^*) \leq (\widehat{\theta}^{\lambda_*}-\theta^*)^\top\Gamma(\widehat{\theta}^{\lambda_*}-\theta^*)\},$$

we can show (Appendix C.1) that any point $\theta \in \mathcal{E}$ cannot be a Lasso solution for any penalty $\lambda > 0$. It is easy to see that the Lasso solution for the oracle penalty λ_*, $\widehat{\theta}^{\lambda_*}$, lies on the boundary of this ellipsoid.

To prove Theorem 1, we want to argue that the event $\widehat{\mathsf{ne}}^{\lambda_*} = \mathsf{ne}^* = [s]$ is unlikely. Let

$$\widehat{\mathsf{ne}}^G = \left\{ i \in [p-1] \mid |G_{\mathbf{X},i}(\widehat{\theta}^{\lambda_*})| = \lambda_* \right\}. \tag{10}$$

We have $\widehat{\mathsf{ne}}^{\lambda_*} \subseteq \widehat{\mathsf{ne}}^G$ by the KKT conditions (9). We further argue in Appendix C.2 that $\widehat{\mathsf{ne}}^G$ has at most one extra element almost surely, and without loss of generality, we may assume that $\widehat{\mathsf{ne}}^G = [s]$ or $\widehat{\mathsf{ne}}^G = [s+1]$. We will consider these two cases separately. Namely, we will bound the following probability:

$$\begin{aligned}&\mathsf{Pr}_{\mathbf{X}\sim\mathcal{N}(0,\Sigma)^n}(\widehat{\mathsf{ne}}^{\lambda_*} = [s]) \\ &\quad = \mathsf{Pr}_{\mathbf{X}\sim\mathcal{N}(0,\Sigma)^n}(\widehat{\mathsf{ne}}^{\lambda_*} = \widehat{\mathsf{ne}}^G = [s]) \\ &\qquad + \mathsf{Pr}_{\mathbf{X}\sim\mathcal{N}(0,\Sigma)^n}(\widehat{\mathsf{ne}}^{\lambda_*} = [s] \wedge \widehat{\mathsf{ne}}^G = [s+1])\end{aligned}$$

Case I: $\widehat{\mathsf{ne}}^G = [s]$. The first step is to show that there exists a line passing through $\widehat{\theta}^{\lambda_*}$ such that any point in the intersection of a small neighborhood of $\widehat{\theta}^{\lambda_*}$ and this line is also a Lasso solution for some penalty λ. See Figure 1 for an illustration of the following argument.

We will do this by defining a line L and show that it satisfies the KKT conditions (9). Consider the following system of equations:

$$\begin{cases} G_{\mathbf{X},i}(\theta) = \mathsf{sign}(\widehat{\theta}_i^{\lambda_*})\lambda & \text{for } i \in [s] \\ \theta_i = 0 & \text{for } i \notin [s] \end{cases} \tag{11}$$

Geometrically, we can view these $p-1$ equations, which are linear in θ and λ, as hyperplanes in the θ-λ space which is a p-dimensional space. It turns out that the intersection of these $p-1$ hyperplanes forms a line almost surely, which is the desired line L. Since $\widehat{\theta}^{\lambda_*}$ clearly is a solution of (11) by the definition of $\widehat{\theta}^{\lambda_*}$, we can write L as

$$L := \left\{ \widehat{\theta}^{\lambda_*} + \delta\theta' \mid \delta \in \mathbb{R} \right\} \tag{12}$$

for some $\theta' \in \mathbb{R}^p$. We will define θ' formally in (19) in the Supplementary Material.

Consider any point $\widetilde{\theta} = \widehat{\theta}^{\lambda_*} + \delta\theta' \in L$ for sufficiently small $|\delta|$. We will check $\widetilde{\theta}$ satisfies the KKT conditions (9). If $|\delta|$ is sufficiently small, $\widetilde{\theta}_i \neq 0$ for $i \in [s]$ since $\widehat{\theta}_i^{\lambda_*} \neq 0$ for $i \in [s]$. By the construction of the system (11), it ensures that all $G_{\mathbf{X},i}(\widetilde{\theta})$ remain equal in magnitude for $i \in [s]$ and the signs are consistent, i.e. $\mathsf{sign}(\widetilde{\theta}_i) = \mathsf{sign}(G_{\mathbf{X},i}(\widetilde{\theta}))$ for $i \in [s]$. It means that $\widetilde{\theta}$ satisfies the first condition in (9). On the other hand, we set $\theta'_i = 0$ for $i \notin [s]$ to ensure $\widetilde{\theta}_i \neq 0$ for $i \notin [s]$. Recall the definition of $\widehat{\mathsf{ne}}^G$, we have $|G_{\mathbf{X},i}(\widehat{\theta}^{\lambda_*})|$ *strictly* less than λ_* for $i \notin [s]$. If $|\delta|$ is sufficiently small, it ensures that $|G_{\mathbf{X},j}(\widetilde{\theta})| \leq |G_{\mathbf{X},i}(\widetilde{\theta})|$ for $i \in [s]$ and $j \notin [s]$. It means that $\widetilde{\theta}$ satisfies the second condition in (9). Hence, for a sufficiently small $|\delta|$, $\widetilde{\theta}$ is a Lasso solution.

Now, if the line L is not a tangent line of $\mathcal{E}$ at $\widehat{\theta}^{\lambda_*}$, then some point in L must be inside the ellipsoid $\mathcal{E}$. This contradicts the observation that no Lasso solution

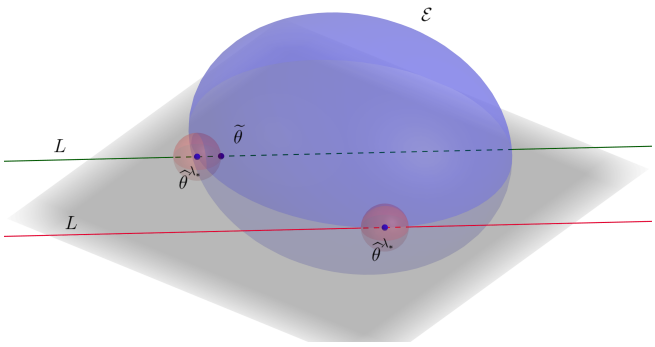

Figure 1: Illustration of the ellipsoid $\mathcal{E}$ and the line L when $p-1=3$ and $\widehat{\mathsf{ne}}^{\lambda_*} = \widehat{\mathsf{ne}}^{G} = \{1,2\}$. (green) There exists another Lasso solution $\widetilde{\theta}$ inside $\mathcal{E}$ when it is not a tangent line. (red) No Lasso solution can be found inside $\mathcal{E}$ when it is a tangent line.

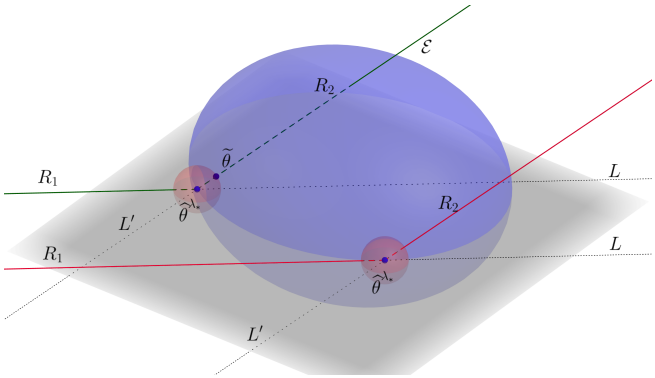

Figure 2: Illustration of the ellipsoid $\mathcal{E}$ and the rays of intersection when $p-1=3$, $\widehat{\mathsf{ne}}^{\lambda_*} = \{1,2\}$ and $\widehat{\mathsf{ne}}^{G} = \{1,2,3\}$. (green) There exists another Lasso solution $\widetilde{\theta}$ inside $\mathcal{E}$ when one of the rays shoots into $\mathcal{E}$. (red) No Lasso solution can be found inside $\mathcal{E}$ when both rays shoot out of $\mathcal{E}$.

can be inside $\mathcal{E}$. Therefore, L must be a tangent line. From here, we can explicitly bound the probability of L being a tangent line and hence also the probability $\mathsf{Pr}_{\mathbf{X}\sim\mathcal{N}(0,\Sigma)^n}(\widehat{\mathsf{ne}}^{\lambda_*} = \widehat{\mathsf{ne}}^{G} = [s])$.

Case II: $\widehat{\mathsf{ne}}^{G} = [s+1]$. In this case, instead of a line, there are two rays shooting from $\widehat{\theta}^{\lambda_*}$ such that any point in the intersection of a small neighborhood of $\widehat{\theta}^{\lambda_*}$ and these two rays is also a Lasso solution for some penalty λ. See Figure 2.

In this case, we have $|G_{\mathbf{X},s+1}(\widehat{\theta}^{\lambda_*})| = \lambda_*$. If we follow the same argument as in the previous case, we can no longer establish $|G_{\mathbf{X},s+1}(\widetilde{\theta})| \le |G_{\mathbf{X},i}(\widetilde{\theta})|$ for $i \in [s]$ *no matter how small* $|\delta|$ *is.* Namely, the second condition in (9) does not hold. Fortunately, it turns out to only cause problems for one of $\delta \ge 0$ or $\delta \le 0$. We can indeed view the line L in (12) as two rays shooting from $\widehat{\theta}^{\lambda_*}$ in the opposite directions which correspond to the cases of $\delta \ge 0$ or $\delta \le 0$. That means the argument in the case of $\widehat{\mathsf{ne}}^{G} = [s]$ still holds for one of these two rays which we denote by R_1. Hence, one of the desired two rays is R_1.

We now only have one side of L which is R_1. The argument of constructing a Lasso solution inside $\mathcal{E}$ when L is not a tangent line does not hold because R_1 probably does not intersect $\mathcal{E}$ (except $\widehat{\theta}^{\lambda_*}$) even when R_1 is not a tangent ray. It is intuitive that there is another ray R_2 instead of the other side of L that any point in the intersection of a small neighborhood of $\widehat{\theta}^{\lambda_*}$ and R_2 is a Lasso solution.

To define another ray R_2, consider the following system of equations:

$$\begin{cases} G_{\mathbf{X},i}(\theta) = \mathsf{sign}(G_{\mathbf{X},i}(\widehat{\theta}^{\lambda_*}))\lambda, & \text{for } i \in [s+1] \\ \theta_i = 0, & \text{for } i \notin [s+1]. \end{cases} \tag{13}$$

By a similar argument as in the previous case, the intersection of these hyperplanes forms a line passing through $\widehat{\theta}^{\lambda_*}$ almost surely which we can write as

$$L' := \left\{ \widehat{\theta}^{\lambda_*} + \delta\theta'' \mid \delta \in \mathbb{R} \right\} \tag{14}$$

for some $\theta'' \in \mathbb{R}^p$. We will define θ'' formally in (20) in the Supplementary Material.

Consider any point $\widetilde{\theta} = \widehat{\theta}^{\lambda_*} + \delta\theta'' \in L'$ for a sufficiently small $|\delta|$. For $i \ne s+1$, $\widetilde{\theta}_i$ satisfies the first or second condition of the KKT conditions (9) accordingly by a similar analysis as in the previous case. For $i = s+1$, $\widetilde{\theta}_{s+1} = \widehat{\theta}^{\lambda_*}_{s+1} + \delta\theta''_{s+1} = \delta\theta''_{s+1}$ is no longer 0 and we need to check if it satisfies the first condition in (9). By the construction of the system (13), it ensures that all $G_{\mathbf{X},i}(\widetilde{\theta})$ remain equal in magnitude for $i \in [s+1]$. We also need to check the sign consistency, i.e. $\mathsf{sign}(\widetilde{\theta}_{s+1}) = \mathsf{sign}(G_{\mathbf{X},s+1}(\widetilde{\theta}))$. We again view L' in (14) as two rays shooting from $\widehat{\theta}^{\lambda_*}$ in the opposite directions. It turns out that only one of them ensures this sign consistency and we choose this ray as the desired second ray R_2.

Now, if one of the rays R_1, R_2 shoots into (i.e. intersects) $\mathcal{E}$, this would contradict the observation that no Lasso solution can be inside $\mathcal{E}$. Therefore, both rays R_1, R_2 shoot out of $\mathcal{E}$. As before, we can bound the probability that both R_1, R_2 shoot out of $\mathcal{E}$ and hence the probability $\mathsf{Pr}_{\mathbf{X}\sim\mathcal{N}(0,\Sigma)^n}(\widehat{\mathsf{ne}}^{\lambda_*} = [s] \wedge \widehat{\mathsf{ne}}^{G} = [s+1])$.

Combining these two cases, we obtain an explicit bound on the probability $\mathsf{Pr}_{\mathbf{X}\sim\mathcal{N}(0,\Sigma)^n}(\widehat{\mathsf{ne}}^{\lambda_*} = [s])$.

7 EXPERIMENTS

In this section, we demonstrate through simulations the failure of CV for structure learning, verifying our main

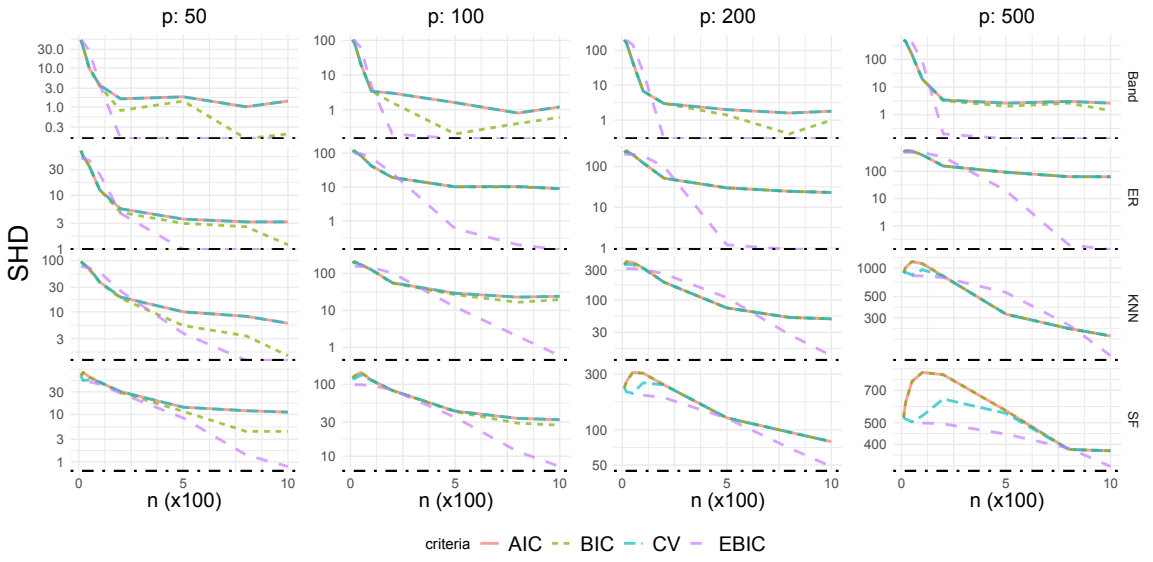

Figure 3: SHD vs. n (in hundreds) and p on different graph types using NS to compare criteria. The black dot-dash line represents zero SHD, i.e. perfect neighborhood selection.

theoretical results. We include here only a snapshot of our results to convey the main point; complete details and results from our exhaustive experiments can be found in Appendix D, including additional experiments on non-Gaussian data.

We use the CV-selected λ_{CV} to approximate λ_* and compare its neighborhood estimate with those selected by several commonly used criteria: Akaike information criterion (AIC) Akaike (1974), Bayesian information criterion (BIC) Schwarz (1978), and Extended Bayesian information criteria (EBIC) Foygel and Drton (2010). The performance is evaluated via

1. The average structural hamming distance (SHD) to measure the number of incorrectly identified neighbors;

2. The average true positive rate (TPR) and false discovery rate (FDR).

We let the number of observations n and the number of observed variables p to vary independently so one can be greatly larger than another. To validate comparisons between criteria, we simulate four different graphs: the Band graph, Scale-Free (SF), Erdös-Rényi (ER) and K-Nearest Neighbor (KNN) graphs. Besides the neighborhood Selection (NS), we also use three popular algorithms for Lasso estimators: Graphical lasso (Glasso) Friedman et al. (2008), constrained ℓ_1-minimization for inverse matrix estimation (CLIME) Cai et al. (2011) and Tuning-Insensitive Graph Estimation and Regression (TIGER) Liu and Wang (2017). They are implemented in the `glasso` and `flare` packages for R.

Results We focus on NS here so as to corroborate our main theoretical results. Exceeding wall time limit (three hours) or undefined FDR value for all zero estimates are marked as missing points for plotting. Performance results with Glasso, Clime and Tiger are postponed to the supplementary materials, with similar conclusions.

In Figure 3, as expected, we see the number of incorrectly identified neighbors decreasing with increasing sample size n given a fixed p. As p increases, the task becomes increasingly harder, which is reflected by greater average SHD. The average SHD of CV always decreases with increasing n but never reaches zero, regardless of how large n gets. This pattern persists for Glasso, Clime and Tiger as well (see the supplementary materials). This confirms Theorem 1. Besides CV, AIC performs poorly as well and never reaches zero average SHD. On the other hand, BIC achieves smaller average SHD, but its performance in high-dimensions is unsatisfactory, which is consistent with known results for BIC (e.g. Mestres et al., 2018). The only candidate with constant decreasing trend with increasing n for all p is EBIC. Specifically, EBIC is always the first one to get closest to or reach zero, regardless of p.

The correctness of EBIC is more obvious when comparing averaged FDR in Figure 4 and 5. Unsurprisingly, we see TPR gradually reaches 100%, depending on the graph type. Specifically, CV and AIC are the first to reach 100% TPR, while EBIC falls behind. However,

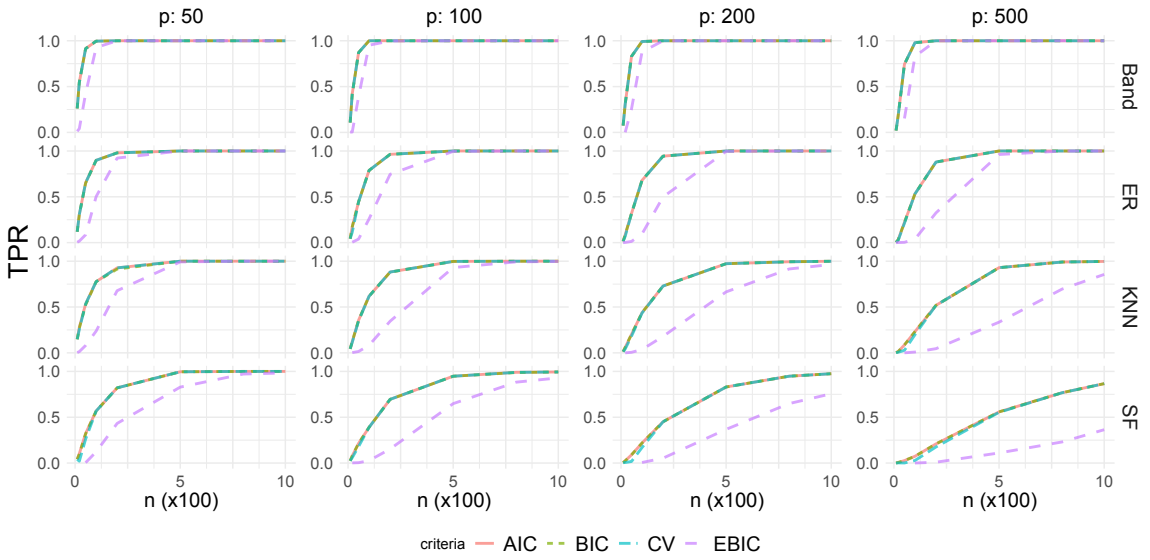

Figure 4: TPR vs. sample size n (in hundreds) and p on different graph types with NS to compare criteria.

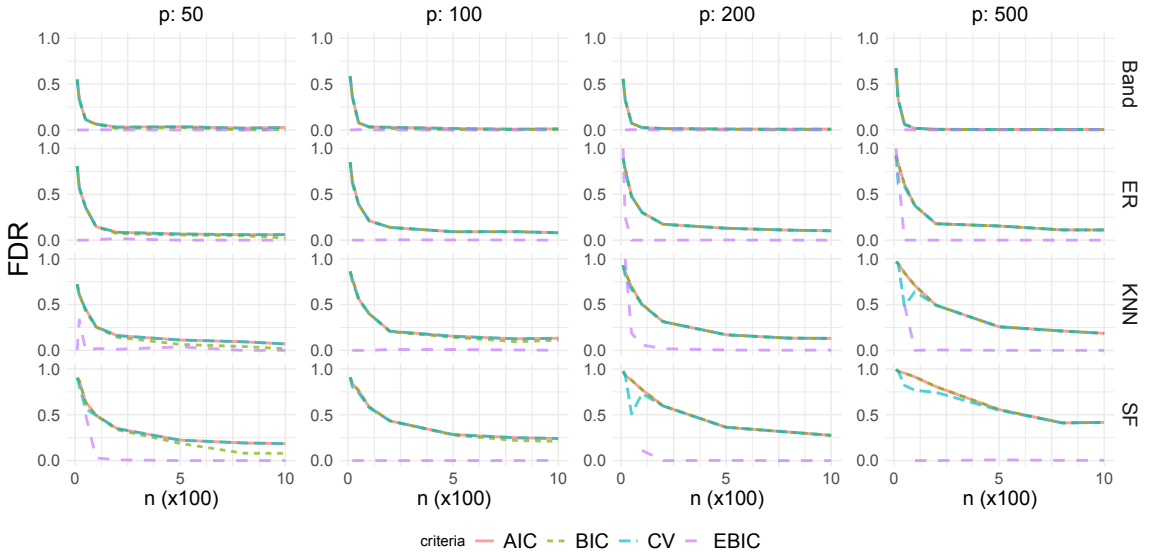

Figure 5: FDR vs. sample size n (in hundreds) and p on different graph types with NS to compare criteria.

this is not whole story: The FDR for CV is problematic, and fails to get near 0% on average. (For the Band graph, we provide a more detailed numerical FDR summary via NS tuned by CV in the supplementary materials.) Moreover, CV fails in average FDR in all other algorithms (see the supplementary materials) as well.

8 CONCLUSION

Cross-validation is the parameter selection criterion of choice in most ML applications, however, its suitability for structure learning problems is not well understood outside of empirical observations. To address this gap, we proved that for a general family of Gaussian graphical models, including DAG models, CV is provably inconsistent for learning the structure of a graph. This shows that using CV as a naive alternative to difficult-to-implement selection criteria is ill-advised. It would be of interest to extend our proofs to non-Gaussian models, where our experiments indeed suggest CV is still inconsistent. On the positive side, our experiments indicate that EBIC is robust across a wide range of settings.

Acknowledgements

The research of MK is supported in part by NSF Grant ECCS-2216912. BA was supported by NSF IIS-1956330, NIH R01GM140467, and the Robert H. Topel Faculty Research Fund at the University of Chicago Booth School of Business. WT was supported by the Singapore MOE AcRF Tier 2 grant MOE-T2EP20122-0001.

References

H. Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19 (6):716–723, 1974.

B. Aragam and Q. Zhou. Concave Penalized Estimation of Sparse Gaussian Bayesian Networks. *Journal of Machine Learning Research*, 16(69):2273–2328, 2015.

S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4: 40–79, 2010.

A. A. Azzalini. *The R package* `sn`*: The skew-normal and related distributions such as the skew-t and the SUN (version 2.1.1).* Università degli Studi di Padova, Italia, 2023. URL `https://cran.r-project.org/package=sn`. Home page: `http://azzalini.stat.unipd.it/SN/`.

A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability theory and related fields*, 113(3):301–413, 1999.

S. Bates, T. Hastie, and R. Tibshirani. Cross-validation: what does it estimate and how well does it do it? *Journal of the American Statistical Association*, (just-accepted):1–22, 2023.

A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

K. Biza, I. Tsamardinos, and S. Triantafillou. Tuning causal discovery algorithms. In *International Conference on Probabilistic Graphical Models*, pages 17–28. PMLR, 2020.

T. Cai, W. Liu, and X. Luo. A constrained l 1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106 (494):594–607, 2011.

J. Chen and Z. Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

W. Chen, M. Drton, and Y. S. Wang. On causal discovery with an equal-variance assumption. *Biometrika*, 106(4):973–980, 2019.

D. Chetverikov, Z. Liao, and V. Chernozhukov. On cross-validated lasso in high dimensions. *The Annals of Statistics*, 49(3):1300–1317, 2021.

M. Chichignoud, J. Lederer, and M. J. Wainwright. A practical scheme and fast algorithm to tune the lasso with optimality guarantees. *The Journal of Machine Learning Research*, 17(1):8162–8181, 2016.

D. M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2003.

D. M. Chickering and C. Meek. Finding optimal Bayesian networks. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 94–102. Morgan Kaufmann Publishers Inc., 2002.

G. Claeskens, N. L. Hjort, et al. Model selection and model averaging. *Cambridge Books*, 2008.

G. R. Cross and A. K. Jain. Markov random field texture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (1):25–39, 1983.

G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. URL `https://igraph.org`.

B. Efron. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American statistical association*, pages 316–331, 1983.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32 (2):407–451, 2004.

R. Foygel and M. Drton. Extended bayesian information criteria for gaussian graphical models. *Advances in neural information processing systems*, 23, 2010.

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

J. Friedman, R. Tibshirani, and T. Hastie. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1): 1–22, 2010. doi: 10.18637/jss.v033.i01.

J. Friedman, T. Hastie, and R. Tibshirani. *glasso: Graphical Lasso: Estimation of Gaussian Graphical Models*, 2019. URL `https://CRAN.R-project.org/package=glasso`. R package version 1.11.

N. Friedman and Z. Yakhini. On the sample complexity of learning bayesian networks. In *Uncertainty in Artifical Intelligence (UAI)*, 02 1996.

F. Fu and Q. Zhou. Learning sparse causal Gaussian networks with experimental intervention: Regularization and coordinate descent. *Journal of the American Statistical Association*, 108(501):288–300, 2013.

S. Geisser. The predictive sample reuse method with applications. *Journal of the American statistical Association*, pages 320–328, 1975.

A. Ghoshal and J. Honorio. Learning identifiable gaussian bayesian networks in polynomial time and sample complexity. *Advances in Neural Information Processing Systems*, 30, 2017.

A. Ghoshal and J. Honorio. Learning linear structural equation models in polynomial time and sample complexity. In *International Conference on Artificial Intelligence and Statistics*, pages 1466–1475. PMLR, 2018.

P. Grünwald and T. van Ommen. Inconsistency of bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4): 1069–1103, 2017.

P. D. Grünwald. Bayesian inconsistency under misspecification. In *Four page abstract of a plenary presentation at the Valencia 8 ISBA conference on Bayesian statistics*, 2006.

P. D. Grünwald. *The minimum description length principle*. MIT press, 2007.

D. M. Haughton. On the choice of a model to fit data from an exponential family. *Annals of Statistics*, 16 (1):342–355, 1988.

A. M. Herzberg and A. Tsukanov. A note on modifications of the jackknife criterion for model selection. *Utilitas Math*, 29:209–216, 1986.

Jorge Parraga-Alava, Pablo Moscato, and Mario Inostroza-Ponta. *mstknnclust: MST-kNN Clustering Algorithm*, 2023. URL `https://CRAN.R-project.org/package=mstknnclust`. R package version 0.3.2.

Y. Kim, S. Kwon, and H. Choi. Consistent model selection criteria on high dimensions. *The Journal of Machine Learning Research*, 13:1037–1057, 2012.

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

J. Lederer and C. Müller. Don't fall for tuning parameters: tuning-free variable selection in high dimensions with the trex. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

J. Lei. Cross-validation with confidence. *Journal of the American Statistical Association*, 115(532):1978–1997, 2020.

K.-C. Li. Asymptotic optimality for cp, cl, cross-validation and generalized cross-validation: discrete index set. *The Annals of Statistics*, pages 958–975, 1987.

X. Li, T. Zhao, L. Wang, X. Yuan, and H. Liu. *flare: Family of Lasso Regression*, 2020. URL `https://CRAN.R-project.org/package=flare`. R package version 1.7.0.

H. Liu and L. Wang. Tiger: A tuning-insensitive approach for optimally estimating gaussian graphical models. 2017.

C. L. Mallows. Some comments on cp. *Technometrics*, 15(4):661–675, 1973. ISSN 00401706. URL `http://www.jstor.org/stable/1267380`.

C. Manning and H. Schutze. *Foundations of statistical natural language processing*. MIT press, 1999.

P. Massart. *Concentration inequalities and model selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003*. Springer, 2007.

C. Meek. *Graphical Models: Selecting causal and statistical models*. PhD thesis, Carnegie Mellon University, 1997.

N. Meinshausen. A note on the lasso for gaussian graphical model selection. *Statistics & Probability Letters*, 78(7):880–884, 2008.

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

P. Menéndez, Y. A. Kourmpetis, C. J. ter Braak, and F. A. van Eeuwijk. Gene regulatory networks from multifactorial perturbations using graphical lasso: application to the dream4 challenge. *PloS one*, 5(12): e14147, 2010.

A. C. Mestres, N. Bochkina, and C. Mayer. Selection of the regularization parameter in graphical models using network characteristics. *Journal of Computational and Graphical Statistics*, 27(2):323–333, 2018.

G. Park and Y. Kim. Identifiability of gaussian linear structural equation models with homogeneous and heterogeneous error variances. *Journal of the Korean Statistical Society*, pages 1–17, 2020.

R. R. Picard and R. D. Cook. Cross-validation of regression models. *Journal of the American Statistical Association*, pages 575–583, 1984.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL `https://www.R-project.org/`.

M. Schmidt, A. Niculescu-Mizil, and K. Murphy. Learning graphical model structure using L1-regularization paths. In *AAAI*, volume 7, pages 1278–1283, 2007.

G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

J. Shao. Linear model selection by cross-validation. *Journal of the American statistical Association*, 88 (422):486–494, 1993.

J. Shao. An asymptotic theory for linear model selection. *Statistica sinica*, pages 221–242, 1997.

A. Shojaie and G. Michailidis. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika*, 97(3):519–538, 2010.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*, 36(2):111–133, 1974.

M. Stone. An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):44–47, 1977.

W. Su, M. Bogdan, and E. Candes. False discoveries occur early on the lasso path. *The Annals of statistics*, pages 2133–2150, 2017.

T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

G. Varoquaux, A. Gramfort, J.-B. Poline, and B. Thirion. Brain covariance selection: better individual functional connectivity models using population prior. *Advances in neural information processing systems*, 23, 2010.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL `https://www.stats.ox.ac.uk/pub/MASS4/`. ISBN 0-387-95457-0.

G. Wahba and S. Wold. A completely automatic french curve: fitting spline functions by cross validation. *Communications in Statistics-Theory and Methods*, 4(1):1–17, 1975.

S. Wang, H. Weng, and A. Maleki. Which bridge estimator is the best for variable selection? *The Annals of Statistics*, 48(5):2791 – 2823, 2020a. doi: 10.1214/19-AOS1906. URL `https://doi.org/10.1214/19-AOS1906`.

Y. Wang, U. Roy, and C. Uhler. Learning high-dimensional gaussian graphical models under total positivity without adjustment of tuning parameters. In *International Conference on Artificial Intelligence and Statistics*, pages 2698–2708. PMLR, 2020b.

A. Wilson, M. Kasy, and L. Mackey. Approximate cross-validation: Guarantees for model assessment and selection. In *International Conference on Artificial Intelligence and Statistics*, pages 4530–4540. PMLR, 2020.

J. Xiang and S. Kim. A* Lasso for learning a sparse Bayesian network structure for continuous variables. In *Advances in Neural Information Processing Systems*, pages 2418–2426, 2013.

Y. Yang. Consistency of cross validation for comparing regression procedures. 2007.

G. Yu and J. Bien. Estimating the error variance in a high-dimensional linear model. *Biometrika*, 106(3): 533–546, 2019.

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

P. Zhang. Model selection via multifold cross validation. *The annals of statistics*, pages 299–313, 1993.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator If your work uses existing assets. [Yes]

(b) The license information of the assets, if applicable. [Not Applicable]

(c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

(d) Information about consent from data providers/curators. [Not Applicable]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

A PROOF OF THEOREM 1

In this section, we will prove Theorem 1. Detailed proof of various supporting lemmas can be found in Appendix C. Before we go into the detail, we present some general observations.

We first state a useful lemma from Meinshausen and Bühlmann (2006) which we will use often: These are the well-known KKT conditions for the Lasso solution. We first define the following notation. For any matrix $\mathbf{X}$ and $i \in [p-1]$, we define

$$G_{\mathbf{X},i}(\theta) := \frac{1}{n}\langle \mathbf{X}_p - \sum_{j=1}^{p-1} \theta_j \mathbf{X}_j, \mathbf{X}_i \rangle. \tag{15}$$

Lemma 7 (KKT conditions, Meinshausen and Bühlmann (2006))**.** *For any matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ and $\theta \in \mathbb{R}^{p-1}$, we have the following KKT conditions. For any $\lambda > 0$,*

$$\theta = \widehat{\theta}^{\lambda}$$

if and only if

$$\begin{cases} \text{if } \theta_i \neq 0, \text{ then } G_{\mathbf{X},i}(\theta) = \mathsf{sign}(\theta_i)\lambda \\ \text{if } \theta_i = 0, \text{ then } |G_{\mathbf{X},i}(\theta)| \leq \lambda \end{cases} \tag{16}$$

Recall the definition of Γ in (3), and define an ellipsoid by

$$\mathcal{E} := \{\theta \in \mathbb{R}^{p-1} : (\theta - \theta^*)^\top \Gamma (\theta - \theta^*) \leq (\widehat{\theta}^{\lambda_*} - \theta^*)^\top \Gamma (\widehat{\theta}^{\lambda_*} - \theta^*)\}.$$

It is clear that the Lasso solution for the oracle penalty λ_*, $\widehat{\theta}^{\lambda_*}$, lies on the boundary of this ellipsoid. Our next observation is that for any matrix $\mathbf{X}$, any point $\theta \in \mathbb{R}^{p-1}$ inside the ellipsoid $\mathcal{E}$ cannot be a Lasso solution for any penalty $\lambda > 0$:

Lemma 8. *For any matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ and $\lambda > 0$, if $\theta \in \mathbb{R}^{p-1}$ satisfies*

$$(\theta - \theta^*)^\top \Gamma (\theta - \theta^*) < (\widehat{\theta}^{\lambda_*} - \theta^*)^\top \Gamma (\widehat{\theta}^{\lambda_*} - \theta^*)$$

then $\theta \neq \widehat{\theta}^{\lambda}$.

The proof of this lemma can be found in Section C.1.

By Lemma 7, for $i \in \widehat{\mathsf{ne}}^{\lambda_*}$, we have $|G_{\mathbf{X},i}(\widehat{\theta}^{\lambda_*})| = \lambda_*$. Let $\widehat{\mathsf{ne}}^{G}$ be the set

$$\widehat{\mathsf{ne}}^{G} = \left\{ i \in [p-1] \mid |G_{\mathbf{X},i}(\widehat{\theta}^{\lambda_*})| = \lambda_* \right\}. \tag{17}$$

Note that $\widehat{\mathsf{ne}}^{\lambda_*} \subseteq \widehat{\mathsf{ne}}^{G}$ by Lemma 7. The following lemma shows that $\widehat{\mathsf{ne}}^{G}$ has at most one more extra element almost surely.

Lemma 9. *Let $\mathbf{X} \in \mathbb{R}^{n\times p}$ be a sample matrix where each row is an i.i.d. sample drawn from $\mathcal{N}(0, \Sigma)$. Suppose the number of samples $n > |\widehat{\mathsf{ne}}^{\lambda_*}| + 2$. Then, we have $|\widehat{\mathsf{ne}}^{G}| \leq |\widehat{\mathsf{ne}}^{\lambda_*}| + 1$ almost surely.*

The proof of this lemma can be found in Section C.2.

To prove Theorem 1, we want to argue that the event $\widehat{\mathsf{ne}}^{\lambda_*} = \mathsf{ne}^*$ is unlikely to happen. Without loss of generality, we assume that $\mathsf{ne}^* = [s]$ for some integer s. If we assume $\mathsf{ne}^* = \widehat{\mathsf{ne}}^{\lambda_*}$, then we have $\widehat{\mathsf{ne}}^{\lambda_*} = [s]$. Also, since $|\widehat{\mathsf{ne}}^G|$ is either $|\widehat{\mathsf{ne}}^{\lambda_*}|$ or $|\widehat{\mathsf{ne}}^{\lambda_*}| + 1$ and $\widehat{\mathsf{ne}}^{\lambda_*} \subseteq \widehat{\mathsf{ne}}^G$, we assume $\widehat{\mathsf{ne}}^G = [s+1]$ if $\widehat{\mathsf{ne}}^G \neq \widehat{\mathsf{ne}}^{\lambda_*}$. We will consider these two cases separately. Namely, we need to bound the following probability

$$\begin{aligned}
&\mathsf{Pr}_{\mathbf{X}\sim\mathcal{N}(0,\Sigma)^n}(\widehat{\mathsf{ne}}^{\lambda_*} = \mathsf{ne}^*) \\
&= \mathsf{Pr}_{\mathbf{X}\sim\mathcal{N}(0,\Sigma)^n}(\widehat{\mathsf{ne}}^{\lambda_*} = [s]) \\
&= \mathsf{Pr}_{\mathbf{X}\sim\mathcal{N}(0,\Sigma)^n}(\widehat{\mathsf{ne}}^{\lambda_*} = \widehat{\mathsf{ne}}^G = [s]) + \mathsf{Pr}_{\mathbf{X}\sim\mathcal{N}(0,\Sigma)^n}(\widehat{\mathsf{ne}}^{\lambda_*} = [s] \wedge \widehat{\mathsf{ne}}^G = [s+1]).
\end{aligned} \tag{18}$$

We preview that the first term can be bounded by $O\big(2^s \cdot \big(p^{-\Omega(\frac{1}{\kappa(\Gamma)})} + spe^{-\Omega(\frac{n}{s^2\kappa(\Gamma)^3})}\big)\big)$ (cf. Section A.1) and the second term can be bounded by $O\big(2^s \cdot \big(p^{-\Omega(\frac{1}{\kappa(\Gamma)})} + spe^{-\Omega(\frac{n}{s^2\kappa(\Gamma)^6})}\big)\big)$ (cf. Section A.2). By plugging these into (18),

$$\begin{aligned}
\mathsf{Pr}_{\mathbf{X}\sim\mathcal{N}(0,\Sigma)^n}(\widehat{\mathsf{ne}}^{\lambda_*} = \mathsf{ne}^*) &\leq O\left(2^s \cdot \left(p^{-\Omega(\frac{1}{\kappa(\Gamma)})} + spe^{-\Omega(\frac{n}{s^2\kappa(\Gamma)^3})}\right)\right) + O\left(2^s \cdot \left(p^{-\Omega(\frac{1}{\kappa(\Gamma)})} + spe^{-\Omega(\frac{n}{s^2\kappa(\Gamma)^6})}\right)\right) \\
&\leq O\left(2^s \cdot \left(p^{-\Omega(\frac{1}{\kappa(\Gamma)})} + spe^{-\Omega(\frac{n}{s^2\kappa(\Gamma)^6})}\right)\right)
\end{aligned}$$

we finish the proof of Theorem 1.

A.1 Case of $\widehat{\mathsf{ne}}^{\lambda_*} = \widehat{\mathsf{ne}}^G = [s]$

In this case, the main idea is to argue that there exists a line passing through $\widehat{\theta}^{\lambda_*}$ such that any point in the intersection of a small neighborhood of $\widehat{\theta}^{\lambda_*}$ and this line is also a Lasso solution for some penalty λ.

To help the construction of this line, we first define the following notations. For any set of n samples $\mathbf{X}$, let $\widehat{\Gamma}$ be the $(p-1)$-by-$(p-1)$ matrix that (r,c)-entry is $\frac{1}{n}\langle \mathbf{X}_r, \mathbf{X}_c\rangle$ for any $r, c \in [p-1]$ and $\widehat{\Gamma}_{[s]}$ be its s-by-s submatrix where the indices are in $[s]$. Also, let $q_{[s]}$ be the s-dimensional vector that the i-th entry is $\mathsf{sign}(\widehat{\theta}_i^{\lambda_*})$ for $i \in [s]$. We claim that the following line L satisfies the above condition.

$$L := \left\{\widehat{\theta}^{\lambda_*} + \delta\theta' \mid \delta \in \mathbb{R}\right\}$$

where θ' is the vector such that the i-the entry of θ' is

$$\theta'_i = \begin{cases} -(\widehat{\Gamma}_{[s]}^{-1} q_{[s]})_i & \text{for } i \in [s] \\ 0 & \text{for } i \notin [s]. \end{cases} \tag{19}$$

The following lemma suggests that any θ in the intersection of a small neighborhood of $\widehat{\theta}^{\lambda_*}$ and the line L is also a Lasso solution for some penalty $\lambda > 0$.

Lemma 10. *For a sufficiently small $|\delta|$, suppose $\widetilde{\theta} = \widehat{\theta}^{\lambda_*} + \delta\theta' \in L$ where θ' is defined in* (19)*. Then, $\widetilde{\theta}$ is a Lasso solution for some penalty $\lambda > 0$.*

The proof of this lemma can be found in Section C.3.

The main idea is to use Lemma 7 and check the KKT conditions. Since $\widehat{\theta}^{\lambda_*}$ by definition is a Lasso solution for λ_*, $\widehat{\theta}^{\lambda_*}$ satisfies (16) or we have

$$\frac{1}{n}\langle \mathbf{X}_p - \sum_{j=1}^{p-1} \widehat{\theta}_j^{\lambda_*} \mathbf{X}_j, \mathbf{X}_i\rangle = \mathsf{sign}(\widehat{\theta}_i^{\lambda_*})\lambda_* \qquad \text{for } i \in [s].$$

Obviously, by the assumption of $\widehat{\mathsf{ne}}^{\lambda_*} = [s]$, it also satisfies

$$\widehat{\theta}_i^{\lambda_*} = 0 \qquad \text{for } i \notin [s].$$

If we consider the following system of linear equations with $\theta \in \mathbb{R}^{p-1}$ and $\lambda \in \mathbb{R}$ as variables

$$\frac{1}{n}\langle \mathbf{X}_p - \sum_{j=1}^{p-1} \theta_j \mathbf{X}_j, \mathbf{X}_i \rangle = \mathsf{sign}(\widehat{\theta}^{\lambda_*}_i)\lambda \qquad \text{for } i \in [s]$$
$$\theta_i = 0 \qquad \text{for } i \notin [s],$$

we can check that $(\widehat{\theta}^{\lambda_*} + \delta\theta', \lambda_* + \delta)$ satisfies this system of linear equations for all $\delta \in \mathbb{R}$. Furthermore, for a sufficiently small $|\delta|$, it satisfies (16) in Lemma 7. Hence, Lemma 10 follows.

Recall that, by Lemma 8, all points inside the ellipsoid $\mathcal{E}$ cannot be a Lasso solution for any penalty. If the line L is not a tangent of $\mathcal{E}$ at $\widehat{\theta}^{\lambda_*}$, there exists a point inside $\mathcal{E}$ such that it is in the intersection of a neighborhood of $\widehat{\theta}^{\lambda_*}$ and the line L. It contradicts Lemma 8 and hence L has to be the tangent of $\mathcal{E}$ at $\widehat{\theta}^{\lambda_*}$. Note that the normal vector of the tangent space of $\mathcal{E}$ at $\widehat{\theta}^{\lambda_*}$ is $\Gamma(\widehat{\theta}^{\lambda_*} - \theta^*)$. Then, L is the tangent line of $\mathcal{E}$ at $\widehat{\theta}^{\lambda_*}$ if and only if ${\theta'}^\top \Gamma(\widehat{\theta}^{\lambda_*} - \theta^*) = 0$. In other words, we have

$$\widehat{\mathsf{ne}}^{\lambda_*} = \widehat{\mathsf{ne}}^{G} = [s] \implies \theta'\Gamma(\widehat{\theta}^{\lambda_*} - \theta^*) = 0.$$

Let θ^Δ be $\widehat{\theta}^{\lambda_*} - \theta^*$. Recall that if $i \notin [s]$ then $\theta^\Delta_i = 0$ from the event $\widehat{\mathsf{ne}}^{\lambda_*} = [s]$ and $\theta'_i = (\widehat{\Gamma}^{-1}q)_i = 0$ from the construction in (19). Namely, we can rewrite the expression $\theta'\Gamma\theta^\Delta$ as

$$\theta'\Gamma\theta^\Delta = q_{[s]}^\top \widehat{\Gamma}^{-1}_{[s]} \Gamma_{[s]} \theta^\Delta_{[s]}.$$

By Lemma 15, we have

$$\mathsf{Pr}_{\mathbf{X}\sim\mathcal{N}(0,\Sigma)^n}(\widehat{\mathsf{ne}}^{\lambda_*} = [s] \wedge q_{[s]}^\top \widehat{\Gamma}^{-1}_{[s]} \Gamma_{[s]} \theta^\Delta_{[s]} = 0) \le O\left(2^s \cdot \left(p^{-\Omega(\frac{1}{\kappa(\Gamma)})} + spe^{-\Omega(\frac{n}{s^2\kappa(\Gamma)^3})}\right)\right)$$

Note that the event $q_{[s]}^\top \widehat{\Gamma}^{-1}_{[s]} \Gamma_{[s]} \theta^\Delta_{[s]} = 0$ is more restrictive and indeed implies $|q_{[s]} \widehat{\Gamma}^{-1}_{[s]} \Gamma_{[s]} \theta^\Delta_{[s]}| \le \frac{1}{100\sqrt{s\sigma_{\max}(\Gamma)}} \sqrt{{\theta^\Delta_{[s]}}^\top \Gamma_{[s]} \theta^\Delta_{[s]}}$. In other words, we have

$$\mathsf{Pr}_{\mathbf{X}\sim\mathcal{N}(0,\Sigma)^n}(\widehat{\mathsf{ne}}^{\lambda_*} = \widehat{\mathsf{ne}}^{G} = [s]) < O\left(2^s \cdot \left(p^{-\Omega(\frac{1}{\kappa(\Gamma)})} + spe^{-\Omega(\frac{n}{s^2\kappa(\Gamma)^3})}\right)\right).$$

A.2 Case of $\widehat{\mathsf{ne}}^{\lambda_*} = [s] \wedge \widehat{\mathsf{ne}}^{G} = [s+1]$

In this case, the main idea is to argue that there exist two rays shooting from $\widehat{\theta}^{\lambda_*}$ such that any point in the intersection of a small neighborhood of $\widehat{\theta}^{\lambda_*}$ and these two rays is also a Lasso solution for some penalty λ.

To help the construction of this line, we define the following notations similar to the notations in last subsection. For any set of n samples $\mathbf{X}$, let $\widehat{\Gamma}$ be the $(p-1)$-by-$(p-1)$ matrix that (r,c)-entry is $\frac{1}{n}\langle \mathbf{X}_r, \mathbf{X}_c \rangle$ for any $r, c \in [p-1]$ and $\widehat{\Gamma}_{[s+1]}$ be its $(s+1)$-by-$(s+1)$ submatrix where the indices are in $[s+1]$. Also, let $q_{[s+1]}$ be the s-dimensional vector that the i-th entry is $\mathsf{sign}(G_{\mathbf{X},i}(\widehat{\theta}^{\lambda_*}))$ for $i \in [s+1]$ where $G_{\mathbf{X},i}$ is defined in (15). Note that $q_i = \mathsf{sign}(\widehat{\theta}^{\lambda_*}_i)$ when $i \in [s]$. We claim that the following rays R_1, R_2 satisfy the above condition.

$$R_1 := \left\{\widehat{\theta}^{\lambda_*} + \delta \cdot \mathsf{sign}(Q)\theta' \mid \delta \ge 0\right\}$$

and

$$R_2 := \left\{\widehat{\theta}^{\lambda_*} + \delta \cdot \mathsf{sign}(Q)\theta'' \mid \delta < 0\right\}$$

where

$$\theta''_i = \begin{cases} -(\widehat{\Gamma}^{-1}_{[s+1]} q_{[s+1]})_i & \text{for } i \in [s+1] \\ 0 & \text{for } i \notin [s+1] \end{cases} \tag{20}$$

and

$$Q = 1 - q_{s+1} \sum_{j=1}^{s} \theta'_j \cdot \frac{1}{n} \langle \mathbf{X}_j, \mathbf{X}_{s+1} \rangle. \tag{21}$$

Recall that θ' is defined in (19). Here we abuse the notation that $\mathsf{sign}(Q) = +1$ even when $Q = 0$.

The following lemma shows that any θ in the intersection of a small neighborhood of $\widehat{\theta}^{\lambda_*}$ and the union of these two rays $R_1 \cup R_2$ is also a Lasso solution for some penalty $\lambda > 0$.

Lemma 11. *For a sufficiently small $|\delta|$, suppose $\widetilde{\theta} = \widehat{\theta}^{\lambda_*} + \delta \cdot \mathsf{sign}(Q)\theta' \in R_1$ for $\delta \geq 0$ and $\widetilde{\theta} = \widehat{\theta}^{\lambda_*} + \delta \cdot \mathsf{sign}(Q)\theta'' \in R_2$ for $\delta < 0$ where θ' and θ'' are defined in* (19) *and* (20) *respectively. Then, $\widetilde{\theta}$ is a Lasso solution for some penalty $\lambda > 0$.*

The proof of this lemma can be found in Section C.4.

The main idea is to use Lemma 7 and check the KKT conditions like Lemma 10. When we consider $\widetilde{\theta} \in R_1$, the analysis is the same as in Lemma 10 except that we need to check R_1 shoots in the correct direction. When we consider $\widetilde{\theta} \in R_2$, the analysis is similar. The key difference is that we need to check the sign of $\widetilde{\theta}_{s+1}$ matches the sign of $G_{\mathbf{X},s+1}(\widetilde{\theta})$ since $\widehat{\theta}^{\lambda_*}_{s+1} = 0$ which may cause a sign mismatch in the perturbation $\widetilde{\theta}$.

Recall that, by Lemma 8, all points inside the ellipsoid $\mathcal{E}$ cannot be a Lasso solution for any penalty. If the rays R_1, R_2 are shooting inside $\mathcal{E}$, there exists a point inside $\mathcal{E}$ such that it is in the intersection of a neighborhood of $\widehat{\theta}^{\lambda_*}$ and the union of two rays $R_1 \cup R_2$. It contradicts Lemma 8 and hence R_1, R_2 have to not shoot into $\mathcal{E}$. Note that the normal vector of the tangent space of $\mathcal{E}$ at $\widehat{\theta}^{\lambda_*}$ is $\Gamma(\widehat{\theta}^{\lambda_*} - \theta^*)$. Also, the direction of the ray R_1 is $\mathsf{sign}(Q)\theta'$ and the direction of the ray R_2 is $-\mathsf{sign}(Q)\theta''$. Then, R_1, R_2 do not shoot into $\mathcal{E}$ if and only if $\mathsf{sign}(Q){\theta'}^\top \Gamma(\widehat{\theta}^{\lambda_*} - \theta^*) \geq 0$ and $-\mathsf{sign}(Q){\theta''}^\top \Gamma(\widehat{\theta}^{\lambda_*} - \theta^*) \geq 0$. The following lemma shows that the probability of R_1, R_2 not shooting into $\mathcal{E}$ is small. In other words, we have

$$\widehat{\mathsf{ne}}^{\lambda_*} = [s] \wedge \widehat{\mathsf{ne}}^{G} = [s+1] \implies \mathcal{A} \wedge \mathcal{B}$$

where $\mathcal{A}$ is the event of $\mathsf{sign}(Q){\theta'}^\top \Gamma(\widehat{\theta}^{\lambda_*} - \theta^*) \geq 0$ and $\mathcal{B}$ is the event of $-\mathsf{sign}(Q){\theta''}^\top \Gamma(\widehat{\theta}^{\lambda_*} - \theta^*) \geq 0$. By Lemma 16, we have

$$\mathsf{Pr}_{\mathbf{X} \sim \mathcal{N}(0,\Sigma)^n}(\widehat{\mathsf{ne}}^{\lambda_*} = [s] \wedge \mathcal{A} \wedge \mathcal{B}) \leq O\left(2^s \cdot \left(p^{-\Omega(\frac{1}{\kappa(\Gamma)})} + spe^{-\Omega(\frac{n}{s^2\kappa(\Gamma)^6})}\right)\right).$$

In other words, we have

$$\mathsf{Pr}_{\mathbf{X} \sim \mathcal{N}(0,\Sigma)^n}(\widehat{\mathsf{ne}}^{\lambda_*} = [s] \wedge \widehat{\mathsf{ne}}^{G} = [s+1]) < O\left(2^s \cdot \left(p^{-\Omega(\frac{1}{\kappa(\Gamma)})} + spe^{-\Omega(\frac{n}{s^2\kappa(\Gamma)^6})}\right)\right).$$

B PROOF OF THEOREM 3

We shorten $\lambda^{\mathbf{X}}_{\mathrm{CV}}$ to be λ_{CV} if there is no ambiguity.

For any $\lambda > 0$, define

$$Q(\lambda, \mathbf{X}) := \frac{1}{2} \mathsf{E}_{Y \sim \mathcal{N}(0,\Sigma)}(Y_p - \sum_{j=1}^{p-1} \widehat{\theta}^{\lambda,\mathbf{X}}_j Y_j)^2.$$

We observe that, by the definition of λ_*,

$$Q(\lambda_*, \mathbf{X}) \leq Q(\lambda_{\mathrm{CV}}, \mathbf{X}).$$

If we manage to prove that

$$Q(\lambda_{\mathrm{CV}}, \mathbf{X}) \leq Q(\lambda_*, \mathbf{X}) + \varepsilon_n$$

where ε_n is a value that $\varepsilon_n \to 0$ as $n \to \infty$, then we can prove $|\lambda_{\mathrm{CV}} - \lambda_*| \to 0$ by the fact that Q is a continuous function.

Before we go into the detail, we first state the following useful inequality. By Chernoff bound and union bound, for any $\varepsilon' > 0$, we have

$$|\mathsf{E}(Y_i Y_j) - \frac{1}{n/K}\langle \mathbf{X}_i^{I_k}, \mathbf{X}_j^{I_k}\rangle| < \varepsilon' \qquad \text{for all } i,j \in [p-1] \text{ and } k \in [K] \tag{22}$$

with probability $1 - O(p^2 K e^{-\Omega(n\varepsilon'^2/K\sigma_{\max}(\Sigma)^2)})$. From now on, our analysis is conditioned on (22).

For any $\lambda > 0$ and any $\mathbf{Z} \in \mathbb{R}^{n\times p}, \mathbf{X} \in \mathbb{R}^{m\times p}$, define

$$Q_{\mathbf{Z}}(\lambda, \mathbf{X}) := \frac{1}{2n}\|\mathbf{Z}_p - \sum_{j=1}^{p-1} \widehat{\theta}_j^{\lambda,\mathbf{X}} \mathbf{Z}_j\|^2.$$

Lemma 12. *For any $\lambda > 0$ and any $\mathbf{Z} \in \mathbb{R}^{n\times p}, \mathbf{X} \in \mathbb{R}^{m\times p}$, we have*

$$|Q(\lambda, \mathbf{X}) - Q_{\mathbf{Z}}(\lambda, \mathbf{X})| = O(\varepsilon' \cdot (1 + \|\widehat{\theta}^{\lambda,\mathbf{X}}\|_1)^2).$$

The proof of this lemma can be found in Section C.5.

Lemma 13. *For any $\lambda > 0$ and any matrix $\mathbf{X} \sim \mathcal{N}(0, \Sigma)$, we have*

$$\|\widehat{\theta}^{\lambda,\mathbf{X}}\|_1 \le O(\sqrt{p}\kappa(\Sigma))$$

with probability $1 - O(e^{-\Omega(n)})$. Recall that $\kappa(\Sigma)$ is the condition number of Σ.

The proof of this lemma can be found in Section C.6.

By Lemma 12 with $\lambda = \lambda_{\mathrm{CV}}$ and $\mathbf{Z} = \mathbf{X}$, we first have

$$Q(\lambda_{\mathrm{CV}}, \mathbf{X}) \le Q_{\mathbf{X}}(\lambda_{\mathrm{CV}}, \mathbf{X}) + O(\varepsilon' \cdot (1 + \|\widehat{\theta}^{\lambda_{\mathrm{CV}},\mathbf{X}}\|_1)^2).$$

Furthermore, by Lemma 13, we have

$$Q(\lambda_{\mathrm{CV}}, \mathbf{X}) \le Q_{\mathbf{X}}(\lambda_{\mathrm{CV}}, \mathbf{X}) + O(\varepsilon' p\kappa(\Sigma)^2). \tag{23}$$

For the term $Q_{\mathbf{X}}(\lambda_{\mathrm{CV}}, \mathbf{X})$, we further have

$$\begin{aligned}
&Q_{\mathbf{X}}(\lambda_{\mathrm{CV}}, \mathbf{X}) \\
&\quad = Q_{\mathbf{X}}(\lambda_{\mathrm{CV}}, \mathbf{X}) + \lambda_{\mathrm{CV}}\|\widehat{\theta}^{\lambda_{\mathrm{CV}},\mathbf{X}}\|_1 - \lambda_{\mathrm{CV}}\|\widehat{\theta}^{\lambda_{\mathrm{CV}},\mathbf{X}}\|_1 \\
&\quad \le Q_{\mathbf{X}}(\lambda_{\mathrm{CV}}, \mathbf{X}^{-I_k}) + \lambda_{\mathrm{CV}}\|\widehat{\theta}^{\lambda_{\mathrm{CV}},\mathbf{X}^{-I_k}}\|_1 - \lambda_{\mathrm{CV}}\|\widehat{\theta}^{\lambda_{\mathrm{CV}},\mathbf{X}}\|_1 \\
&\quad \le Q_{\mathbf{X}^{I_k}}(\lambda_{\mathrm{CV}}, \mathbf{X}^{-I_k}) + O(\varepsilon' \cdot (1 + \|\widehat{\theta}^{\lambda_{\mathrm{CV}},\mathbf{X}^{-I_k}}\|_1)^2) + \lambda_{\mathrm{CV}}\|\widehat{\theta}^{\lambda_{\mathrm{CV}},\mathbf{X}^{-I_k}}\|_1 - \lambda_{\mathrm{CV}}\|\widehat{\theta}^{\lambda_{\mathrm{CV}},\mathbf{X}}\|_1 \\
&\quad \le Q_{\mathbf{X}^{I_k}}(\lambda_{\mathrm{CV}}, \mathbf{X}^{-I_k}) + O(\varepsilon' p\kappa(\Sigma)^2) + \lambda_{\mathrm{CV}}(\|\widehat{\theta}^{\lambda_{\mathrm{CV}},\mathbf{X}^{-I_k}}\|_1 - \|\widehat{\theta}^{\lambda_{\mathrm{CV}},\mathbf{X}}\|_1)
\end{aligned} \tag{24}$$

for $k \in [K]$. The penultimate inequality is due to the fact that $\widehat{\theta}^{\lambda_{\mathrm{CV}},\mathbf{X}} = \arg\min_{\theta\in\mathbb{R}^{p-1}} \frac{1}{2n}\|\mathbf{X}_p - \sum_{j=1}^{p-1}\theta_j \mathbf{X}_j\|^2 + \lambda_{\mathrm{CV}}\|\theta\|_1$ and the last inequality is due to applying Lemma 12 twice and triangle inequality.

In (24), the first term $Q_{\mathbf{X}^{I_k}}(\lambda_{\mathrm{CV}}, \mathbf{X}^{-I_k})$ is what we are looking for. By the definition of λ_{CV}, observe that

$$\frac{1}{K}\sum_{k=1}^{K} Q_{\mathbf{X}^{I_k}}(\lambda_{\mathrm{CV}}, \mathbf{X}^{-I_k}) \le \frac{1}{K}\sum_{k=1}^{K} Q_{\mathbf{X}^{I_k}}(\lambda_*, \mathbf{X}^{-I_k}). \tag{25}$$

Moreover, for each term $Q_{\mathbf{X}^{I_k}}(\lambda_*, \mathbf{X}^{-I_k})$, we have

$$\begin{aligned}
&Q_{\mathbf{X}^{I_k}}(\lambda_*, \mathbf{X}^{-I_k}) \\
&\quad = Q_{\mathbf{X}^{-I_k}}(\lambda_*, \mathbf{X}^{-I_k}) + O(\varepsilon' p\kappa(\Sigma)^2) \\
&\quad \le Q_{\mathbf{X}^{-I_k}}(\lambda_*, \mathbf{X}) + \lambda_*\|\widehat{\theta}^{\lambda_*,\mathbf{X}}\|_1 - \lambda_*\|\widehat{\theta}^{\lambda_*,\mathbf{X}^{-I_k}}\|_1 + O(\varepsilon' p\kappa(\Sigma)^2) \\
&\quad \le Q(\lambda_*, \mathbf{X}) + O(\varepsilon' p\kappa(\Sigma)^2) + \lambda_*(\|\widehat{\theta}^{\lambda_*,\mathbf{X}}\|_1 - \|\widehat{\theta}^{\lambda_*,\mathbf{X}^{-I_k}}\|_1)
\end{aligned} \tag{26}$$

If we plug (26) into (25) and further plug it and (24) into (23), we have

$$Q(\lambda_{\mathrm{CV}}, \mathbf{X}) \le Q(\lambda_*, \mathbf{X}) + O(\varepsilon' p \kappa(\Sigma)^2) + \lambda_{\mathrm{CV}}(\|\widehat{\theta}^{\lambda_{\mathrm{CV}}, \mathbf{X}^{-I_k}}\|_1 - \|\widehat{\theta}^{\lambda_{\mathrm{CV}}, \mathbf{X}}\|_1) + \lambda_* \cdot \frac{1}{K}\sum_{k=1}^{K}(\|\widehat{\theta}^{\lambda_*, \mathbf{X}}\|_1 - \|\widehat{\theta}^{\lambda_*, \mathbf{X}^{-I_k}}\|_1).$$

Lemma 14. *For any $\lambda > 0$ and any matrices $\mathbf{X}, \mathbf{Z} \sim \mathcal{N}(0, \Sigma)$, we have*

$$|\|\widehat{\theta}^{\lambda, \mathbf{X}}\|_1 - \|\widehat{\theta}^{\lambda, \mathbf{Z}}\|_1| \le O(\sqrt{\varepsilon' \cdot \frac{p^2\kappa(\Sigma)^2}{\sigma_{\min}(\Sigma)}}) \qquad \text{and} \qquad \lambda \le O(\sqrt{p}\sigma_{\max}(\Sigma)\kappa(\Sigma))$$

as long as λ is not too large such that $\widehat{\theta}^{\lambda} \neq 0$.

The proof of this lemma can be found in Section C.7.

By Lemma 14, we have

$$Q(\lambda_{\mathrm{CV}}, \mathbf{X}) \le Q(\lambda_*, \mathbf{X}) + \sqrt{\varepsilon'} \cdot \mathsf{poly}(p) \cdot C_\Sigma$$

where C_Σ is a constant depending only on Σ. Taking $\varepsilon' = \frac{1}{n^{1/10}\mathsf{poly}(p) \cdot C_\Sigma^2}$, we have

$$Q(\lambda_{\mathrm{CV}}, \mathbf{X}) \le Q(\lambda_*, \mathbf{X}) + \frac{1}{n^{1/20}}.$$

In other words, we have

$$|Q(\lambda_{\mathrm{CV}}, \mathbf{X}) - Q(\lambda_*, \mathbf{X})| < \frac{1}{n^{1/20}}$$

with probability $1 - O(p^2 K e^{-\Omega(\mathsf{poly}(n) \cdot \mathsf{poly}(1/p) C'_\Sigma)})$ for some constant C'_Σ depending only on Σ. Using the fact that Q is a continuous function, we can conclude our result.

C OMITTED PROOFS

C.1 Proof of Lemma 8

Lemma 8. *For any matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ and $\lambda > 0$, if $\theta \in \mathbb{R}^{p-1}$ satisfies*

$$(\theta - \theta^*)^\top \Gamma (\theta - \theta^*) < (\widehat{\theta}^{\lambda_*} - \theta^*)^\top \Gamma (\widehat{\theta}^{\lambda_*} - \theta^*)$$

then $\theta \neq \widehat{\theta}^{\lambda}$.

Proof. Recall that the definition of a, v and Γ from (1). Also, since Γ is a positive definite matrix, there exists a matrix H such that $\Gamma = HH^\top$. For any $\theta \in \mathbb{R}^{p-1}$, we first expand $\mathsf{E}_{Y \sim \mathcal{N}(0,\Sigma)}(Y_p - \sum_{j=1}^{p-1}\theta_j Y_j)^2$ as

$$\begin{aligned}
&\mathsf{E}_{Y \sim \mathcal{N}(0,\Sigma)}(Y_p - \sum_{j=1}^{p-1}\theta_j Y_j)^2 \\
&= \theta^\top \Gamma \theta - 2\theta^\top v + a \\
&= \theta^\top \Gamma \theta - 2\theta^\top H H^{-1} v + v^\top (H^{-1})^\top H^{-1} v - v^\top (H^{-1})^\top H^{-1} v + a \\
&= \|H^\top \theta - H^{-1} v\|_2^2 + a - v^\top \Gamma^{-1} v.
\end{aligned}$$

Recall that $\theta^* = \Gamma^{-1} v$. By plugging it into the above equation, we have

$$\begin{aligned}
\mathsf{E}_{Y \sim \mathcal{N}(0,\Sigma)}(Y_p - \sum_{j=1}^{p-1}\theta_j Y_j)^2 &= \|H^\top \theta - H^\top \theta^*\|_2^2 + a - v^\top \Gamma^{-1} v \\
&= (\theta - \theta^*)^\top \Gamma (\theta - \theta^*) + a - v^\top \Gamma^{-1} v
\end{aligned} \tag{27}$$

By the definition of λ_*, we have the following inequality

$$\mathsf{E}_{Y\sim\mathcal{N}(0,\Sigma)}(Y_p - \sum_{j=1}^{p-1}\widehat{\theta}_j^{\lambda_*}Y_j)^2 \le \mathsf{E}_{Y\sim\mathcal{N}(0,\Sigma)}(Y_p - \sum_{j=1}^{p-1}\widehat{\theta}_j^{\lambda}Y_j)^2 \qquad \text{for any } \lambda > 0.$$

By (27), it implies

$$(\widehat{\theta}^{\lambda_*} - \theta^*)^\top \Gamma(\widehat{\theta}^{\lambda_*} - \theta^*) \le (\widehat{\theta}^{\lambda} - \theta^*)^\top \Gamma(\widehat{\theta}^{\lambda} - \theta^*).$$

□

C.2 Proof of Lemma 9

Lemma 9. *Let $\mathbf{X} \in \mathbb{R}^{n\times p}$ be a sample matrix where each row is an i.i.d. sample drawn from $\mathcal{N}(0,\Sigma)$. Suppose the number of samples $n > |\widehat{\mathsf{ne}}^{\lambda_*}| + 2$. Then, we have $|\widehat{\mathsf{ne}}^{G}| \le |\widehat{\mathsf{ne}}^{\lambda_*}| + 1$ almost surely.*

Proof. Suppose $|\widehat{\mathsf{ne}}^{G}| \ge |\widehat{\mathsf{ne}}^{\lambda_*}| + 2$. Without loss of generality, we assume $\widehat{\mathsf{ne}}^{\lambda_*} = [s]$ for some integer s and $[s+2] \subseteq \widehat{\mathsf{ne}}^{G}$. Consider the following system of linear equations with $\theta \in \mathbb{R}^{p-1}$ and $\lambda \in \mathbb{R}$ as variables.

$$\begin{cases} \frac{1}{n}\langle \mathbf{X}_p - \sum_{j=1}^{p-1}\theta_j\mathbf{X}_j, \mathbf{X}_i\rangle = \mathsf{sign}(G_{\mathbf{X},i}(\widehat{\theta}^{\lambda_*}))\lambda & \text{for } i \in [s+2], \\ \theta_i = 0 & \text{for } i \notin [s]. \end{cases} \tag{28}$$

By the definition of $\widehat{\mathsf{ne}}^{\lambda_*}$, $(\widehat{\theta}^{\lambda_*}, \lambda_*)$ is a solution of this system. It means this system of linear equations (28) has a solution. Since (28) has a solution, the determinant of matrix $\widehat{\Gamma}_{[s+2],[s]\cup\{p,*\}}$ is 0 where $\widehat{\Gamma}_{[s+2],[s]\cup\{p,*\}}$ is the $(s+2)$-by-$(s+2)$ matrix that the rows are indexed by $[s+2]$ and the columns are indexed by $[s]\cup\{p,*\}$ and the (r,c)-entry is $\begin{cases} \frac{1}{n}\langle \mathbf{X}_r, \mathbf{X}_c\rangle & \text{for } r \in [s+2],\ c \in [s]\cup\{p\} \\ \mathsf{sign}(G_{\mathbf{X},r}(\widehat{\theta}^{\lambda_*})) & \text{for } r \in [s+2],\ c = * \end{cases}$, i.e.

$$\widehat{\Gamma}_{[s+2],[s]\cup\{p,*\}} = \begin{bmatrix} \frac{1}{n}\langle \mathbf{X}_1, \mathbf{X}_1\rangle & \cdots & \frac{1}{n}\langle \mathbf{X}_1, \mathbf{X}_s\rangle & \frac{1}{n}\langle \mathbf{X}_1, \mathbf{X}_p\rangle & \mathsf{sign}(G_{\mathbf{X},1}(\widehat{\theta}^{\lambda_*})) \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \frac{1}{n}\langle \mathbf{X}_{s+2}, \mathbf{X}_1\rangle & \cdots & \frac{1}{n}\langle \mathbf{X}_{s+2}, \mathbf{X}_s\rangle & \frac{1}{n}\langle \mathbf{X}_{s+2}, \mathbf{X}_p\rangle & \mathsf{sign}(G_{\mathbf{X},s+2}(\widehat{\theta}^{\lambda_*})) \end{bmatrix}.$$

Now, we project $\mathbf{X}_p$ onto the subspace spanned by $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{s+2}\}$ and write the projection as $\sum_{j=1}^{s+2}\alpha_j\mathbf{X}_j$ for some α_j. Plugging it into $\det(A) = 0$ and using the properties of determinant, we have

$$\alpha_{s+1}\det(\widehat{\Gamma}_{[s+2],[s]\cup\{s+1,*\}}) + \alpha_{s+2}\det(\widehat{\Gamma}_{[s+2],[s]\cup\{s+2,*\}}) = 0 \tag{29}$$

where

$$\widehat{\Gamma}_{[s+2],[s]\cup\{j,*\}} = \begin{bmatrix} \frac{1}{n}\langle \mathbf{X}_1, \mathbf{X}_1\rangle & \cdots & \frac{1}{n}\langle \mathbf{X}_1, \mathbf{X}_s\rangle & \frac{1}{n}\langle \mathbf{X}_1, \mathbf{X}_j\rangle & \mathsf{sign}(G_{\mathbf{X},1}(\widehat{\theta}^{\lambda_*})) \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \frac{1}{n}\langle \mathbf{X}_{s+2}, \mathbf{X}_1\rangle & \cdots & \frac{1}{n}\langle \mathbf{X}_{s+2}, \mathbf{X}_s\rangle & \frac{1}{n}\langle \mathbf{X}_{s+2}, \mathbf{X}_j\rangle & \mathsf{sign}(G_{\mathbf{X},s+2}(\widehat{\theta}^{\lambda_*})) \end{bmatrix} \quad \text{for } j = \{s+1, s+2\}.$$

Note that, conditioned on $\mathbf{X}_1, \ldots, \mathbf{X}_{s+2}$, $\alpha_{s+1}, \alpha_{s+2}$ are distributed as Gaussian if $n > s+2$. If one of $\det(\widehat{\Gamma}_{[s+2],[s]\cup\{s+1,*\}})$ and $\det(\widehat{\Gamma}_{[s+2],[s]\cup\{s+2,*\}})$ is not zero, the expression $\alpha_{s+1}\det(\widehat{\Gamma}_{[s+2],[s]\cup\{s+1,*\}}) + \alpha_{s+2}\det(\widehat{\Gamma}_{[s+2],[s]\cup\{s+2,*\}})$ is distributed as Gaussian conditioned on $\mathbf{X}_1, \ldots, \mathbf{X}_{s+2}$ and hence is non-zero almost surely which contradicts (29). Therefore, $|\widehat{\mathsf{ne}}^{G}| \le |\widehat{\mathsf{ne}}^{\lambda_*}| + 1$.

It remains to show that $\det(\widehat{\Gamma}_{[s+2],[s]\cup\{s+2,*\}})$ (or $\det(\widehat{\Gamma}_{[s+2],[s]\cup\{s+1,*\}})$) is non-zero almost surely. By Cramer's rule, $\frac{\det(\widehat{\Gamma}_{[s+2],[s]\cup\{s+2,*\}})}{\det(\widehat{\Gamma}_{[s+2]})}$ is the $(s+2)$-th entry of $\widehat{\Gamma}_{[s+2]}^{-1}q_{[s+2]}$ where $\widehat{\Gamma}_{[s+2]}$ is the $(s+2)$-by-$(s+2)$ matrix whose

(r,c)-entry is $\frac{1}{n}\langle \mathbf{X}_r, \mathbf{X}_c\rangle$ for $r,c \in [s+2]$ and $q_{[s+2]}$ is the $(s+2)$-dimensional vector whose i-th entry is $\mathsf{sign}(G_{\mathbf{X},i})$ for $i \in [s+2]$. Note that $\det(\widehat{\Gamma}_{[s+2]}) \neq 0$ almost surely if $n > s+2$. By the block matrix calculation, we have

$$(\widehat{\Gamma}^{-1}_{[s+2]} q_{[s+2]})_{s+2} = \frac{q_{s+2} - \widehat{u}^\top \widehat{\Gamma}^{-1}_{[s+1]} q_{[s+1]}}{\widehat{\Gamma}_{s+2,s+2} - \widehat{u}^\top \widehat{\Gamma}^{-1}_{[s+1]} \widehat{u}}$$

where $\widehat{u}$ is the $(s+1)$-dimensional vector whose i-th entry is $\frac{1}{n}\langle \mathbf{X}_i, \mathbf{X}_{s+2}\rangle$ for $i \in [s+1]$, $\widehat{\Gamma}_{[s+1]}$ is the $(s+1)$-by-$(s+1)$ submatrix of $\widehat{\Gamma}_{[s+2]}$ whose indices are in $[s+1]$ and $q_{[s+1]}$ is the $(s+1)$-dimensional subvector of $q_{[s+2]}$ whose indices are in $[s+1]$. Suppose $(\widehat{\Gamma}^{-1}_{[s+2]} q_{[s+2]})_{s+2} = 0$. We have

$$q_{s+2} - \widehat{u}^\top \widehat{\Gamma}^{-1}_{[s+1]} q_{[s+1]} = 0. \tag{30}$$

Conditioned on $\mathbf{X}_1, \ldots, \mathbf{X}_{s+1}$, the term $\widehat{u}^\top \widehat{\Gamma}^{-1}_{[s+1]} q_{[s+1]}$ depends linearly on $\mathbf{X}_{s+2}$ and hence it is distributed as Gaussian which is almost surely not 1 or -1. It contradicts (30). Therefore, $\det(\widehat{\Gamma}_{[s+2],[s]\cup\{s+2,*\}}) \neq 0$.

□

C.3 Proof of Lemma 10

Lemma 10. *For a sufficiently small $|\delta|$, suppose $\widetilde{\theta} = \widehat{\theta}^{\lambda_*} + \delta\theta' \in L$ where θ' is defined in* (19)*. Then, $\widetilde{\theta}$ is a Lasso solution for some penalty $\lambda > 0$.*

Proof. To use Lemma 7, we examine $G_{\mathbf{X},i}(\widehat{\theta}^{\lambda_*} + \delta\theta')$ for any i and we need to check the two cases of coordinates of $\widehat{\theta}^{\lambda_*} + \delta\theta'$ being non-zero or not in Lemma 7 which correspond to $i \in [s]$ and $i \notin [s]$.

We first examine it for $i \in [s]$. By direct calculation, we have

$$G_{\mathbf{X},i}(\widehat{\theta}^{\lambda_*} + \delta\theta') = \frac{1}{n}\langle \mathbf{X}_p - \sum_{j=1}^{p-1} (\widehat{\theta}^{\lambda_*}_j + \delta\theta'_j)\mathbf{X}_j, \mathbf{X}_i\rangle = q_i(\lambda_* + \delta) \qquad \text{for } i \in [s] \tag{31}$$

where q_i is the i-th entry of $q_{[s]}$.

We now examine it for $i \notin [s]$. By direct calculation, we have

$$\begin{aligned}
|G_{\mathbf{X},i}(\widehat{\theta}^{\lambda_*} + \delta\theta')| &= |\frac{1}{n}\langle \mathbf{X}_p - \sum_{j=1}^{p-1} (\widehat{\theta}^{\lambda_*}_j + \delta\theta'_j)\mathbf{X}_j, \mathbf{X}_i\rangle| \\
&\leq |\frac{1}{n}\langle \mathbf{X}_p - \sum_{j=1}^{p-1} \widehat{\theta}^{\lambda_*}_j \mathbf{X}_j, \mathbf{X}_i\rangle| + C \cdot |\delta| \\
&= |G_{\mathbf{X},i}(\widehat{\theta}^{\lambda_*})| + C \cdot |\delta| \qquad \text{for } i \notin [s]
\end{aligned}$$

where $C = |\sum_{j=1}^{p-1} \theta'_j \langle \mathbf{X}_j, \mathbf{X}_i\rangle|$. If we combine with the fact that $|G_{\mathbf{X},i}(\widehat{\theta}^{\lambda_*})| < \lambda_*$ for $i \notin [s]$, for a sufficiently small $|\delta|$, we have

$$|G_{\mathbf{X},i}(\widehat{\theta}^{\lambda_*} + \delta\theta')| < |G_{\mathbf{X},i}(\widehat{\theta}^{\lambda_*})| + C \cdot |\delta| < \lambda_* + \delta. \tag{32}$$

Using Lemma 7 with (31) and (32), we conclude that $\widehat{\theta}^{\lambda_*} + \delta\theta'$ is a Lasso solution for $\lambda = \lambda_* + \delta$ for a sufficiently small $|\delta|$.

□

C.4 Proof of Lemma 11

Lemma 11. *For a sufficiently small $|\delta|$, suppose $\widetilde{\theta} = \widehat{\theta}^{\lambda_*} + \delta \cdot \mathsf{sign}(Q)\theta' \in R_1$ for $\delta \geq 0$ and $\widetilde{\theta} = \widehat{\theta}^{\lambda_*} + \delta \cdot \mathsf{sign}(Q)\theta'' \in R_2$ for $\delta < 0$ where θ' and θ'' are defined in* (19) *and* (20) *respectively. Then, $\widetilde{\theta}$ is a Lasso solution for some penalty $\lambda > 0$.*

Proof. To use Lemma 7, we examine both $G_{\mathbf{X},i}(\widehat{\theta}^{\lambda_*} + \delta \cdot \mathsf{sign}(Q)\theta')$ when $\delta \geq 0$ and $G_{\mathbf{X},i}(\widehat{\theta}^{\lambda_*} + \delta \cdot \mathsf{sign}(Q)\theta'')$ when $\delta < 0$ for any i.

We first analyze $G_{\mathbf{X},i}(\widehat{\theta}^{\lambda_*} + \delta \cdot \mathsf{sign}(Q)\theta')$ when $\delta \geq 0$ and the analysis is similar to the analysis in Lemma 10. We need to check the two cases of coordinates of $\widehat{\theta}^{\lambda_*} + \delta \cdot \mathsf{sign}(Q)\theta'$ being non-zero or not which correspond to $i \in [s]$ and $i \notin [s]$. For $i \in [s]$, by the same calculation as in Lemma 10, we have

$$G_{\mathbf{X},i}(\widehat{\theta}^{\lambda_*} + \delta \cdot \mathsf{sign}(Q)\theta') = q_i(\lambda_* + \delta \cdot \mathsf{sign}(Q)) \qquad \text{for } i \in [s]. \tag{33}$$

For $i \notin [s]$, we further split into two cases: $i \notin [s+1]$ and $i = s+1$. For $i \notin [s+1]$, by the same calculation as in Lemma 10, we have

$$|G_{\mathbf{X},i}(\widehat{\theta}^{\lambda_*} + \delta \cdot \mathsf{sign}(Q)\theta')| \leq |G_{\mathbf{X},i}(\widehat{\theta}^{\lambda_*})| + C \cdot |\delta|. \qquad \text{for } i \notin [s+1]$$

where $C = |\sum_{j=1}^{p-1} \theta'_j \langle \mathbf{X}_j, \mathbf{X}_i \rangle|$ and hence

$$|G_{\mathbf{X},i}(\widehat{\theta}^{\lambda_*} + \delta\theta')| < \lambda_* + \delta \cdot \mathsf{sign}(Q) \tag{34}$$

for a sufficiently small $|\delta|$. For $i = s+1$, we have

$$\begin{aligned}
&G_{\mathbf{X},s+1}(\widehat{\theta}^{\lambda_*} + \delta \cdot \mathsf{sign}(Q)\theta') \\
&= q_{s+1}\lambda_* + \delta \cdot \mathsf{sign}(Q) \sum_{j=1}^{s} \theta'_j \cdot \frac{1}{n}\langle \mathbf{X}_j, \mathbf{X}_{s+1} \rangle \\
&= q_{s+1}\big(\lambda_* + \delta \cdot \mathsf{sign}(Q) - \delta \cdot \mathsf{sign}(Q)\big(1 - q_{s+1} \sum_{j=1}^{s} \theta'_j \cdot \frac{1}{n}\langle \mathbf{X}_j, \mathbf{X}_{s+1} \rangle\big)\big) \\
&= q_{s+1}(\lambda_* + \delta \cdot \mathsf{sign}(Q) - \mathsf{sign}(Q)Q \cdot \delta) \qquad \text{recall that } Q = 1 - q_{s+1} \sum_{j=1}^{s} \theta'_j \cdot \frac{1}{n}\langle \mathbf{X}_j, \mathbf{X}_{s+1} \rangle.
\end{aligned}$$

and hence, using the fact that $\mathsf{sign}(Q)Q = |Q| > 0$ and $\delta \geq 0$,

$$|G_{\mathbf{X},s+1}(\widehat{\theta}^{\lambda_*} + \delta\theta')| = \lambda_* + \delta \cdot \mathsf{sign}(Q) - |Q| \cdot \delta \leq \lambda_* + \delta \cdot \mathsf{sign}(Q) \tag{35}$$

for a sufficiently small $|\delta|$. Using Lemma 7 with (33), (34) and (35), we conclude that $\widehat{\theta} + \delta \cdot \mathsf{sign}(Q)\theta'$ is a Lasso solution for $\lambda = \lambda_* + \delta \cdot \mathsf{sign}(Q)$ for a sufficiently small $|\delta|$ and $\delta \geq 0$.

We now analyze $G_{\mathbf{X},i}(\widehat{\theta}^{\lambda_*} + \delta \cdot \mathsf{sign}(Q)\theta'')$ when $\delta < 0$. We need to check the two cases of coordinates of $\widehat{\theta}^{\lambda_*} + \delta \cdot \mathsf{sign}(Q)\theta''$ being non-zero or not which correspond to $i \in [s+1]$ and $i \notin [s+1]$. For $i \in [s+1]$, by a similar calculation as in Lemma 10, we have

$$G_{\mathbf{X},i}(\widehat{\theta}^{\lambda_*} + \delta \cdot \mathsf{sign}(Q)\theta'') = q_i(\lambda_* + \delta \cdot \mathsf{sign}(Q)) \qquad \text{for } i \in [s+1]. \tag{36}$$

We further split into two cases: $i \in [s]$ and $i = s+1$. For $i \in [s]$, we note that $q_i = \mathsf{sign}(\widehat{\theta}_i^{\lambda_*}) = \mathsf{sign}(\widehat{\theta}_i^{\lambda_*} + \delta \cdot \mathsf{sign}(Q)\theta''_i)$ for a sufficiently small $|\delta|$. For $i = s+1$, we check that $\widehat{\theta}_{s+1}^{\lambda_*} + \delta \cdot \mathsf{sign}(Q)\theta''_{s+1} = \delta \cdot \mathsf{sign}(Q)\theta''_{s+1}$ by the assumption on $\widehat{\theta}^{\lambda_*}$. From the definition of θ'' in (20) and direct block matrix calculation, we have

$$\theta''_{s+1} = \frac{-q_{s+1}(1 - q_{s+1}\widehat{u}^\top \widehat{\Gamma}_{[s]}^{-1} q_{[s]})}{\widehat{\Gamma}_{s+1,s+1} - \widehat{u}^\top \widehat{\Gamma}_{[s]}^{-1}\widehat{u}} = \frac{-q_{s+1} \cdot Q}{\widehat{\Gamma}_{s+1,s+1} - \widehat{u}^\top \widehat{\Gamma}_{[s]}^{-1}\widehat{u}}$$

where $\widehat{u}$ is the s-dimensional vector whose i-th entry is $\widehat{\Gamma}_{i,s+1} = \frac{1}{n}\langle \mathbf{X}_i, \mathbf{X}_{s+1} \rangle$ for $i \in [s]$. Since the denominator $\widehat{\Gamma}_{s+1,s+1} - \widehat{u}^\top \widehat{\Gamma}_{[s]}^{-1}\widehat{u}$ is positive, $\mathsf{sign}(Q)Q = |Q| > 0$ and $\delta < 0$, we have

$$\mathsf{sign}(\widehat{\theta}_{s+1}^{\lambda_*} + \delta \cdot \mathsf{sign}(Q)\theta''_{s+1}) = \mathsf{sign}(\delta \cdot \mathsf{sign}(Q)\theta''_{s+1}) = q_{s+1}. \tag{37}$$

For $i \notin [s+1]$, by a similar calculation as in Lemma 10, we have

$$|G_{\mathbf{X},i}(\widehat{\theta}^{\lambda_*} + \delta \cdot \mathsf{sign}(Q)\theta'')| \leq |G_{\mathbf{X},i}(\widehat{\theta}^{\lambda_*})| + C_1 \cdot |\delta|. \qquad \text{for } i \notin [s+1]$$

where $C_1 = |\sum_{j=1}^{p-1} \theta_j'' \langle \mathbf{X}_j, \mathbf{X}_i \rangle|$ and hence

$$|G_{\mathbf{X},i}(\widehat{\theta}^{\lambda_*} + \delta \cdot \mathsf{sign}(Q)\theta')| < \lambda_* + \delta \cdot \mathsf{sign}(Q) \tag{38}$$

for a sufficiently small $|\delta|$. Using Lemma 7 with (36), (37) and (38), we conclude that $\widehat{\theta}^{\lambda_*} + \delta \cdot \mathsf{sign}(Q)\theta''$ is a Lasso solution for $\lambda = \lambda_* + \delta \cdot \mathsf{sign}(Q)$ for a sufficiently small $|\delta|$ and $\delta < 0$.

□

C.5 Proof of Lemma 12

Lemma 12. *For any $\lambda > 0$ and any $\mathbf{Z} \in \mathbb{R}^{n\times p}, \mathbf{X} \in \mathbb{R}^{m\times p}$, we have*

$$|Q(\lambda, \mathbf{X}) - Q_{\mathbf{Z}}(\lambda, \mathbf{X})| = O(\varepsilon' \cdot (1 + \|\widehat{\theta}^{\lambda,\mathbf{X}}\|_1)^2).$$

Proof. By the definition of $Q(\lambda, \mathbf{X})$ and $Q_{\mathbf{Z}}(\lambda, \mathbf{X})$, we have

$$\begin{aligned} Q(\lambda, \mathbf{X}) - Q_{\mathbf{Z}}(\lambda, \mathbf{X}) &= \frac{1}{2}\mathsf{E}(Y_p - \sum_{j=1}^{p-1} \widehat{\theta}_j^{\lambda,\mathbf{X}} Y_j)^2 - \frac{1}{2n}\|\mathbf{Z}_p - \sum_{j=1}^{p-1} \widehat{\theta}_j^{\lambda,\mathbf{X}} \mathbf{Z}_j\|_2^2 \\ &= \frac{1}{2}\mathsf{E}(Y_p^2 - 2Y_p \sum_{j=1}^{p-1} \widehat{\theta}_j^{\lambda,\mathbf{X}} Y_j + \sum_{i=1}^{p-1}\sum_{j=1}^{p-1} \widehat{\theta}_i^{\lambda,\mathbf{X}} \widehat{\theta}_j^{\lambda,\mathbf{X}} Y_i Y_j) \\ &\quad - \frac{1}{2}\big(\frac{1}{n}\langle \mathbf{Z}_p, \mathbf{Z}_p \rangle - 2\sum_{j=1}^{p-1} \widehat{\theta}_j^{\lambda,\mathbf{X}} \frac{1}{n}\langle \mathbf{Z}_p, \mathbf{Z}_j \rangle + \sum_{i=1}^{p-1}\sum_{j=1}^{p-1} \widehat{\theta}_i^{\lambda,\mathbf{X}} \widehat{\theta}_j^{\lambda,\mathbf{X}} \frac{1}{n}\langle \mathbf{Z}_i, \mathbf{Z}_j \rangle\big) \end{aligned}$$

Recall that $|\mathsf{E}(Y_i Y_i) - \frac{1}{n}\langle \mathbf{Z}_i, \mathbf{Z}_j \rangle| < \varepsilon'$. Hence, we have

$$\begin{aligned} |Q(\lambda, \mathbf{X}) - Q_{\mathbf{Z}}(\lambda, \mathbf{X})| &< \frac{\varepsilon'}{2}\big(1 + 2\sum_{j=1}^{p-1} |\widehat{\theta}_j^{\lambda,\mathbf{X}}| + \sum_{i=1}^{p-1}\sum_{j=1}^{p-1} |\widehat{\theta}_i^{\lambda,\mathbf{X}}||\widehat{\theta}_j^{\lambda,\mathbf{X}}|\big) \\ &= \varepsilon' \cdot (1 + \|\widehat{\theta}^{\lambda,\mathbf{X}}\|_1)^2 \end{aligned}$$

□

C.6 Proof of Lemma 13

Lemma 13. *For any $\lambda > 0$ and any matrix $\mathbf{X} \sim \mathcal{N}(0, \Sigma)$, we have*

$$\|\widehat{\theta}^{\lambda,\mathbf{X}}\|_1 \le O(\sqrt{p}\kappa(\Sigma))$$

with probability $1 - O(e^{-\Omega(n)})$. Recall that $\kappa(\Sigma)$ is the condition number of Σ.

Proof. It is easy to see that

$$\|\widehat{\theta}^{\lambda,\mathbf{X}}\|_1 \le \|\widehat{\theta}^{0,\mathbf{X}}\|_1 = \|\widehat{\Gamma}^{-1}\widehat{v}\|_1.$$

Furthermore,

$$\|\widehat{\Gamma}^{-1}\widehat{v}\|_1 \le \sqrt{p}\|\widehat{\Gamma}^{-1}\widehat{v}\|_2 \le \sqrt{p}\|\widehat{\Gamma}^{-1}\|_2\|\widehat{v}\|_2$$

where $\widehat{v}$ is the $(p-1)$-dimensional vector whose i-th entry is $\frac{1}{n}\langle \mathbf{X}_i, \mathbf{X}_p \rangle$ for $i \in [p-1]$.

By matrix Chernoff bound, we have

$$\|\widehat{\Gamma} - \Gamma\|_2 < O(\|\Gamma\|_2) \qquad \text{and} \qquad \|\widehat{v} - v\|_2 < O(\|v\|_2)$$

with probability $1 - O(e^{-\Omega(n)})$. Hence, we have

$$\|\widehat{\theta}^{\lambda,\mathbf{X}}\|_1 \le O(\sqrt{p}\|\Gamma^{-1}\|_2\|v\|_2) \le O(\sqrt{p}\kappa(\Sigma)).$$

□

C.7 Proof of Lemma 14

Lemma 14. *For any $\lambda > 0$ and any matrices $\mathbf{X}, \mathbf{Z} \sim \mathcal{N}(0, \Sigma)$, we have*

$$|\|\widehat{\theta}^{\lambda,\mathbf{X}}\|_1 - \|\widehat{\theta}^{\lambda,\mathbf{Z}}\|_1| \le O(\sqrt{\varepsilon' \cdot \frac{p^2\kappa(\Sigma)^2}{\sigma_{\min}(\Sigma)}}) \qquad \text{and} \qquad \lambda \le O(\sqrt{p}\sigma_{\max}(\Sigma)\kappa(\Sigma))$$

as long as λ is not too large such that $\widehat{\theta}^{\lambda} \neq 0$.

Proof. For any $\lambda > 0$ and any matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$, define

$$F_{\lambda,\mathbf{X}}(\theta) := \frac{1}{2n}\|\mathbf{X}_p - \sum_{j=1}^{p-1}\theta_j\mathbf{X}_j\|^2 + \lambda\|\theta\|_1.$$

For any $\theta \in \mathbb{R}^{p-1}$, By Taylor expansion, we have

$$\begin{aligned} F_{\lambda,\mathbf{X}}(\theta) = F_{\lambda,\mathbf{X}}(\widehat{\theta}^{\lambda,\mathbf{X}}) &+ \nabla F_{\lambda,\mathbf{X}}(\widehat{\theta}^{\lambda,\mathbf{X}})^\top(\theta - \widehat{\theta}^{\lambda,\mathbf{X}}) \\ &+ \frac{1}{2}(\theta - \widehat{\theta}^{\lambda,\mathbf{X}})^\top\nabla^2 F_{\lambda,\mathbf{X}}(\widehat{\theta}^{\lambda,\mathbf{X}})(\theta - \widehat{\theta}^{\lambda,\mathbf{X}}) \end{aligned}$$

Note that $\nabla F_{\lambda,\mathbf{X}}(\widehat{\theta}^{\lambda,\mathbf{X}})^\top(\theta - \widehat{\theta}^{\lambda,\mathbf{X}}) \ge 0$; otherwise it contradicts the fact that $\widehat{\theta}^{\lambda,\mathbf{X}}$ is $\arg\min_{\theta\in\mathbb{R}^{p-1}} F_{\lambda,\mathbf{X}}(\theta)$. It means that

$$\begin{aligned} F_{\lambda,\mathbf{X}}(\theta) &\ge F_{\lambda,\mathbf{X}}(\widehat{\theta}^{\lambda,\mathbf{X}}) + \frac{1}{2}(\theta - \widehat{\theta}^{\lambda,\mathbf{X}})^\top\nabla^2 F_{\lambda,\mathbf{X}}(\widehat{\theta}^{\lambda,\mathbf{X}})(\theta - \widehat{\theta}^{\lambda,\mathbf{X}}) \\ &\ge F_{\lambda,\mathbf{X}}(\widehat{\theta}^{\lambda,\mathbf{X}}) + \frac{1}{2}\sigma_{\min}(\widehat{\Gamma})\|\theta - \widehat{\theta}^{\lambda,\mathbf{X}}\|_2^2 \end{aligned}$$

By matrix Chernoff bound, we have

$$\|\widehat{\Gamma} - \Gamma\|_2 < O(\|\Gamma\|_2)$$

with probability $1 - O(e^{-\Omega(n)})$. Namely, we have

$$\begin{aligned} F_{\lambda,\mathbf{X}}(\theta) &\ge F_{\lambda,\mathbf{X}}(\widehat{\theta}^{\lambda,\mathbf{X}}) + \Omega(\sigma_{\min}(\Gamma)\cdot\|\theta - \widehat{\theta}^{\lambda,\mathbf{X}}\|_2^2) \\ &\ge F_{\lambda,\mathbf{X}}(\widehat{\theta}^{\lambda,\mathbf{X}}) + \Omega(\frac{\sigma_{\min}(\Gamma)}{p}\cdot\|\theta - \widehat{\theta}^{\lambda,\mathbf{X}}\|_1^2). \end{aligned} \tag{39}$$

On the other hand,

$$\begin{aligned} F_{\lambda,\mathbf{X}}(\widehat{\theta}^{\lambda,\mathbf{Z}}) &= Q_{\mathbf{X}}(\lambda,\mathbf{Z}) + \lambda\|\widehat{\theta}^{\lambda,\mathbf{Z}}\|_1 && \text{recall the definition of } F_{\lambda,\mathbf{X}} \text{ and } Q_{\mathbf{X}} \\ &\le Q_{\mathbf{Z}}(\lambda,\mathbf{Z}) + O(\varepsilon' p\kappa(\Sigma)^2) + \lambda\|\widehat{\theta}^{\lambda,\mathbf{Z}}\|_1 && \text{by Lemma 12 and 13} \\ &\le Q_{\mathbf{Z}}(\lambda,\mathbf{X}) + O(\varepsilon' p\kappa(\Sigma)^2) + \lambda\|\widehat{\theta}^{\lambda,\mathbf{X}}\|_1 && \widehat{\theta}^{\lambda,\mathbf{Z}} \text{ is the minimum point} \\ &\le Q_{\mathbf{X}}(\lambda,\mathbf{X}) + O(\varepsilon' p\kappa(\Sigma)^2) + \lambda\|\widehat{\theta}^{\lambda,\mathbf{X}}\|_1 && \text{by Lemma 12 and 13} \\ &= F_{\lambda,\mathbf{X}}(\widehat{\theta}^{\lambda,\mathbf{X}}) + O(\varepsilon' p\kappa(\Sigma)^2) && \text{recall the definition of } F_{\lambda,\mathbf{X}} \text{ and } Q_{\mathbf{X}} \end{aligned} \tag{40}$$

Plugging (40) into (39) with $\theta = \widehat{\theta}^{\lambda,\mathbf{Z}}$, we have

$$\|\widehat{\theta}^{\lambda,\mathbf{Z}} - \widehat{\theta}^{\lambda,\mathbf{X}}\|_1^2 \le O(\varepsilon' \cdot \frac{p^2\kappa(\Sigma)^2}{\sigma_{\min}(\Sigma)})$$

or

$$|\|\widehat{\theta}^{\lambda,\mathbf{Z}}\|_1 - \|\widehat{\theta}^{\lambda,\mathbf{X}}\|_1| \le O(\sqrt{\varepsilon' \cdot \frac{p^2\kappa(\Sigma)^2}{\sigma_{\min}(\Sigma)}}).$$

By Lemma 7, we have

$$\frac{1}{n}\langle \mathbf{X}_p - \sum_{j=1}^{p-1} \widehat{\theta}_j^{\lambda,\mathbf{X}} \mathbf{X}_j, \mathbf{X}_i \rangle = \mathsf{sign}(\widehat{\theta}_i^{\lambda,\mathbf{X}})\lambda$$

for $i \in [p-1]$ that $\widehat{\theta}_i^{\lambda,\mathbf{X}} \neq 0$. Hence,

$$\lambda \leq M \cdot (1 + \|\widehat{\theta}^{\lambda,\mathbf{X}}\|_1)$$

where $M = \max\langle \mathbf{X}_i, \mathbf{X}_j \rangle$ as long as λ is not too large such as $\widehat{\theta}^{\lambda} \neq 0$. By Lemma 13, we have

$$\lambda \leq O(M \cdot \sqrt{p}\kappa(\Sigma)) \leq O(\sqrt{p}\sigma_{\max}(\Sigma)\kappa(\Sigma)).$$

□

C.8 Proof of Lemma 15

Lemma 15. *Recall that*

- $q_{[s]}$ *is the* s*-dimensional vector whose* i*-th entry is* $\mathsf{sign}(\widehat{\theta}_i^{\lambda_*})$ *for* $i \in [s]$
- $\widehat{\Gamma}_{[s]}$ *is the* s*-by-*s *matrix whose* (r,c)*-entry is* $\frac{1}{n}\langle \mathbf{X}_r, \mathbf{X}_c \rangle$ *for* $r, c \in [s]$
- $\Gamma_{[s]}$ *is the* s*-by-*s *submatrix of* Γ *in* (1) *whose indices are in* $[s]$
- $\theta^{\Delta} = \widehat{\theta}^{\lambda_*} - \theta^*$

Let $\mathcal{C}$ *be the event of* $|q_{[s]}\widehat{\Gamma}_{[s]}^{-1}\Gamma_{[s]}\theta_{[s]}^{\Delta}| \leq \frac{1}{100\sqrt{s\sigma_{\max}(\Gamma)}}\sqrt{\theta_{[s]}^{\Delta\top}\Gamma_{[s]}\theta_{[s]}^{\Delta}}$. *Then, we have*

$$\mathsf{Pr}_{\mathbf{X}\sim\mathcal{N}(0,\Sigma)^n}(\widehat{\mathsf{ne}}^{\lambda_*} = [s] \wedge \mathcal{C}) \leq O\left(2^s \cdot \left(p^{-\Omega(\frac{1}{\kappa(\Gamma)})} + spe^{-\Omega(\frac{n}{s^2\kappa(\Gamma)^3})}\right)\right).$$

Proof. From Lemma 7, we have

$$\begin{cases} \text{if } i \in [s], \text{ then } G_{\mathbf{X},i}(\widehat{\theta}^{\lambda_*}) = \frac{1}{n}\langle \mathbf{X}_p - \sum_{j=1}^{p-1} \widehat{\theta}_j^{\lambda_*}\mathbf{X}_j, \mathbf{X}_i \rangle = \mathsf{sign}(\widehat{\theta}_i^{\lambda_*})\lambda_* \\ \text{if } i \notin [s], \text{ then } G_{\mathbf{X},i}(\widehat{\theta}^{\lambda_*}) = \frac{1}{n}\langle \mathbf{X}_p - \sum_{j=1}^{p-1} \widehat{\theta}_j^{\lambda_*}\mathbf{X}_j, \mathbf{X}_i \rangle = \lambda_{i,*} \end{cases}$$

where $\lambda_{i,*}$ are some values whose absolute value is less than λ_*, i.e. $|\lambda_{i,*}| \leq \lambda_*$. Let $\widehat{w}$ be the $(p-1)$-dimensional vector whose i-th entry is $\frac{1}{n}\langle \mathbf{X}_p - \sum_{j=1}^{p-1} \theta_j^* \mathbf{X}_j, \mathbf{X}_i \rangle$ for $i \in [p-1]$ and z be the $(p-1)$-dimensional vector whose i-th entry is $\begin{cases} \mathsf{sign}(\widehat{\theta}_i^{\lambda_*})\lambda_* & \text{for } i \in [s] \\ \lambda_{i,*} & \text{for } i \notin [s] \end{cases}$. We can write it in the matrix form.

$$\widehat{w} - \widehat{\Gamma}\theta^{\Delta} = z \tag{41}$$

Recall that Γ is a positive definite matrix which means Γ can be decomposed as

$$\Gamma = HH^{\top} \qquad \text{for some matrix } H.$$

We now multiple the both sides of (41) by H^{-1} and we have

$$H^{-1}\widehat{w} - H^{-1}\widehat{\Gamma}\theta^{\Delta} = H^{-1}z. \tag{42}$$

for any H that satisfies $\Gamma = HH^{\top}$.

Note that there are infinitely many such decomposition by introducing an orthonormal matrix, i.e.

$$\Gamma = HU(HU)^\top \qquad \text{for some orthonormal matrix } U.$$

In other words, we can always introduce an orthonormal matrix to ensure H satisfies certain properties. We now multiple both sides by $(HU)^{-1}$ and we have

$$(HU)^{-1}\widehat{w} - (HU)^{-1}\widehat{\Gamma}\theta^{\Delta} = (HU)^{-1}z \tag{43}$$

There exists an orthonormal matrix U such that

$$\begin{cases} ((HU)^{-1})_{r,c} = 0 \text{ for } r \in [s] \text{ and } c \notin [s] \\ ((HU)^{-1}z)_i \text{ are positive and the same for } i \in [s]. \end{cases} \tag{44}$$

That is,

$$(HU)^{-1} = \begin{bmatrix} * & O_{s\times(p-1-s)} \\ * & * \end{bmatrix} \qquad \text{and} \qquad (HU)^{-1}z = \frac{1}{\sqrt{s}}\|((HU)^{-1}z)_{[s]}\|_2 \begin{bmatrix} 1 \\ \vdots \\ 1 \\ * \\ \vdots \\ * \end{bmatrix}$$

where $O_{s\times(p-1-s)}$ is the s-by-$(p-1-s)$ zero matrix and $((HU)^{-1}z)_{[s]}$ is the s-dimensional subvector of $(HU)^{-1}z$ whose indices are in $[s]$. Note that U depends on the samples by the second condition. Moreover, by the first condition, U only depends on $\mathsf{sign}(\widehat{\theta}_i^{\lambda_*})$ for $i \in [s]$. Hence, there are 2^s possibilities and we will take union bound over all of them.

Let H' be HU such that U satisfies (44). Observe that

$$\|(H'^{-1}z)_{-[s]}\|_2^2 = \sum_{j=s+1}^{p-1} (H'^{-1}z)_j^2$$

where $(H'^{-1}z)_{-[s]}$ is the $(p-1-s)$-dimensional subvector of $H'^{-1}z$ whose indices are not in $[s]$. Then, there are at least $\frac{p-1-s}{2}$ of $(H'^{-1}z)_j^2$ less than $\frac{2}{p-1-s}\|(H'^{-1}z)_{-[s]}\|_2^2$. Also, we can bound the term $\|(H'^{-1}z)_{-[s]}\|_2$ by

$$\|(H'^{-1}z)_{-[s]}\|_2 \le \|H'^{-1}\|_2 \cdot \|z\|_2 \le \sigma_{\max}(H'^{-1}) \cdot \sqrt{p-1}\lambda_* = \sqrt{\frac{p-1}{\sigma_{\min}(\Gamma)}}\lambda_*.$$

In other words, at least $\frac{p-1-s}{2}$ of $i \notin [s]$ such that

$$(H'^{-1}z)_i \le \sqrt{\frac{2(p-1)}{p-1-s}\frac{1}{\sigma_{\min}(\Gamma)}}\lambda_*$$

On the other hand, all of $(H'^{-1}z)_i$ are positive and the same for $i \in [s]$ and hence

$$\begin{aligned} (H'^{-1}z)_i &= \frac{1}{\sqrt{s}}\|(H'^{-1}z)_{[s]}\|_2 = \frac{1}{\sqrt{s}}\|(H'^{-1})_{[s]}z_{[s]}\|_2 \qquad \text{by (44)} \\ &\ge \frac{1}{\sqrt{s}}\sigma_{\min}(H'^{-1})\|z_{[s]}\|_2 \\ &= \sqrt{\frac{1}{\sigma_{\max}(\Gamma)}}\lambda_* \end{aligned} \tag{45}$$

where $(H'^{-1})_{[s]}$ is the s-by-s submatrix of H'^{-1} whose indices are in $[s]$ and $(H'^{-1}z)_{[s]}$ (resp. $z_{[s]}$) is the s-dimensional subvector of $H'^{-1}z$ (resp. z) whose indices are in $[s]$.

From (43), we have at least $\frac{p-1-s}{2}$ of $j \notin [s]$ such that, for all $i \in [s]$,

$$(H'^{-1}\widehat{w} - H'^{-1}\widehat{\Gamma}H'^{-\top}H'^{\top}\theta^{\Delta})_i \geq \eta \cdot (H'^{-1}\widehat{w} - H'^{-1}\widehat{\Gamma}H'^{-\top}H'^{\top}\theta^{\Delta})_j \tag{46}$$

where $\eta = \sqrt{\frac{p-1-s}{2(p-1)}\frac{1}{\kappa(\Gamma)}}$ and $\kappa(\Gamma)$ is the condition number of Γ, i.e. $\kappa(\Gamma) = \frac{\sigma_{\max}(\Gamma)}{\sigma_{\min}(\Gamma)}$.

By matrix Chernoff bound, we have

$$\|\widehat{\Gamma}_{[s]} - \Gamma_{[s]}\|_2 < t\|\Gamma_{[s]}\|_2 \tag{47}$$

with probability $1 - O(e^{-\Omega(t^2 n)})$ for any $t > 0$. Here, $\Gamma_{[s]}$ (resp. $\widehat{\Gamma}_{[s]}$) is the s-by-s submatrix of Γ (resp. $\widehat{\Gamma}$) whose indices are in $[s]$. By Weyl's inequality, we further have

$$\begin{aligned}
|\|\widehat{\Gamma}_{[s]}^{-1}\|_2 - \|\Gamma_{[s]}^{-1}\|_2| &= |\frac{1}{\|\widehat{\Gamma}_{[s]}^{-1}\|_2} - \frac{1}{\|\Gamma_{[s]}^{-1}\|_2}| \cdot \|\widehat{\Gamma}_{[s]}^{-1}\|_2\|\Gamma_{[s]}^{-1}\|_2 \\
&= |\sigma_{\min}(\widehat{\Gamma}_{[s]}) - \sigma_{\min}(\Gamma_{[s]})| \cdot \|\widehat{\Gamma}_{[s]}^{-1}\|_2\|\Gamma_{[s]}^{-1}\|_2 \\
&\leq \|\widehat{\Gamma}_{[s]} - \Gamma_{[s]}\|_2 \cdot \|\widehat{\Gamma}_{[s]}^{-1}\|_2\|\Gamma_{[s]}^{-1}\|_2 \leq t\kappa(\Gamma)\|\widehat{\Gamma}_{[s]}^{-1}\|_2.
\end{aligned} \tag{48}$$

which implies that

$$\|\Gamma_{[s]}\widehat{\Gamma}_{[s]}^{-1} - I\|_2 \leq \|\Gamma_{[s]} - \widehat{\Gamma}_{[s]}\|_2\|\widehat{\Gamma}_{[s]}^{-1}\|_2 \leq t\|\Gamma_{[s]}\|_2 \cdot \frac{1}{1 - t\kappa(\Gamma)}\|\Gamma_{[s]}^{-1}\|_2 = \frac{t\kappa(\Gamma)}{1 - t\kappa(\Gamma)}.$$

Let $H'_{[s]}$ be the s-by-s submatrix of H' whose indices are in $[s]$ and $z_{[s]}$ be the s-dimensional subvector of z whose indices are in $[s]$. Now, we have

$$\begin{aligned}
|({H'_{[s]}}^{-1}\Gamma_{[s]}\widehat{\Gamma}_{[s]}^{-1}z_{[s]})_i - ({H'_{[s]}}^{-1}z_{[s]})_i| &\leq \|{H'_{[s]}}^{-1}(\Gamma_{[s]}\widehat{\Gamma}_{[s]}^{-1} - I)z_{[s]}\|_2 \\
&\leq \|{H'_{[s]}}^{-1}\|_2\|(\Gamma_{[s]}\widehat{\Gamma}_{[s]}^{-1} - I)\|_2\|z_{[s]}\|_2 \\
&\leq \sqrt{\frac{1}{\sigma_{\min}(\Gamma)}} \cdot \frac{t\kappa(\Gamma)}{1 - t\kappa(\Gamma)} \cdot \sqrt{s}\lambda_*.
\end{aligned}$$

If we pick $t = \frac{1}{\kappa(\Gamma)(1+\sqrt{s\kappa(\Gamma)})} = \Theta(\frac{1}{\sqrt{s\kappa(\Gamma)^3}})$ then we have

$$\begin{aligned}
&|({H'_{[s]}}^{-1}\Gamma_{[s]}\widehat{\Gamma}_{[s]}^{-1}z_{[s]})_i - ({H'_{[s]}}^{-1}z_{[s]})_i| \\
&\quad\leq \frac{\lambda_*}{50\sqrt{\sigma_{\max}(\Gamma)}} \\
&\quad\leq \frac{1}{50}({H'_{[s]}}^{-1}z_{[s]})_i \qquad \text{recall that, from (45), } ({H'_{[s]}}^{-1}z_{[s]})_i = (H'^{-1}z)_i \geq \sqrt{\frac{1}{\sigma_{\max}(\Gamma)}}\lambda_* \text{ for } i \in [s]
\end{aligned} \tag{49}$$

with probability $1 - O(e^{-\Omega(\frac{n}{s\kappa(\Gamma)^3})})$.

Now, we can analyze the event $\mathcal{C}$ which is $|q_{[s]}\widehat{\Gamma}_{[s]}^{-1}\Gamma_{[s]}\theta^{\Delta}_{[s]}| \leq \frac{1}{100\sqrt{s\sigma_{\max}(\Gamma)}}\sqrt{\theta^{\Delta}_{[s]}{}^{\top}\Gamma_{[s]}\theta^{\Delta}_{[s]}}$. Note that

$$\begin{aligned}
\lambda_* \cdot q_{[s]}\widehat{\Gamma}_{[s]}^{-1}\Gamma_{[s]}\theta^{\Delta}_{[s]} &= \lambda_* \cdot q_{[s]}^{\top}\widehat{\Gamma}_{[s]}^{-1}\Gamma_{[s]}{H'_{[s]}}^{-\top}{H'_{[s]}}^{\top}\theta^{\Delta}_{[s]} \\
&= \lambda_* \cdot \sum_{j=1}^{s}({H'_{[s]}}^{-1}\Gamma_{[s]}\widehat{\Gamma}_{[s]}^{-1}q_{[s]})_j({H'_{[s]}}^{\top}\theta^{\Delta}_{[s]})_j \\
&\leq \sum_{j \in I_+}\frac{51}{50}({H'_{[s]}}^{-1}z_{[s]})_j({H'_{[s]}}^{\top}\theta^{\Delta}_{[s]})_j + \sum_{j \in I_-}\frac{49}{50}({H'_{[s]}}^{-1}z_{[s]})_j({H'_{[s]}}^{\top}\theta^{\Delta}_{[s]})_j
\end{aligned}$$

where I_+ (resp. I_-) is the subset of $[s]$ that $({H'_{[s]}}^\top \theta^\Delta_{[s]})_j$ is larger (resp. smaller) than 0 for $j \in I_+$ (resp. $j \in I_-$). Recall that $({H'_{[s]}}^{-1} z_{[s]})_j = \frac{1}{\sqrt{s}} \|(H'^{-1})_{[s]} z_{[s]}\|_2$ for all $j \in [s]$ and hence

$$\lambda_* \cdot q_{[s]} \widehat{\Gamma}^{-1}_{[s]} \Gamma_{[s]} \theta^\Delta_{[s]} \le \frac{1}{\sqrt{s}} \|(H'^{-1})_{[s]} z_{[s]}\|_2 \bigg(\sum_{j \in I_+} \frac{51}{50} ({H'_{[s]}}^\top \theta^\Delta_{[s]})_j + \sum_{j \in I_-} \frac{49}{50} ({H'_{[s]}}^\top \theta^\Delta_{[s]})_j \bigg) \tag{50}$$

On the other hand, note that $\frac{\lambda_*}{\sqrt{\sigma_{\max}(\Gamma)}} \le \frac{1}{\sqrt{s}} \|(H'^{-1})_{[s]} z_{[s]}\|_2$ and $\sqrt{{\theta^\Delta_{[s]}}^\top \Gamma_{[s]} \theta^\Delta_{[s]}} = \|{H'_{[s]}}^\top \theta^\Delta_{[s]}\|_2 \le \sqrt{s} \bigg(|({H'_{[s]}}^\top \theta^\Delta_{[s]})_{i_0}| + |({H'_{[s]}}^\top \theta^\Delta_{[s]})_{i_1}| \bigg)$ where i_0 is the index such that $i_0 = \arg\max_{i \in I_+} |({H'_{[s]}}^\top \theta^\Delta_{[s]})_i|$ (the largest positive value) and i_1 is the index such that $i_1 = \arg\max_{i \in I_-} |({H'_{[s]}}^\top \theta^\Delta_{[s]})_i|$ (the largest negative value). We have

$$\begin{aligned} \lambda_* \cdot q_{[s]} \widehat{\Gamma}^{-1}_{[s]} \Gamma_{[s]} \theta^\Delta_{[s]} &\ge -\frac{\lambda_*}{100\sqrt{s\sigma_{\max}(\Gamma)}} \sqrt{{\theta^\Delta_{[s]}}^\top \Gamma_{[s]} \theta^\Delta_{[s]}} \\ &\ge -\frac{1}{100} \cdot \frac{1}{\sqrt{s}} \|(H'^{-1})_{[s]} z_{[s]}\|_2 \cdot \bigg(|({H'_{[s]}}^\top \theta^\Delta_{[s]})_{i_0}| + |({H'_{[s]}}^\top \theta^\Delta_{[s]})_{i_1}| \bigg). \end{aligned} \tag{51}$$

By comparing (50) and (51), we have

$$-\frac{1}{100} \bigg(|({H'_{[s]}}^\top \theta^\Delta_{[s]})_{i_0}| + |({H'_{[s]}}^\top \theta^\Delta_{[s]})_{i_1}| \bigg) \le \sum_{j \in I_+} \frac{51}{50} ({H'_{[s]}}^\top \theta^\Delta_{[s]})_j + \sum_{j \in I_-} \frac{49}{50} ({H'_{[s]}}^\top \theta^\Delta_{[s]})_j$$

and it implies

$$\sum_{j \in I_-} \left| \frac{({H'_{[s]}}^\top \theta^\Delta_{[s]})_j}{({H'_{[s]}}^\top \theta^\Delta_{[s]})_{i_0}} \right| < \frac{103}{97} s \tag{52}$$

where i_0 is the index such that $i_0 = \arg\max_i ({H'_{[s]}}^\top \theta^\Delta_{[s]})_i$.

Recall that $\theta^\Delta_i = 0$ for $i \notin [s]$ and $(H'^\top)_{r,c} = 0$ for $r \notin [s]$ and $s \in [s]$. Then, we have

$$H'^{-1} \widehat{\Gamma} H'^{-\top} H'^\top \theta^\Delta = (H'^{-1} \widehat{\Gamma} H'^{-\top})_{[p-1],[s]} H'^\top_{[s]} \theta^\Delta_{[s]}$$

where $(H'^{-1} \widehat{\Gamma} H'^{-\top})_{[p-1],[s]}$ is the $(p-1)$-by-s submatrix of $H'^{-1} \widehat{\Gamma} H'^{-\top}$ whose row indices are in $[p-1]$ and column indices are in $[s]$.

Moreover, recall that

$$\widehat{w}_i = \frac{1}{n} \langle \mathbf{X}_p - \sum_{j=1}^{p-1} \theta^*_j \mathbf{X}_j, \mathbf{X}_i \rangle = \frac{1}{n} \mathbf{X}^\top b$$

and we can rewrite it as

$$H^{-1} \widehat{w} = \frac{1}{n} \sum_{j=1}^{n} b_j H^{-1} {\mathbf{X}^{(j)}}^\top$$

where b is the n-dimensional vector $\mathbf{X}_p - \sum_{j=1}^{p-1} \theta^*_j \mathbf{X}_j$ and $\mathbf{X}^{(i)}$ be the i-th row of $\mathbf{X}$. Note that $H^{-1} {\mathbf{X}^{(i)}}^\top$ are distributed as $\mathcal{N}(0, I)$. Indeed, it is easy to check that

$$\mathsf{E}_{\mathbf{X}^{(i)} \sim \mathcal{N}(0,\Sigma)} \left(H^{-1} {\mathbf{X}^{(i)}}^\top \mathbf{X}^{(i)} H^{-\top} \right) = I \qquad \text{for any } i \in [n]. \tag{53}$$

Recall that $\mathbf{X}^{(i)}$ is the i-th row of $\mathbf{X}$. By (53), $H'^{-1} {\mathbf{X}^{(i)}}^\top$ are distributed as $\mathcal{N}(0, I)$. Consider the entries of $(H'^{-1} \widehat{\Gamma} H'^{-\top})_{r,c}$ for $r \in [p-1]$ and $c \in [s]$. By Chernoff bound and union bound, for all $r \in [p-1]$ and $c \in [s]$, we have

$$|(H'^{-1} \widehat{\Gamma} H'^{-\top})_{r,c} - 1_{r=c}| < t \tag{54}$$

with probability $1 - O(spe^{-\Omega(t^2 n)})$ for any $t > 0$. Here, $1_{r=c} = \begin{cases} 1 & \text{if } r = c \\ 0 & \text{if } r \neq c. \end{cases}$

Combining (54) and (52), there exists an index $i_0 \in [s]$ such that, for all $i \notin [s]$, we have

$$\begin{aligned}
&(H'^{-1}\widehat{\Gamma}H'^{-\top}H'^{\top}\theta^{\Delta})_{i_0} - \eta \cdot (H'^{-1}\widehat{\Gamma}H'^{-\top}H'^{\top}\theta^{\Delta})_i \\
&\quad = ((H'^{-1}\widehat{\Gamma}H'^{-\top})_{[p-1],[s]}H'^{\top}_{[s]}\theta^{\Delta}_{[s]})_{i_0} - \eta \cdot ((H'^{-1}\widehat{\Gamma}H'^{-\top})_{[p-1],[s]}H'^{\top}_{[s]}\theta^{\Delta}_{[s]})_i \\
&\quad = \sum_{j=1}^{s}(H'^{-1}\widehat{\Gamma}H'^{-\top})_{i_0,j}(H'^{\top}_{[s]}\theta^{\Delta}_{[s]})_j - \eta \cdot \sum_{j=1}^{s}(H'^{-1}\widehat{\Gamma}H'^{-\top})_{i,j}(H'^{\top}_{[s]}\theta^{\Delta}_{[s]})_j \\
&\quad = (H'^{\top}_{[s]}\theta^{\Delta}_{[s]})_{i_0}\left(\sum_{j=1}^{s}(H'^{-1}\widehat{\Gamma}H'^{-\top})_{i_0,j}\frac{(H'^{\top}_{[s]}\theta^{\Delta}_{[s]})_j}{(H'^{\top}_{[s]}\theta^{\Delta}_{[s]})_{i_0}} - \eta \cdot \sum_{j=1}^{s}(H'^{-1}\widehat{\Gamma}H'^{-\top})_{i,j}\frac{(H'^{\top}_{[s]}\theta^{\Delta}_{[s]})_j}{(H'^{\top}_{[s]}\theta^{\Delta}_{[s]})_{i_0}}\right) \\
&\quad \geq (H'^{\top}_{[s]}\theta^{\Delta}_{[s]})_{i_0}\left((1-t) - (t(s-1) + t\frac{103}{97}s) - \eta \cdot (ts + t\frac{103}{97}s)\right) \\
&\quad \geq 0
\end{aligned} \tag{55}$$

with probability $1 - O(spe^{-\Omega(\frac{n}{s^2})})$ if we pick $t = \frac{1}{1000s}$.

If we plug (55) into (46), there exists an index $i_0 \in [s]$ such that, for at least $\frac{p-1-s}{2}$ of $i \notin [s]$, we have

$$(H'^{-1}\widehat{w})_{i_0} \geq \eta \cdot (H'^{-1}\widehat{w})_i. \tag{56}$$

Recall that

$$\widehat{w}_i = \frac{1}{n}\langle \mathbf{X}_p - \sum_{j=1}^{p-1}\theta^*_j \mathbf{X}_j, \mathbf{X}_i\rangle \qquad \text{and} \qquad H'^{-1}\widehat{w} = \sum_{j=1}^{n} b_j H'^{-1}\mathbf{X}^{(j)\top}$$

where b is the n-dimensional vector $\mathbf{X}_p - \sum_{j=1}^{p-1}\theta^*_j \mathbf{X}_j$. The entries of b are distributed as $\mathcal{N}(0, a - v^\top \Gamma v)$ independently and independent to the entries of $H'^{-1}\mathbf{X}^{(j)\top}$ for $j \in [n]$. If we further rewrite (56) as

$$\sum_{j=1}^{n}\frac{b_j}{\|b\|_2}\frac{(H'^{-1}\mathbf{X}^{(j)\top})_{i_0}}{\sqrt{n}} \geq \eta \cdot \left(\sum_{j=1}^{n}\frac{b_j}{\|b\|_2}\frac{(H'^{-1}\mathbf{X}^{(j)\top})_i}{\sqrt{n}}\right).$$

Hence, we can view the term $\frac{b}{\|b\|_2}$ as a random projection and both side are just a Gaussian variable from $\mathcal{N}(0,1)$. The probability of this event is

$$\int_{-\infty}^{\infty}\left(\mathsf{erf}(\frac{x}{\eta})\right)^{\frac{p-1-s}{2}}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\mathsf{d}x$$

where $\mathsf{erf}(*) = \int_{-\infty}^{*}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\mathsf{d}x$. To bound this probability, we first see that

$$\mathsf{erf}(\frac{x}{\eta}) = \int_{-\infty}^{\frac{x}{\eta}}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}\mathsf{d}y = 1 - \int_{\frac{x}{\eta}}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}\mathsf{d}y < 1 - \Omega(e^{-O((\frac{x}{\eta})^2)})$$

Let ξ be the value such that $\mathsf{erf}(\frac{\xi}{\eta}) < 1 - \frac{1}{\sqrt{\frac{p-1-s}{2}}}$ which means $\xi = \Theta(\eta\sqrt{\log p})$. Then, the probability can be further expressed as

$$\begin{aligned}
&\int_{-\infty}^{\infty}\left(\mathsf{erf}(\frac{x}{\eta})\right)^{\frac{p-1-s}{2}}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\mathsf{d}x \\
&\quad = \int_{-\infty}^{\xi}\left(\mathsf{erf}(\frac{x}{\eta})\right)^{\frac{p-1-s}{2}}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\mathsf{d}x + \int_{\xi}^{\infty}\left(\mathsf{erf}(\frac{x}{\eta})\right)^{\frac{p-1-s}{2}}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\mathsf{d}x.
\end{aligned}$$

For the first term,

$$\int_{-\infty}^{\xi}\left(\mathsf{erf}(\frac{x}{\eta})\right)^{\frac{p-1-s}{2}}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\mathsf{d}x<\int_{-\infty}^{\xi}\left(1-\frac{1}{\sqrt{\frac{p-1-s}{2}}}\right)^{\frac{p-1-s}{2}}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\mathsf{d}x$$
$$<e^{-\frac{1}{\sqrt{\frac{p-1-s}{2}}}}.$$

For the second term,

$$\int_{\xi}^{\infty}\left(\mathsf{erf}(\frac{x}{\eta})\right)^{\frac{p-1-s}{2}}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\mathsf{d}x<O(e^{-\Omega(\xi^2)})=O(p^{-\Omega(\eta^2)}).$$

By combining these two terms and recalling that $\eta=\sqrt{\frac{p-1-s}{2(p-1)}\frac{1}{\kappa(\Gamma)}}$, we have

$$\int_{-\infty}^{\infty}\left(\mathsf{erf}(\frac{x}{\eta})\right)^{\frac{p-1-s}{2}}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\mathsf{d}x\leq e^{-\frac{1}{\sqrt{\frac{p-1-s}{2}}}}+O(p^{-\Omega(\eta^2)})\leq O(p^{\Omega(-\frac{1}{\kappa(\Gamma)})}).\tag{57}$$

Finally, combining the failure probabilities in (49), (55) and (57) and taking union bound over all 2^s choices of H' for satisfying (44), we have

$$\begin{aligned}\mathsf{Pr}_{\mathbf{X}\sim\mathcal{N}(0,\Sigma)^n}(\widehat{\mathsf{ne}}^{\lambda_*}=[s]\wedge\mathcal{C})&\leq O\bigg(2^s\cdot\Big(p^{-\Omega(\frac{1}{\kappa(\Gamma)})}+spe^{-\Omega(\frac{n}{s^2})}+e^{-\Omega(\frac{n}{s\kappa(\Gamma)^3})}\Big)\bigg)\\&\leq O\bigg(2^s\cdot\Big(p^{-\Omega(\frac{1}{\kappa(\Gamma)})}+spe^{-\Omega(\frac{n}{s^2\kappa(\Gamma)^3})}\Big)\bigg).\end{aligned}$$

□

C.9 Proof of Lemma 16

Lemma 16. *Recall that*

- *the event* $\mathcal{A}$ *is* $\mathsf{sign}(Q)\theta'^{\top}\Gamma(\widehat{\theta}^{\lambda_*}-\theta^*)\geq 0$
- *the event* $\mathcal{B}$ *is* $-\mathsf{sign}(Q)\theta''^{\top}\Gamma(\widehat{\theta}^{\lambda_*}-\theta^*)\geq 0$

where Q, θ' *and* θ'' *are defined in* (21), (19) *and* (20) *respectively. Then, we have*

$$\mathsf{Pr}_{\mathbf{X}\sim\mathcal{N}(0,\Sigma)^n}(\widehat{\mathsf{ne}}^{\lambda_*}=[s]\wedge\mathcal{A}\wedge\mathcal{B})<O\bigg(2^s\cdot\Big(p^{-\Omega(\frac{1}{\kappa(\Gamma)})}+spe^{-\Omega(\frac{n}{s^2\kappa(\Gamma)^6})}\Big)\bigg).$$

Proof. Recall that $\theta^{\Delta}=\widehat{\theta}^{\lambda_*}-\theta^*$. We first expand $\theta''^{\top}\Gamma\theta^{\Delta}$. By the fact $\theta_i^{\Delta}=0$ for $i\notin[s]$ from $\mathsf{ne}^*=\widehat{\mathsf{ne}}^{\lambda_*}=[s]$ and the definition of θ'' in (20), we have

$$\theta''^{\top}\Gamma\theta^{\Delta}=-q_{[s+1]}^{\top}\widehat{\Gamma}_{[s+1]}^{-1}\Gamma_{[s+1],[s]}\theta_{[s]}^{\Delta}$$

where $\Gamma_{[s+1],[s]}$ is the $(s+1)$-by-s sub-matrix of Γ whose row indices are in $[s+1]$ and column indices are in $[s]$. By direct block matrix calculation, we have

$$\begin{aligned}&q_{[s+1]}^{\top}\widehat{\Gamma}_{[s+1]}^{-1}\Gamma_{[s+1],[s]}\\&=q_{[s]}^{\top}\widehat{\Gamma}_{[s]}^{-1}\Gamma_{[s]}+\frac{q_{[s]}^{\top}\widehat{\Gamma}_{[s]}^{-1}\widehat{u}}{\widehat{\Gamma}_{s+1}-\widehat{u}^{\top}\widehat{\Gamma}_{[s]}^{-1}\widehat{u}}\widehat{u}^{\top}\widehat{\Gamma}_{[s]}^{-1}\Gamma_{[s]}-\frac{q_{s+1}}{\widehat{\Gamma}_{s+1}-\widehat{u}^{\top}\widehat{\Gamma}_{[s]}^{-1}\widehat{u}}\widehat{u}^{\top}\widehat{\Gamma}_{[s]}^{-1}\Gamma_{[s]}\\&\quad-\frac{q_{[s]}^{\top}\widehat{\Gamma}_{[s]}^{-1}\widehat{u}}{\widehat{\Gamma}_{s+1}-\widehat{u}^{\top}\widehat{\Gamma}_{[s]}^{-1}\widehat{u}}u^{\top}+\frac{q_{s+1}}{\widehat{\Gamma}_{s+1}-\widehat{u}^{\top}\widehat{\Gamma}_{[s]}^{-1}\widehat{u}}u^{\top}\\&=q_{[s]}^{\top}\widehat{\Gamma}_{[s]}^{-1}\Gamma_{[s]}-\frac{q_{[s]}^{\top}\widehat{\Gamma}_{[s]}^{-1}\widehat{u}-q_{s+1}}{\widehat{\Gamma}_{s+1}-\widehat{u}^{\top}\widehat{\Gamma}_{[s]}^{-1}\widehat{u}}(u^{\top}-\widehat{u}^{\top}\widehat{\Gamma}_{[s]}^{-1}\Gamma_{[s]})\end{aligned}$$

where $\widehat{u}$ (resp. u) is the s-dimensional vector whose i-th entry is $\widehat{\Gamma}_{i,s+1}$ (resp. $\Gamma_{i,s+1}$) for $i \in [s]$. In other words, we have

$$-\mathsf{sign}(Q)\theta''^{\top}\Gamma\theta^{\Delta} = \mathsf{sign}(Q)q_{[s]}^{\top}\widehat{\Gamma}_{[s]}^{-1}\Gamma_{[s]}\theta_{[s]}^{\Delta} - \mathsf{sign}(Q)\frac{q_{[s]}^{\top}\widehat{\Gamma}_{[s]}^{-1}\widehat{u} - q_{s+1}}{\widehat{\Gamma}_{s+1} - \widehat{u}^{\top}\widehat{\Gamma}_{[s]}^{-1}\widehat{u}}(u^{\top} - \widehat{u}^{\top}\widehat{\Gamma}_{[s]}^{-1}\Gamma_{[s]})\theta_{[s]}^{\Delta}. \tag{58}$$

By the fact $\theta_i^{\Delta} = 0$ for $i \notin [s]$ from $\widehat{\mathsf{ne}}^{*} = \widehat{\mathsf{ne}}^{\lambda_*} = [s]$ and the definition of θ' in (19), we have

$$\theta'^{\top}\Gamma\theta^{\Delta} = -q_{[s]}^{\top}\widehat{\Gamma}_{[s]}^{-1}\Gamma_{[s]}\theta_{[s]}^{\Delta}.$$

We plug it into (58) and use the assumption that $-\mathsf{sign}(Q)\theta''^{\top}\Gamma\theta^{\Delta} \geq 0$. Then, we have

$$\mathsf{sign}(Q)\theta'^{\top}\Gamma\theta^{\Delta} \leq -\mathsf{sign}(Q)\frac{q_{[s]}^{\top}\widehat{\Gamma}_{[s]}^{-1}\widehat{u} - q_{s+1}}{\widehat{\Gamma}_{s+1} - \widehat{u}^{\top}\widehat{\Gamma}_{[s]}^{-1}\widehat{u}}(u^{\top} - \widehat{u}^{\top}\widehat{\Gamma}_{[s]}^{-1}\Gamma_{[s]})\theta_{[s]}^{\Delta}. \tag{59}$$

We now further bound the terms in the RHS.

By matrix Chernoff bound, we first have

$$\|\widehat{\Gamma}_{[s+1]} - \Gamma_{[s+1]}\|_2 < t\|\Gamma_{[s+1]}\|_2$$

with probability $1 - O(e^{-\Omega(t^2 n)})$ for any $t > 0$. Here, $\Gamma_{[s+1]}$ (resp. $\widehat{\Gamma}_{[s+1]}$) is the $(s+1)$-by-$(s+1)$ submatrix of Γ (resp. $\widehat{\Gamma}$) whose indices are in $[s+1]$. By the similar argument as in (48), it implies

$$\begin{aligned}
|\|\widehat{\Gamma}_{[s]}^{-1}\|_2 - \|\Gamma_{[s]}^{-1}\|_2| &\leq t\kappa(\Gamma)\|\widehat{\Gamma}_{[s]}^{-1}\|_2, \\
|\|\widehat{\Gamma}_{[s+1]}^{-1}\|_2 - \|\Gamma_{[s+1]}^{-1}\|_2|, &\leq t\kappa(\Gamma)\|\widehat{\Gamma}_{[s+1]}^{-1}\|_2 \\
\|\widehat{\Gamma}_{[s]}^{-1} - \Gamma_{[s]}^{-1}\|_2 &\leq t\kappa(\Gamma)\|\widehat{\Gamma}_{[s]}^{-1}\|_2.
\end{aligned}$$

For the term $q_{[s]}^{\top}\widehat{\Gamma}_{[s]}^{-1}\widehat{u} - q_{s+1}$, we have

$$|q_{[s]}^{\top}\widehat{\Gamma}_{[s]}^{-1}\widehat{u} - q_{s+1}| \leq \|q_{[s]}\|_2\|\widehat{\Gamma}_{[s]}^{-1}\|_2\|\widehat{u}\|_2 + 1$$

Recall that the entries of $q_{[s]}$ has absolute values 1 and hence $\|q_{[s]}\|_2 = \sqrt{s}$. The term $\|\widehat{\Gamma}_{[s]}^{-1}\|_2$ is bounded by $\frac{1}{(1-t\kappa(\Gamma))\sigma_{\min}(\Gamma)}$ and the term $\|\widehat{u}\|_s$ is bounded by $\|\widehat{\Gamma}_{[s+1]}\|_2 \leq t\sigma_{\max}(\Gamma)$. We have

$$|q_{[s]}^{\top}\widehat{\Gamma}_{[s]}^{-1}\widehat{u} - q_{s+1}| \leq \sqrt{s} \cdot \frac{1}{(1-t\kappa(\Gamma))\sigma_{\min}(\Gamma)} \cdot t\sigma_{\max}(\Gamma) + 1 = \frac{\sqrt{s}t\kappa(\Gamma)}{1-t\kappa(\Gamma)} + 1 \tag{60}$$

For the term $\widehat{\Gamma}_{s+1} - \widehat{u}^{\top}\widehat{\Gamma}_{[s]}\widehat{u}$, we have

$$|\widehat{\Gamma}_{s+1} - \widehat{u}^{\top}\widehat{\Gamma}_{[s]}^{-1}\widehat{u}| \geq \sigma_{\min}(\widehat{\Gamma}_{[s+1]}) \geq (1-t\kappa(\Gamma))\sigma_{\min}(\Gamma). \tag{61}$$

For the term $(u^{\top} - \widehat{u}^{\top}\widehat{\Gamma}_{[s]}^{-1}\Gamma_{[s]})\theta_{[s]}^{\Delta}$, we further expand it as

$$(u^{\top} - \widehat{u}^{\top}\widehat{\Gamma}_{[s]}^{-1}\Gamma_{[s]})\theta_{[s]}^{\Delta} = u^{\top}(\Gamma_{[s]}^{-\top} - \widehat{\Gamma}_{[s]}^{-\top})H'_{[s]}{H'_{[s]}}^{\top}\theta_{[s]}^{\Delta} + (u^{\top} - \widehat{u}^{\top})\widehat{\Gamma}_{[s]}^{-1}H'_{[s]}{H'_{[s]}}^{\top}\theta_{[s]}^{\Delta}$$

where H' is the matrix satisfying $\Gamma = H'H'^{\top}$ and (44) and $H'_{[s]}$ is its s-by-s submatrix whose indices are in $[s]$. For the first term, we have

$$\begin{aligned}
|u^{\top}(\Gamma_{[s]}^{-\top} - \widehat{\Gamma}_{[s]}^{-\top})H'_{[s]}{H'_{[s]}}^{\top}\theta_{[s]}^{\Delta}| &\leq \|u\|_2 \cdot \|\Gamma_{[s]}^{-1} - \widehat{\Gamma}_{[s]}^{-1}\|_2 \cdot \|H'_{[s]}\|_2 \cdot \|{H'_{[s]}}^{\top}\theta_{[s]}^{\Delta}\|_2 \\
&\leq \sigma_{\max}(\Gamma) \cdot \frac{t\kappa(\Gamma)}{\sigma_{\min}(\Gamma)} \cdot \sqrt{\sigma_{\max}(\Gamma)} \cdot \|{H'_{[s]}}^{\top}\theta_{[s]}^{\Delta}\|_2 \\
&= t\kappa(\Gamma)^2\sqrt{\sigma_{\max}(\Gamma)} \cdot \|{H'_{[s]}}^{\top}\theta_{[s]}^{\Delta}\|_2.
\end{aligned}$$

For the second term, we have

$$\begin{aligned}|(u^\top - \widehat{u}^\top)\widehat{\Gamma}_{[s]}^{-1} {H'_{[s]}}^\top \theta^\Delta_{[s]}| &\le \|u - \widehat{u}\|_2 \cdot \|\widehat{\Gamma}_{[s]}^{-1}\|_2 \cdot \|{H'_{[s]}}^\top \theta^\Delta_{[s]}\|_2 \\ &\le t\sigma_{\max}(\Gamma) \cdot \frac{1}{(1-t\kappa(\Gamma))\sigma_{\min}(\Gamma)} \cdot \sqrt{\sigma_{\max}(\Gamma)} \cdot \|{H'_{[s]}}^\top \theta^\Delta_{[s]}\|_2 \\ &= \frac{t\kappa(\Gamma)\sqrt{\sigma_{\max}(\Gamma)}}{1-t\kappa(\Gamma)} \cdot \|{H'_{[s]}}^\top \theta^\Delta_{[s]}\|_2.\end{aligned}$$

It means that

$$|(u^\top - \widehat{u}^\top \widehat{\Gamma}_{[s]}^{-1}\Gamma_{[s]})\theta^\Delta_{[s]}| \le \left(\kappa(\Gamma) + \frac{1}{1-t\kappa(\Gamma)}\right) t\kappa(\Gamma)\sqrt{\sigma_{\max}(\Gamma)} \|{H'_{[s]}}^\top \theta^\Delta_{[s]}\|_2 \tag{62}$$

Plugging (60), (61) and (62) into the RHS of (59), we have

$$\begin{aligned}\mathsf{sign}(Q){\theta'}^\top \Gamma \theta^\Delta &\le |\frac{q_{[s]}^\top \widehat{\Gamma}_{[s]}^{-1}\widehat{u} - q_{s+1}}{\widehat{\Gamma}_{s+1} - \widehat{u}^\top \widehat{\Gamma}_{[s]}^{-1}\widehat{u}}(u^\top - \widehat{u}^\top \widehat{\Gamma}_{[s]}^{-1}\Gamma_{[s]})\theta^\Delta_{[s]}| \\ &\le \frac{\left(\frac{\sqrt{s}t\kappa(\Gamma)}{1-t\kappa(\Gamma)} + 1\right)\left(\kappa(\Gamma) + \frac{1}{1-t\kappa(\Gamma)}\right) t\kappa(\Gamma)\sqrt{\sigma_{\max}(\Gamma)}}{(1-t\kappa(\Gamma))\sigma_{\min}(\Gamma)} \|{H'_{[s]}}^\top \theta^\Delta_{[s]}\|_2 \\ &= \frac{\left(\frac{\sqrt{s}t\kappa(\Gamma)}{1-t\kappa(\Gamma)} + 1\right)\left(\kappa(\Gamma) + \frac{1}{1-t\kappa(\Gamma)}\right) t\kappa(\Gamma)^2}{(1-t\kappa(\Gamma))\sqrt{\sigma_{\max}(\Gamma)}} \|{H'_{[s]}}^\top \theta^\Delta_{[s]}\|_2\end{aligned}$$

If we pick $t = \frac{1}{1000\sqrt{s}\kappa(\Gamma)^3} = \Theta(\frac{1}{\sqrt{s}\kappa(\Gamma)^3})$, then we have $1 - t\kappa(\Gamma) = 1 - \frac{1}{1000\sqrt{s}\kappa(\Gamma)^2} \ge \frac{999}{1000}$, $\sqrt{s}t\kappa(\Gamma) = \frac{1}{1000\kappa(\Gamma)^2} \le \frac{1}{1000}$ and $t\kappa(\Gamma)^2 = \frac{1}{1000\sqrt{s}\kappa(\Gamma)}$. It means that we have

$$\begin{aligned}\frac{\sqrt{s}t\kappa(\Gamma)}{1-t\kappa(\Gamma)} + 1 &\le \frac{\frac{1}{1000}}{\frac{999}{1000}} + 1 = \frac{1000}{999} \\ \kappa(\Gamma) + \frac{1}{1-t\kappa(\Gamma)} &\le \kappa(\Gamma) + \frac{1000}{999} \le \frac{1999}{999}\kappa(\Gamma) \\ \frac{t\kappa(\Gamma)^2}{1-t\kappa(\Gamma)} &\le \frac{\frac{1}{1000\sqrt{s}\kappa(\Gamma)}}{\frac{999}{1000}} = \frac{1}{999\sqrt{s}\kappa(\Gamma)}\end{aligned}$$

which implies

$$\mathsf{sign}(Q){\theta'}^\top \Gamma \theta^\Delta \le \frac{1}{100\sqrt{s\sigma_{\max}(\Gamma)}} \|{H'_{[s]}}^\top \theta^\Delta_{[s]}\|_2 = \frac{1}{100\sqrt{s\sigma_{\max}(\Gamma)}} \sqrt{{\theta^\Delta_{[s]}}^\top \Gamma_{[s]} \theta^\Delta_{[s]}} \tag{63}$$

with probability $1 - O(e^{-\Omega(\frac{n}{s\kappa(\Gamma)^6})})$.

From the event of $\mathsf{sign}(Q){\theta'}^\top \Gamma \theta^\Delta \ge 0$, it implies $\mathsf{sign}(Q){\theta'}^\top \Gamma \theta^\Delta = |{\theta'}^\top \Gamma \theta^\Delta|$ and we have

$$|{\theta'}^\top \Gamma \theta^\Delta| \le \frac{1}{100\sqrt{s\sigma_{\max}(\Gamma)}} \sqrt{{\theta^\Delta_{[s]}}^\top \Gamma_{[s]} \theta^\Delta_{[s]}}. \tag{64}$$

By Lemma 15, the probability of (64) is less than

$$O\left(2^s \cdot \left(p^{-\Omega(\frac{1}{\kappa(\Gamma)})} + spe^{-\Omega(\frac{n}{s^2\kappa(\Gamma)^3})}\right)\right)$$

Combining the failure probability of (63), the probability of $\widehat{\mathsf{ne}}^{\lambda_*} = [s] \wedge \mathcal{A} \wedge \mathcal{B}$ is bounded by

$$O\left(2^s \cdot \left(p^{-\Omega(\frac{1}{\kappa(\Gamma)})} + spe^{-\Omega(\frac{n}{s^2\kappa(\Gamma)^3})} + e^{-\Omega(\frac{n}{s\kappa(\Gamma)^6})}\right)\right) < O\left(2^s \cdot \left(p^{-\Omega(\frac{1}{\kappa(\Gamma)})} + spe^{-\Omega(\frac{n}{s^2\kappa(\Gamma)^6})}\right)\right).$$

□

D EXPERIMENT DETAILS

D.1 Set-up

Computing All experiments were conducted on an 8-core Intel Xeon processor E5-2680v4 with 2.40 GHz frequency, and 16GB of memory. All experiments were set three hours wall time limit. Exceeding time limit or undefined FDR value for all zero estimates are marked as NULL for data recording and as missing points for plotting.

Graph Models We include four common graphs: the Band graph, Scale-Free (SF), Erdös-Rényi (ER) and K-Nearest Neighbor (KNN) graphs for GGM simulation. Details for non-Gaussian simulations refer to D.2. The Band and SF graphs were generated via flare Li et al. (2020) package, and the rest ER and KNN were implemented via i-graph Csardi and Nepusz (2006) and mstknnclust Jorge Parraga-Alava et al. (2023) in R. Specifically, graphs are initialized as following

- Band graphs: These are single chain graph given the number of observations n and the number of variables p, and u, v, g are set as default in the r-flare package.
- Scale-free (SF) graphs: These are generated by Barabási–Albert model; The graph is initialized with two connected nodes, and the probability of a new node connecting to one existing node is proportional to the degree of the existing node.
- Erdös-Rény (ER) graphs: These are random undirected graph with n number of nodes, and $p-1$ number of edges, which are selected uniformly and randomly from the set of possible edges.
- K-nearest neighbor (KNN) graphs: There are random graphs where nodes are connected only if they are one of the k-nearest neighbors based on corresponding distance between them. A uniformly distributed $p \times p$ matrix is generated as some random data to calculate euclidean distances between nodes.

Methods For each graph model, each of the following method was implemented provided a penalty parameter.

- Neighbourhood selection (NS): The neighbourhood selection method in Meinshausen and Bühlmann (2006) was implemented, and code is available at: `https://anonymous.link`.
- Graphical Lasso (Glasso) in Friedman et al. (2008) was implemented based on the glasso R package Friedman et al. (2019), and code is available at: `https://github.com/cran/glasso`.
- Constrained ℓ_1-minimization for inverse matrix estimation (CLIME) in Cai et al. (2011) was implemented based on the flare R package Li et al. (2020), and mirror code is available at: `https://github.com/cran/clime`.
- Tuning-Insensitive Graph Estimation and Regression (TIGER) in Liu and Wang (2017) was implemented based on flare R package Li et al. (2020), and mirror code is available at: `https://github.com/cran/tiger`.

Metrics For each graph model, we evaluate the performance of each method with the penalty parameter selected by the following criteria: Akaike information criterion (AIC, Akaike, 1974), Bayesian information criterion (BIC, Schwarz, 1978), extended Bayesian information criterion (EBIC, Foygel and Drton, 2010), and 5-fold cross-validation (CV). The performance was measured by the following metrics:

- Structured Hamming distance (SHD): The number of edge insertions, deletions or flips (in directed graph) that is needed to transform the estimated graph to the true graph.
- True Positive Rate (TPR):The proportion of correctly identified edges to the total number of edges in the true graph.
- False Discovery Rate (FDR): The proportion of incorrectly identified edges to the total number of edges in the estimated graph.

Remarks Experiment jobs were auto-written given configs (the graph type, the method, n and p) and submitted via slurm. Experiments and analysis were conducted using R (R Core Team (2021)) and full code can be found at: `https://github.com/zhao-lyu/GGM`.

D.2 Non-Gaussian Simulation

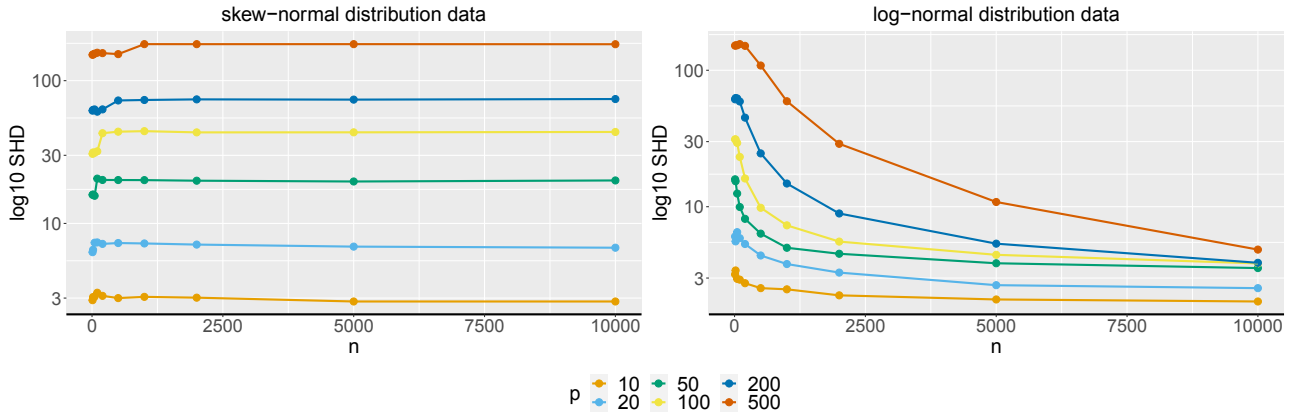

Figure 6: Log-SHD vs. sample size n and the number of variables p on Skew-normal and log-normal data via CV. The bottom black line denotes zero SHD, which is never attained.

Although our theoretical results are specific to the Gaussian setting, we can also demonstrate the fallibility of CV for non-Gaussian data. Specifically, we randomly generate n i.i.d samples $\mathbf{X} \in \mathbb{R}^{n\times p}$ from skew-normal and log-normal distributions via sn Azzalini (2023) and MASS packages Venables and Ripley (2002) in R, respectively. The true coefficients $\theta_i \sim \text{Unif}([-2,-1]\cup[1,2])$ if it's in the neighborhood, otherwise, we set it to zero. The response variable is set by $\mathbf{Y} = \mathbf{X}\theta + \varepsilon$, with $\varepsilon \sim N(0, 0.1)$. We implement glmnet Friedman et al. (2010) package for R to test with 5-fold CV to select the penalty parameter and obtain the estimates. We repeat this 100 times, and take the average SHD for each n and p. Although Figure 6 shows a decreasing trend in SHD with increasing n for all p, CV plateaus and never achieves perfect selection (highlighted with the black line for zero SHD). Even when $p = 10$ and $n = 10000$, CV fails to correctly select all neighbors. This once again indicates that CV is suboptimal for structure learning.

D.3 GGM Simulation

For each graph type, we construct the covariance and precision matrices Σ^* and K^*, and simulate data generated from $N(0, \Sigma^*)$. We choose the largest penalty parameter $\lambda_{\max}$ (the smallest value which will result in a null, no-edge model) and the smallest penalty parameter $\lambda_{\min}$ (the largest value which will result in a model whose number of edge is less than $2\|K^*\|_1$). 100 penalty parameters λ are chosen logarithmically evenly spaced between $\lambda_{\max}$ and $\lambda_{\min}$. At each penalty λ, an estimate $\widehat{K}_\lambda$ is computed via a given algorithm (NS, Glasso, CLIME or TIGER), and is used to model the edge set E_λ. We next compute the unpenalized maximum likelihood estimate with the same support as E_λ. We can compute all criteria and choose λs corresponding to the lowest criteria among all λs. This is repeated for 10 trials for each combination of n, p, graph type and algorithms.

We perform 5-fold CV with generated data. For each λ and each training set and test set, we fit the model with training set using glasso, and evaluate the performance via samples in the test set by calculating the Gaussian log-likelihood. The penalty value with the maximum averaged Gaussian log-likelihood is selected as λ_{CV}. We measure its performance via the average SHD, TPR and FDR to compare with other criteria.

D.4 Additional Results

As noted in Corollary 2, the sparsity level s also plays a role in determining the probability of correct recovery of the neighborhood. To emprically illustrate this point, we set $p = 50, 500$ and simulated 5000 data from a Gaussian linear model. The sparsity s is enforced by randomly setting s out of p coefficients to be exactly 0. The result shown in Figure 7 indeed confirms our theoretical result: a larger s implies a lower chance of recovery.

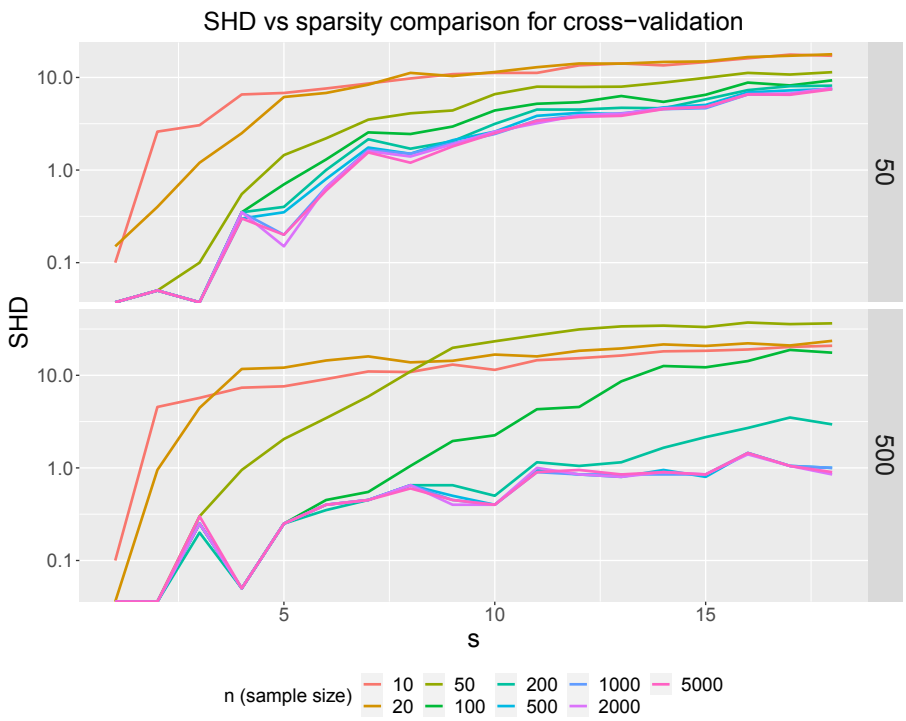

Figure 7: SHD vs sparsity comparison for CV for $p = 50$ and 50 respectively for varying n.

Below we provide remaining experimental results for 10 runs when utilizing NS, Glasso, CLIME, and TIGER; see Figures 8, 9, 10, 11 respectively. For each run, a random seed is set to generate varying datasets. We also provide detailed numerical results of FDR in Table 1 and average SHD in Table 2 for NS method tuned by CV for the Band graph, which clearly suggests CV neither reaches 0% FDR nor obtain fully correct graph (0 SHD).

To further investigate the behaviour of CV for large n, and to verify that it indeed fails to achieve exact recovery (i.e. zero average SHD), we provide additional experimental results with larger n over 100 runs in Figures 12, 13, 14 and 15.

	$n = 10$	$n = 20$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 800$	$n = 1000$
$p = 100$	0.587	0.369	0.0782	0.0329	0.0289	0.0158	0.00796	0.0119
$p = 200$	0.558	0.327	0.0738	0.0256	0.0148	0.0099	0.00792	0.0089
$p = 500$	0.674	0.333	0.0602	0.0176	0.0068	0.0052	0.00597	0.0051

Table 1: Average FDR for the Band graph via NS method tuned by CV

	n=10	n=20	n=50	n=100	n=200	n=500	n=800	n=1000
p=100	103.2	81.8	20.2	3.4	3.0	1.6	0.8	1.2
p=200	202.8	168.4	47.2	6.8	3.0	2.0	1.6	1.8
p=500	510.0	454.2	151.4	19.2	3.4	2.6	3.0	2.6

Table 2: Average SHD for the Band graph via NS method tuned by CV

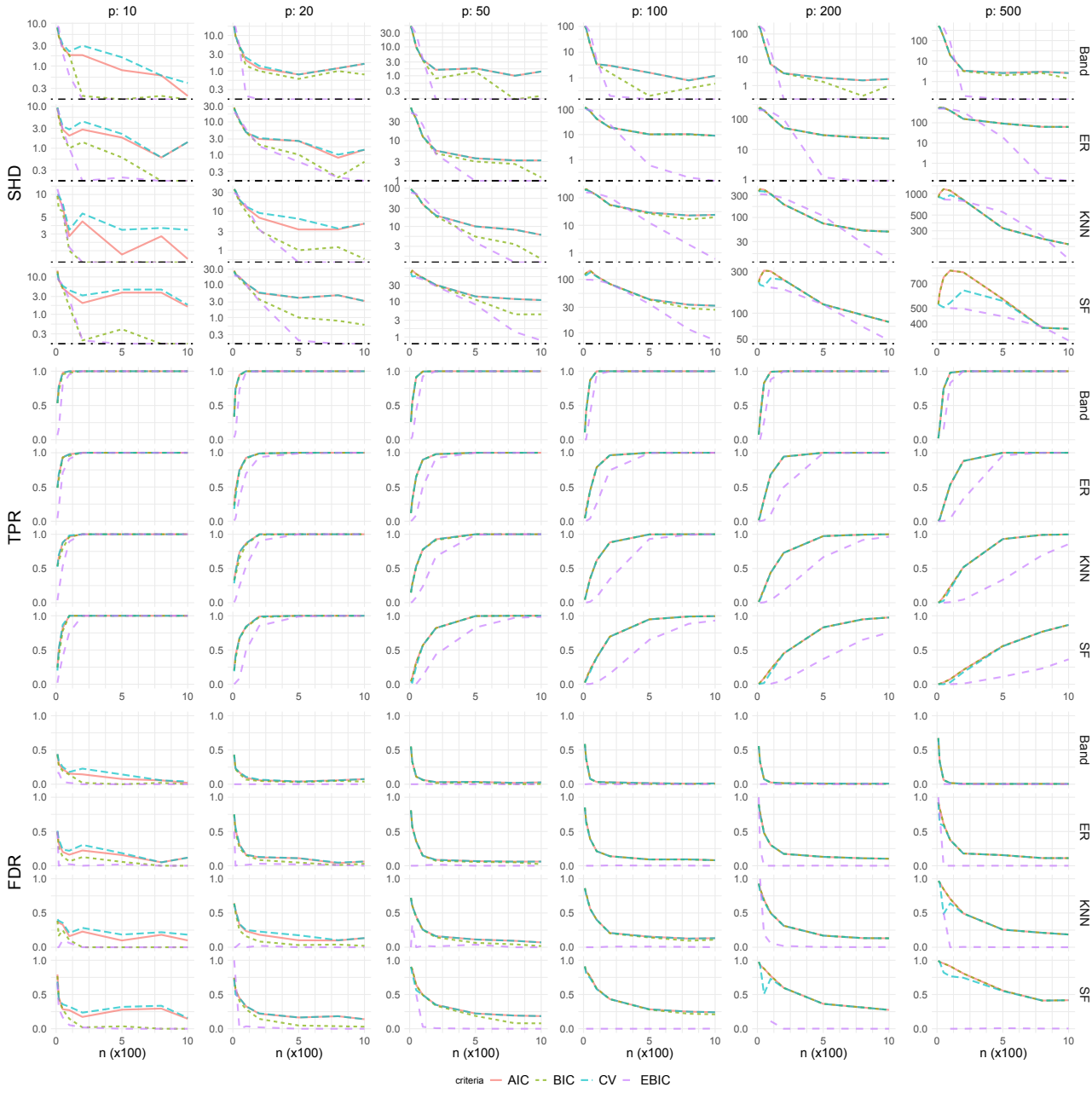

Figure 8: Average SHD, TPR and FDR over varying sample size n (in hundreds) and the number of variables p on groups of graph types using Neighbourhood Selection (NS) to compare criteria. The dotdash line represents the 0-SHD, i.e. perfect neighborhood selection. Wall time limit was set to three hours. Exceeding time limit or undefined FDR value for all zero estimates are marked as missing points for plotting.

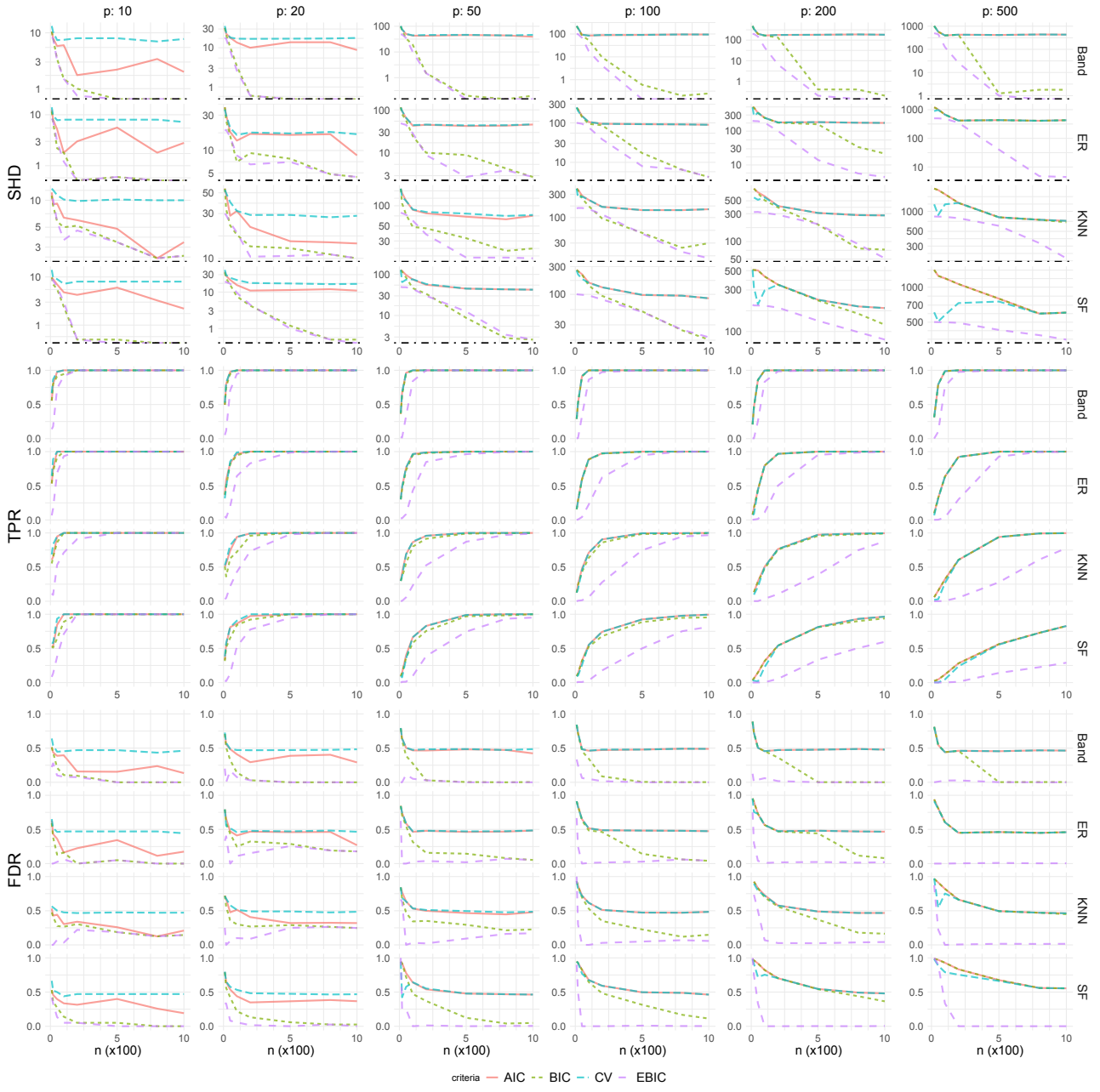

Figure 9: Average SHD, TPR and FDR over varying sample size n (in hundreds) and the number of variables p on groups of graph types using Glasso to compare criteria. The dotdash line represents the 0-SHD, i.e. perfect neighborhood selection. Wall time limit was set to three hours. Exceeding time limit or undefined FDR value for all zero estimates are marked as missing points for plotting.

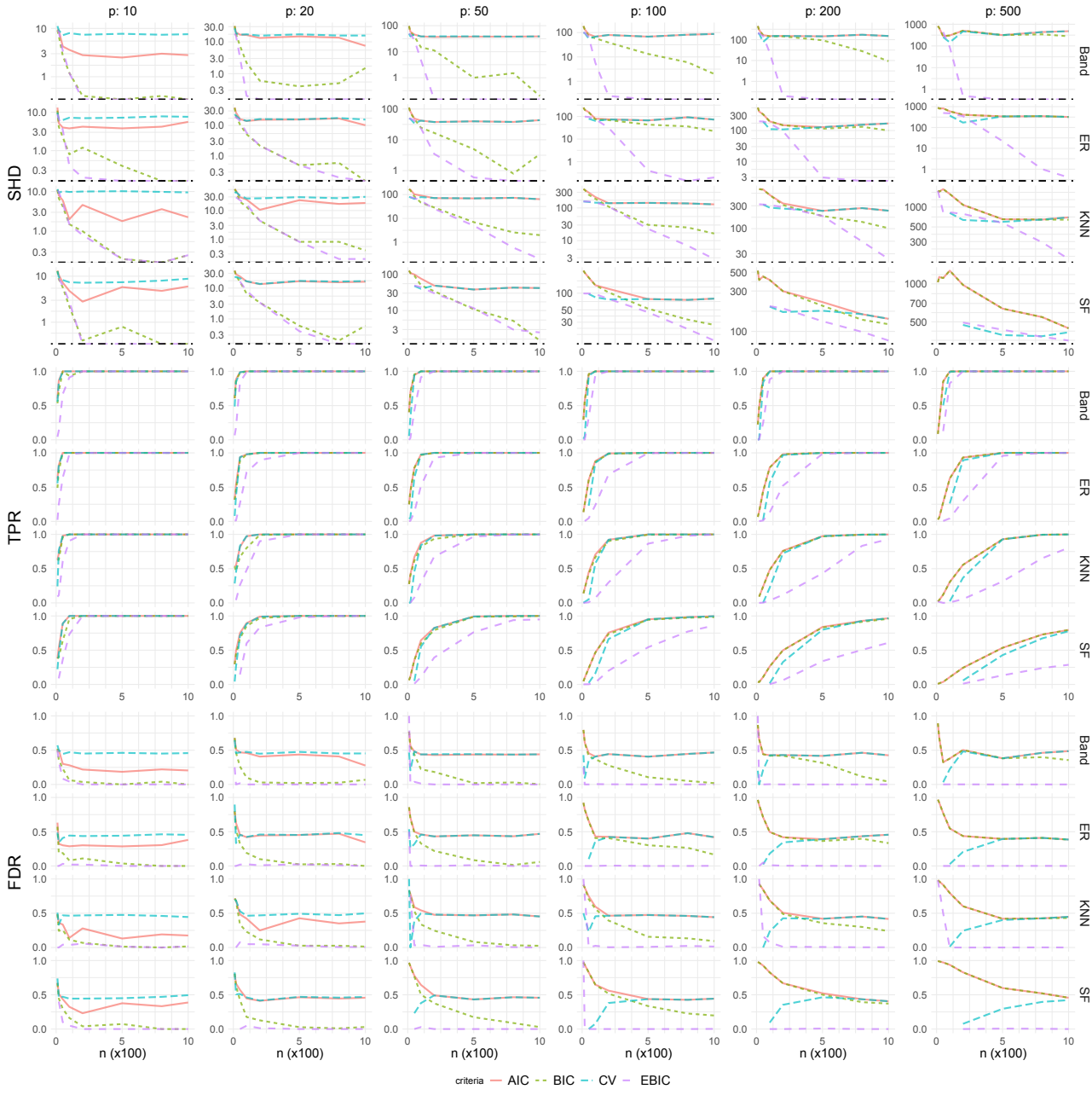

Figure 10: Average SHD, TPR and FDR over varying sample size n (in hundreds) and the number of variables p on groups of graph types using CLIME to compare criteria. The dotdash line represents the 0-SHD, i.e. perfect neighborhood selection. Wall time limit was set to three hours. Exceeding time limit or undefined FDR value for all zero estimates are marked as missing points for plotting.

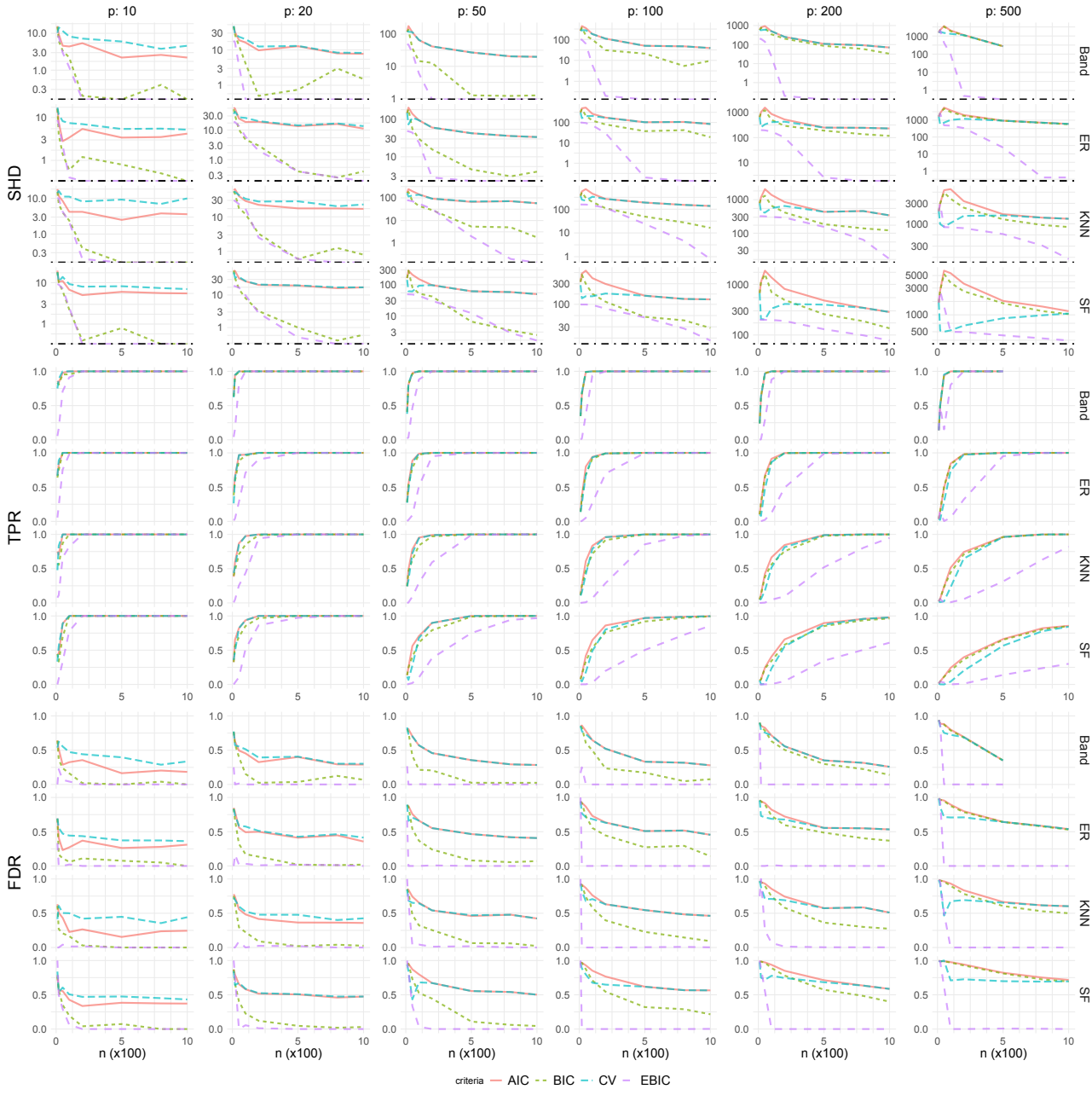

Figure 11: Average SHD, TPR and FDR over varying sample size n (in hundreds) and the number of variables p on groups of graph types using TIGER to compare criteria. The dotdash line represents the 0-SHD, i.e. perfect neighborhood selection. Wall time limit was set to three hours. Exceeding time limit or undefined FDR value for all zero estimates are marked as missing points for plotting.

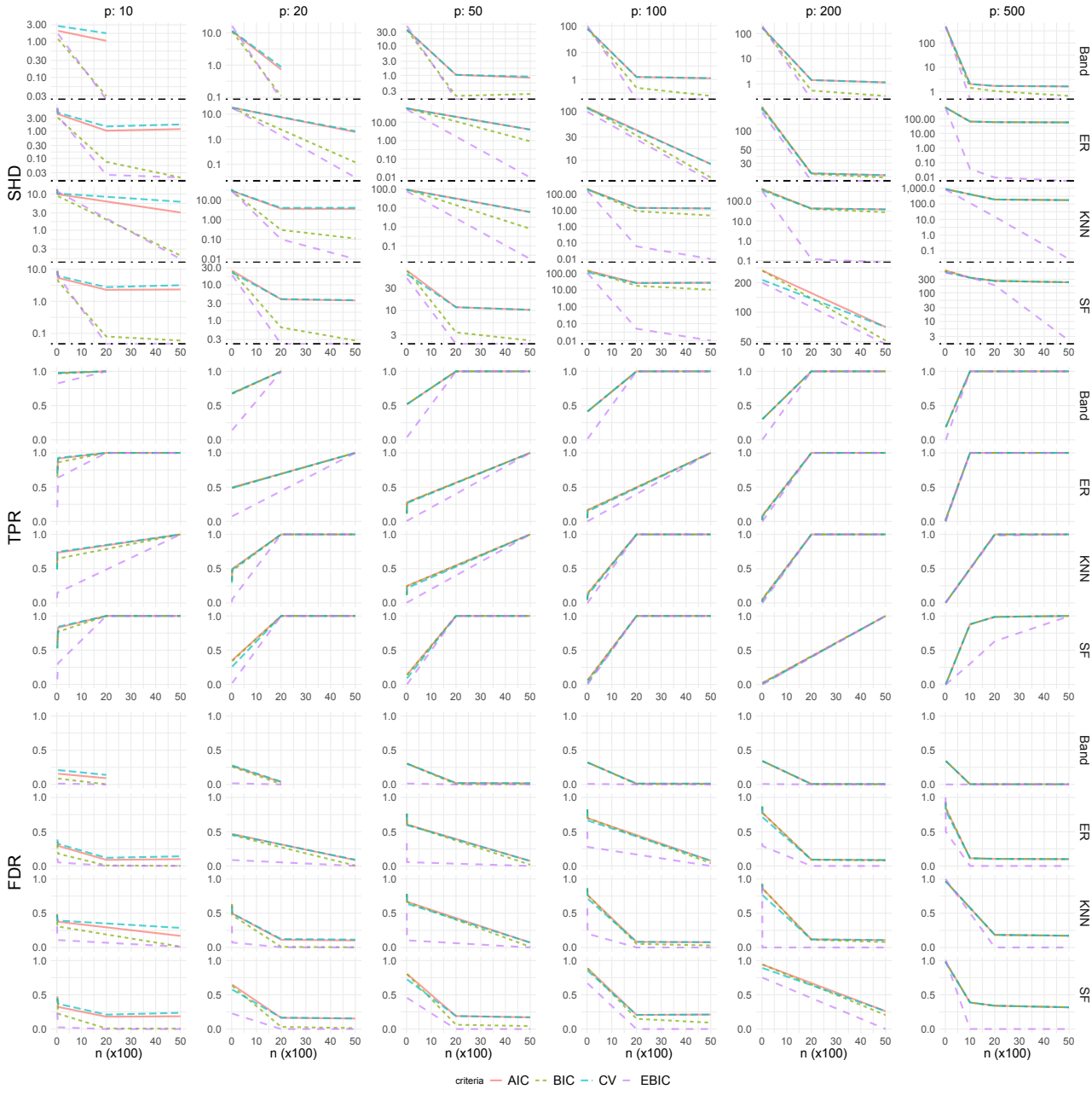

Figure 12: Average SHD, TPR and FDR over varying sample size n (in hundreds) and the number of variables p on groups of graph types using Neighbourhood Selection (NS) to compare criteria on 100 runs. The dotdash line represents the 0-SHD, i.e. perfect neighborhood selection. Wall time limit was set to three hours. Exceeding time limit or undefined FDR value for all zero estimates are marked as missing points for plotting.

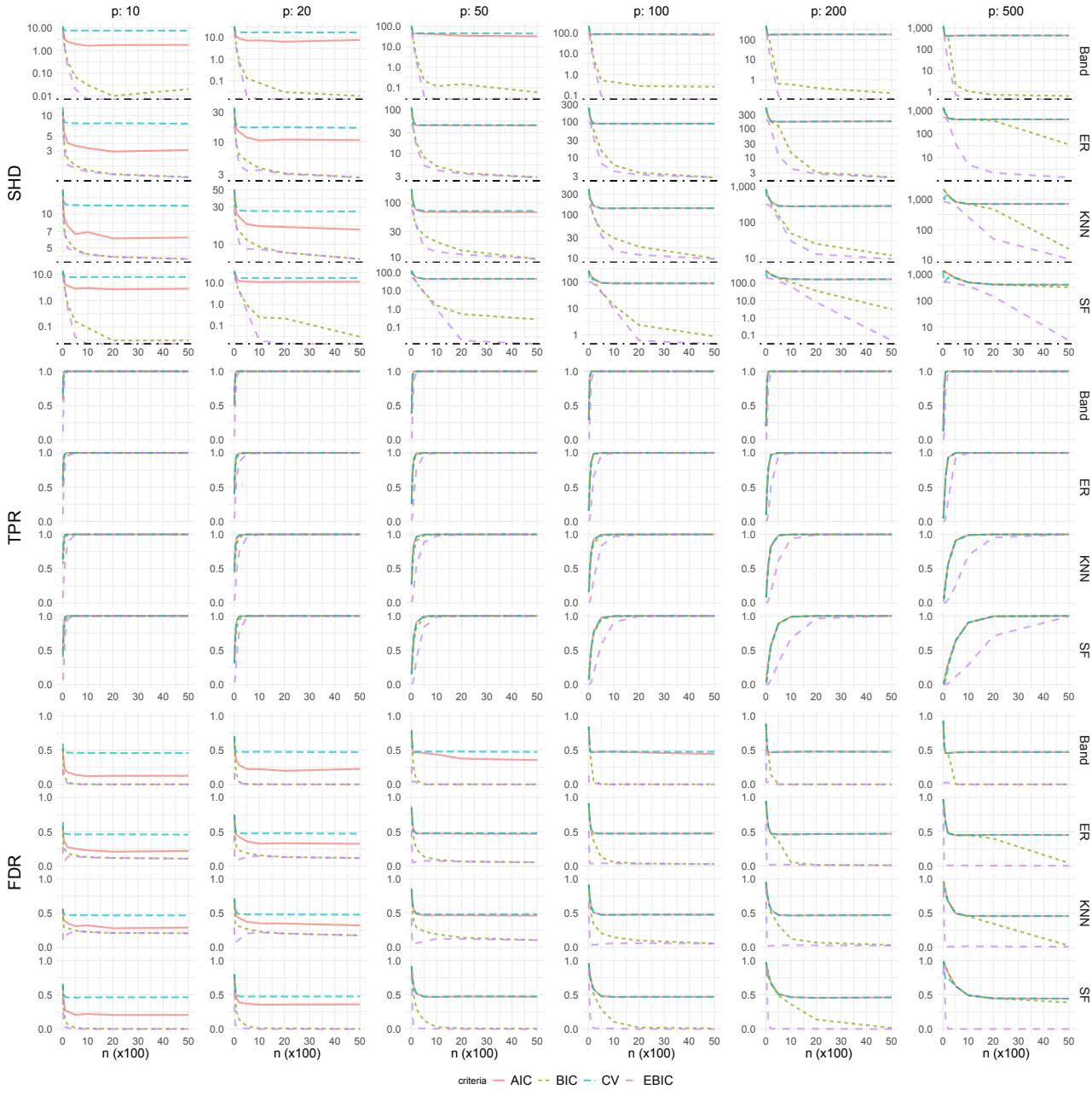

Figure 13: Average SHD, TPR and FDR over varying sample size n (in hundreds) and the number of variables p on groups of graph types using Glasso to compare criteria on 100 runs. The dotdash line represents the 0-SHD, i.e. perfect neighborhood selection. Wall time limit was set to three hours. Exceeding time limit or undefined FDR value for all zero estimates are marked as missing points for plotting.

Figure 14: Average SHD, TPR and FDR over varying sample size n (in hundreds) and the number of variables p on groups of graph types using CLIME to compare criteria on 100 runs. The dotdash line represents the 0-SHD, i.e. perfect neighborhood selection. Wall time limit was set to three hours. Exceeding time limit or undefined FDR value for all zero estimates are marked as missing points for plotting.

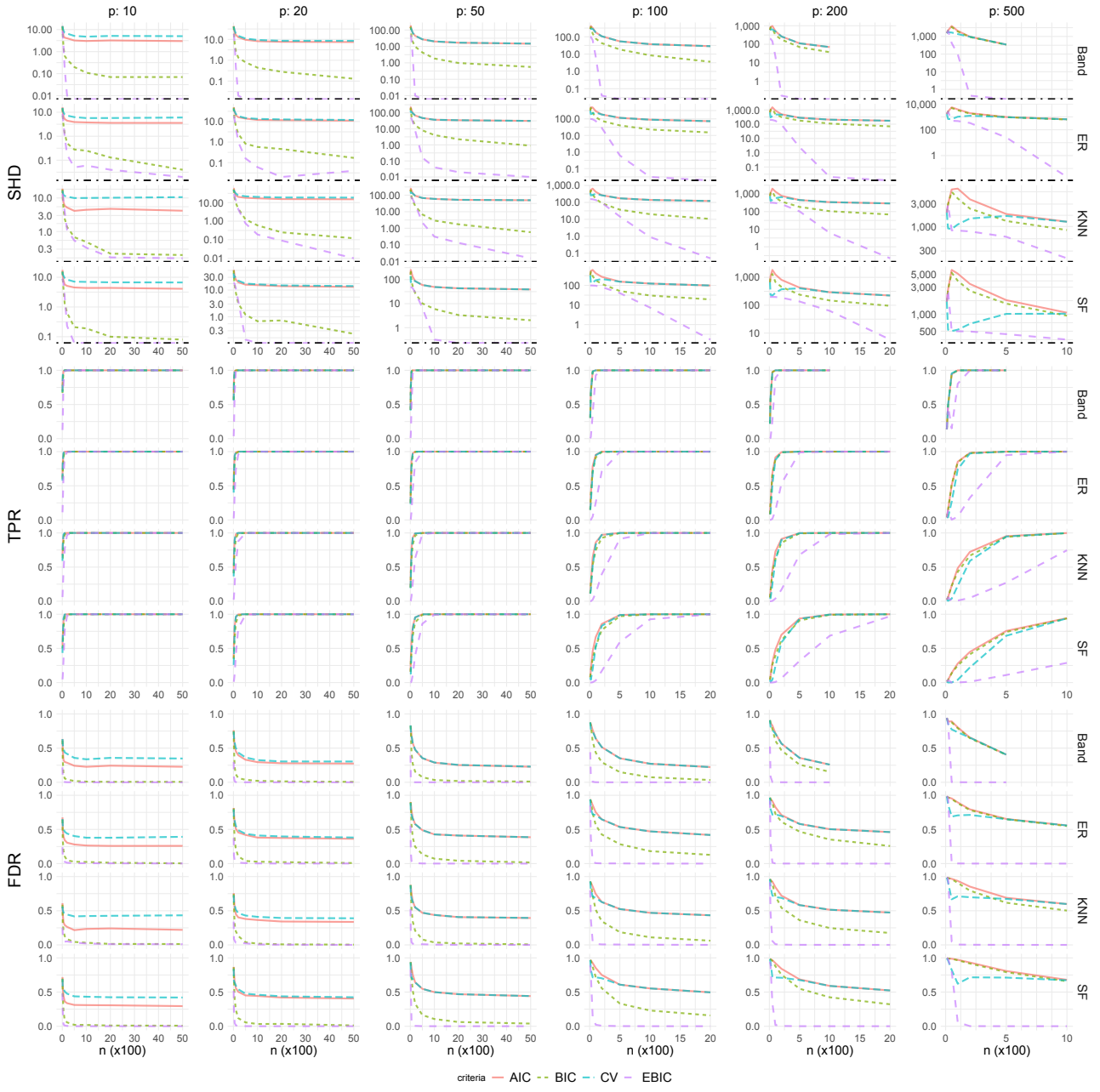

Figure 15: Average SHD, TPR and FDR over varying sample size n (in hundreds) and the number of variables p on groups of graph types using TIGER to compare criteria on 100 runs. The dotdash line represents the 0-SHD, i.e. perfect neighborhood selection. Wall time limit was set to three hours. Exceeding time limit or undefined FDR value for all zero estimates are marked as missing points for plotting.