UCLA
UCLA Previously Published Works

Title

Learning Directed Acyclic Graphs with Penalized Neighbourhood Regression

Permalink

https://escholarship.org/uc/item/8jn5v3d2

Authors

Aragam, Bryon
Amini, Arash A
Zhou, Qing

Publication Date

2015-11-28

LEARNING DIRECTED ACYCLIC GRAPHS WITH PENALIZED NEIGHBOURHOOD REGRESSION

BRYON ARAGAM,[1] ARASH A. AMINI,[2] AND QING ZHOU[3]

University of California, Los Angeles

ABSTRACT. We study a family of regularized score-based estimators for learning the structure of a directed acyclic graph (DAG) for a multivariate normal distribution from high-dimensional data with $p \gg n$. Our main results establish support recovery guarantees and deviation bounds for a family of penalized least-squares estimators under concave regularization without assuming prior knowledge of a variable ordering. These results apply to a variety of practical situations that allow for arbitrary nondegenerate covariance structures as well as many popular regularizers including the MCP, SCAD, ℓ_0 and ℓ_1. The proof relies on interpreting a DAG as a recursive linear structural equation model, which reduces the estimation problem to a series of neighbourhood regressions. We provide a novel statistical analysis of these neighbourhood problems, establishing uniform control over the superexponential family of neighbourhoods associated with a Gaussian distribution. We then apply these results to study the statistical properties of score-based DAG estimators, learning causal DAGs, and inferring conditional independence relations via graphical models. Our results yield—for the first time—finite-sample guarantees for structure learning of Gaussian DAGs in high-dimensions via score-based estimation.

1. INTRODUCTION

The problem of learning directed acyclic graphs from observational data has received increased attention recently due to its many applications in causal inference, genomics, machine learning, and theoretical computer science. While there are several competing strategies for solving this problem, a well-established line of work starting with Heckerman et al. (1995) has illustrated that so-called *score-based learning* outperforms existing approaches and scales to high-dimensional datasets with thousands of variables (Aragam and Zhou, 2015; Ramsey et al., 2016). The basic idea behind score-based learning is to formulate the problem as a statistical M-estimation problem of the form

$$(1.1) \qquad \widehat{B} \in \underset{B \in \mathbb{D}_p}{\arg\min}\, Q(B),$$

E-mail address: [1]naragam@cs.cmu.edu, [2]aaamini@stat.ucla.edu, [3]zhou@stat.ucla.edu.

where $\mathbb{D}_p$ is the set of $p \times p$ matrices representing the weighted adjacency matrix of a directed acyclic graph (DAG) and Q is a *score function*, usually combining a data-dependent loss function, based on a sample of size n, and a model complexity regularizer.

Unfortunately, due to the nonconvex and combinatorial nature of the program (1.1), very little is known about the statistical properties of score-based estimators in the high-dimensional regime with $p \gg n$. In contrast, the development and application of algorithms for high-dimensional score-based learning has surged in recent years. This has created a gap in the literature, whereby score-based methods based on (1.1) are commonly used by practitioners despite an absence of guarantees regarding the quality of such estimators. A goal of this paper is to study the statistical properties of $\widehat{B}$, motivated by applications to the theory of high-dimensional M-estimation, causal inference, and graphical modeling. Our main results establish—for the first time—structure learning and estimation guarantees for the score-based estimator $\widehat{B}$ in high-dimensions that avoid commonly used but uncheckable assumptions found in the literature, giving a sense of what happens when a score-based method is naïvely applied to high-dimensional datasets.

1.1. **Background.** Our approach is based on the well-known structural equation model (SEM) interpretation of Gaussian DAGs (Wright, 1934; Drton et al., 2011). Suppose $X = (X_1, \dots, X_p)$ is a random vector satisfying

$$(1.2) \qquad X = \widetilde{B}^T X + \widetilde{\varepsilon}, \quad \widetilde{\varepsilon} \sim \mathcal{N}_p(0, \widetilde{\Omega}),$$

where $\widetilde{B} \in \mathbb{D}_p$ and $\widetilde{\Omega}$ is a $p \times p$ positive diagonal matrix of variances. One can interpret $\widetilde{B}$ as the weighted adjacency matrix of a graph. We will say that $\widetilde{B}$ is, or represents, a DAG if this graph is directed and acyclic. Given an $n \times p$ random matrix $\mathbf{X}$ whose rows are i.i.d. drawn according to the model (1.2), define a penalized least-squares (PLS) score function by

$$(1.3) \qquad Q(B) = \frac{1}{2n} \|\mathbf{X} - \mathbf{X}B\|_2^2 + \rho_\lambda(B),$$

where $\|\cdot\|_2$ is the matrix Frobenius norm and ρ_λ belongs to a class of regularizers, parametrized by the *regularization parameter* λ. The corresponding score-based estimator, defined by (1.1), is the main focus of this article. In the sequel, $\widehat{B}$ always refers to the estimator defined by using (1.3) in (1.1).

The objective function (1.3) is not new—there is a long line of previous work based on the same objective function (e.g. Schmidt et al., 2007; Shojaie and Michailidis, 2010; Xiang and Kim, 2013; Han et al., 2016). Unfortunately, these works are mainly empirical and leave many open questions regarding the statistical properties of $\widehat{B}$ unanswered. Given its popularity in practice, our goal here is to analyze the statistical properties of $\widehat{B}$.

1.2. **Motivation.** The decomposition (1.2) is often referred to as a linear *structural equation model* for X, and extends well beyond the Gaussian setting considered here. Structural equations have a long history in the

statistical literature, dating all the way back to Wright (1921, 1934). More recently, structural equations have become an essential quantitative analysis tool in the social sciences (Yuan and Bentler, 2007). Structural equations have also played an important role in machine learning and causal inference (Bühlmann et al., 2014; Mooij et al., 2014) as well as covariance selection in statistics (Pourahmadi, 2013; Wermuth, 1980). These models have been applied recently in biology (Cai et al., 2013), network modeling (Horvath, 2011), and psychology (Guay et al., 2015; Morin et al., 2015).

Motivated by these applications and recent interest in developing fast algorithms for learning DAGs (Schmidt et al., 2007; Teyssier and Koller, 2012; Fu and Zhou, 2013; Xiang and Kim, 2013; Aragam and Zhou, 2015; Gu et al., 2016; Ramsey et al., 2016), we seek to address three fundamental questions regarding score-based estimators:

(1) *Score-based learning.* It is well-known that DAG models are nonidentifiable (Section 2.1)—nonetheless, is it possible to obtain statistical guarantees for $\widehat{B}$ when $p \gg n$?

(2) *Causal DAG learning.* In some situations, there exists a unique causal DAG that represents the true, underlying distribution. In such situations, can $\widehat{B}$ recover the structure of the causal DAG?

(3) *Conditional independence learning.* What can be said about the conditional independence (CI) relations inferred by $\widehat{B}$, particularly when there is no faithful DAG that represents the joint distribution?

Crucially, in all three problems, we are interested in *structure learning* (also known as *support recovery* and *model selection consistency*), i.e. recovering the correct graph. In graphical modeling, stucture learning is of fundamental importance since it is a critical step for learning the conditional independence structure of the underlying distribution; ℓ_2-consistency is insufficient for this purpose. Typically, the *faithfulness* assumption is needed (Section 5.3), which is a strong assumption that rarely holds in practice (Lin et al., 2014; Uhler et al., 2013). Our analysis avoids this assumption entirely, and to the best of our knowledge, provides the first structure learning guarantees for DAGs *without faithfulness* in high-dimensions.

The organization of the rest of this paper is as follows: In the rest of this section, we review previous work and outline our main contributions. Section 2 provides some intuition behind the estimator (1.1) and covers necessary preliminaries. Section 3 sets up a framework for neighbourhood regression that is fundamental to our main results. In Section 4, we present our main results, establishing uniform nonasymptotic bounds on the probability of false selection and the estimation error of a family of penalized least-squares estimators. In Section 5, we apply these results to answer each of the three questions outlined above. Finally, we compare our results to the literature in Section 6 before concluding with some extensions and generalizations in Section 7. As the proofs are quite involved and invoke novel

technical machinery, we outline the main ideas in Section 8 and postpone detailed proofs of the various technical results to the Appendix.

1.3. **Previous work.** Despite substantial methodological and algorithmic progress towards solving (1.1) when p is large, theoretical progress in high-dimensions has been slower. Chickering (2003) showed that in a low-dimensional setting with p fixed, score-based learning is sufficient to learn a so-called *faithful* DAG for X in the asymptotic limit $n \to \infty$, where n is the number of i.i.d. samples. The first truly high-dimensional results for DAG learning were for the PC algorithm (Kalisch and Bühlmann, 2007), however, this algorithm is not based on (1.1) and requires strong assumptions such as faithfulness. van de Geer and Bühlmann (2013) proved the first such results for a score-based estimator, establishing bounds on the ℓ_2-error and the number of edges in $\widehat{B}$ for a thresholded ℓ_0-penalized maximum likelihood estimator. Notably, their results fall short of providing consistency in structure learning. One of the main contributions of the current work is a unified regression framework for analyzing score-based estimators that provides guarantees on support recovery, parameter estimation, and sparsity. Furthermore, our work covers a wide spectrum of nonconvex regularizers including both ℓ_0 and ℓ_1 as boundary cases. A more detailed comparison to the aforementioned results is provided in Section 6.

Finally, we note that—perhaps surprisingly—there has been much more theoretical progress for nonlinear and non-Gaussian models (Peters et al., 2012, 2014; Mooij et al., 2014). There has also been progress using multi-stage methods that separate the learning procedure into several decoupled steps (Loh and Bühlmann, 2014; Bühlmann et al., 2014). These results are largely due to various assumptions that guarantee identifiability (e.g. Shimizu et al., 2006; Hoyer et al., 2009; Zhang and Hyvärinen, 2009). Under such assumptions, the analysis becomes substantially easier. By contrast, the present work studies a fully nonidentifiable model as in van de Geer and Bühlmann (2013), and as such the main complication in the theoretical analysis is dealing with the existence of equivalent DAGs as discussed in Section 2.1. Furthermore, our approach via score-based learning consists of a single learning step that effectively combines neighbourhood search, order search, and regression by minimizing a single penalized loss function.

1.4. **Contributions.** Our main results will be organized into two sections: In Section 4, we develop our technical results for neighbourhood regression problems, and then in Section 5 we apply these results to address the three main questions raised in Section 1.2.

On the technical side, we provide simultaneous, uniform control over the class of DAGs satisfying (1.2) for a joint Gaussian distribution, including both support recovery guarantees (Section 4.2) and deviation bounds (Section 4.3), by leveraging a neighbourhood regression interpretation of the program (1.1). As the cardinality of the class of such DAGs is $O(p!)$, it is very challenging to obtain any uniform control result. We overcome this

technical difficulty by introducing a novel notion of monotonicity among the neighbourhood regressions associated with a DAG. This proof technique is completely different from existing work and makes an explicit connection between graphical models and existing concepts from the regression literature. In this way, it sheds new light on the interplay between the covariance matrix Σ and its regression coefficients through the intuitive concept of neighbourhood regression.

Besides the technical novelty, the significance of this work is seen from its direct contributions to the following three statistical applications which motivate our analysis:

(1) *Score-based learning* (Section 5.1). Without any identifiability assumptions, we show that $\widehat{B}$ consistently estimates the structure of a sparse DAG representation of the joint Gaussian distribution via (1.2), thus providing useful theoretical guarantees for the practical application of score-based DAG learning methods. Our result is intriguing more generally: Whereas much of the literature on high-dimensional statistics has focused on identifiable models, our results indicate that much of this theory carries over to the nonidentifiable setting, and more importantly, it is possible to *adaptively* estimate a well-behaved parameter instead of an arbitrary parameter.

(2) *Causal DAG learning* (Section 5.2). Under an identifiability assumption, we establish that $\widehat{B}$ can be used to learn the "true" causal DAG. We generalize existing results on estimating causal DAGs with equal error variances to what we call *minimum-trace* DAGs. Furthermore, we strengthen existing results to include support recovery in the high-dimensional setting.

(3) *Conditional independence learning* (Section 5.3). We show that it is possible to use our framework to consistently estimate the full set of CI relations in a Gaussian distribution, even in the absence of a faithful DAG representation. Detecting CIs among a set of random variables from data is a key motivation for structure learning of graphical models, including DAGs. Achieving this task without the faithfulness assumption distinguishes our work from the existing literature on graphical models.

Notwithstanding these interesting applications to causal inference and graphical models, the first application is interesting in and of itself from a purely statistical standpoint: $\widehat{B}$ is a popular, useful estimator and yet we have an extremely limited theoretical understanding of its properties. In this paper we attempt to provide a comprehensive portrait of the behaviour of $\widehat{B}$, providing much needed justification—and caution—for its use in applications.

Notation and terminology. Universal constants which may not be the same from line to line will be denoted by c_1, c_2, etc. The standard ℓ_q norms for $q \in [0, \infty]$ will be denoted by $\|\cdot\|_q$. For a matrix $A \in \mathbb{R}^{n \times p}$, $\|A\|_q$ is its ℓ_q norm, viewing A as a vector in $\mathbb{R}^{np}$, and so, in particular, the Frobenius norm on matrices is denoted by $\|\cdot\|_2$. The maximum and minimum eigenvalues of a matrix A are denoted by $r_{\max}(A)$ and $r_{\min}(A)$, respectively. The support

of $A = (a_{ij})$ is defined by $\mathrm{supp}(A) := \{(i,j) : a_{ij} \neq 0\}$ and the cardinality of a set by $|\cdot|$. We also make use of the column notation $A = [\, a_1 \mid \cdots \mid a_p \,] \in \mathbb{R}^{n \times p}$ to denote the columns of A. Given a permutation π, P_π denotes the associated permutation operator on matrices: For any matrix A, $P_\pi A$ is the matrix obtained by permuting the rows and columns of A according to π, so that $(P_\pi A)_{ij} = a_{\pi(i)\pi(j)}$. For any integer m, we define $[m] = \{1, \ldots, m\}$ and $[m]_j = [m] - \{j\}$. For a vector $v \in \mathbb{R}^m$ and a subset $S \subset [m]$, we let $v_S \in \mathbb{R}^{|S|}$ denote the restriction of v to the components in S. For a matrix $A \in \mathbb{R}^{n \times m}$, $A_S \in \mathbb{R}^{n \times |S|}$ denotes *column-wise* restriction to the columns in S. Given two quantities X and Y, which may depend on n and Σ, we will write $X \lesssim Y$ to mean there exists a constant $a > 0$—independent of n and Σ—such that $X \leq aY$, and analogously for $X \gtrsim Y$. A positive definite matrix is denoted as $\Sigma \succ 0$.

We write $\mathbf{X} \stackrel{\text{iid}}{\sim} \mathcal{N}_p(0, \Sigma)$ for a random matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ to mean that the *rows* are i.i.d. draws from $\mathcal{N}_p(0, \Sigma)$. Boldface type is reserved for *random* quantities that depend on the sample size n; for example a random matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ or the response $\mathbf{y} \in \mathbb{R}^n$ in a linear model. Events defined on some probability space will usually be denoted by script font, e.g. $\mathcal{E}$, $\mathcal{F}$, etc.

2. Preliminaries

In this section we discuss some preliminary notions and develop some intuition for the problem we study. Throughout this section, it is important to keep in mind our main goal: To study the consistency of the estimator $\widehat{B}$ by decomposing the estimation problem into a series of simpler, neighbourhood regression problems. The present section sets the stage, while the technical details of neighbourhood regression are developed in the next section (Section 3). The setup outlined here, including much of the notation, is the same as van de Geer and Bühlmann (2013). In other words, we work with the same statistical model as in van de Geer and Bühlmann (2013), however, our results are more general and our proof technique is completely different; for a comparison, see Section 6.

2.1. **Nonidentifiability and permutation equivalence.** A fundamental property of the model (1.2) is that the parameters $(\widetilde{B}, \widetilde{\Omega})$ are statistically nonidentifiable. It follows from (1.2) that $X \sim \mathcal{N}_p(0, \Sigma(\widetilde{B}, \widetilde{\Omega}))$, where

$$(2.1) \qquad \Sigma(\widetilde{B}, \widetilde{\Omega}) := (I - \widetilde{B})^{-T} \widetilde{\Omega} (I - \widetilde{B})^{-1}.$$

This yields a parametrization of a multivariate normal distribution in terms of a DAG $\widetilde{B}$ and its conditional variances $\widetilde{\Omega}$. As it turns out, the mapping $(\widetilde{B}, \widetilde{\Omega}) \mapsto \Sigma(\widetilde{B}, \widetilde{\Omega})$ is not one-to-one. In fact, given $\Sigma \succ 0$, we can show that for each order of the variables, there is a DAG $(\widetilde{B}, \widetilde{\Omega})$ such that $\Sigma(\widetilde{B}, \widetilde{\Omega}) = \Sigma$. This is most easily illustrated with an example:
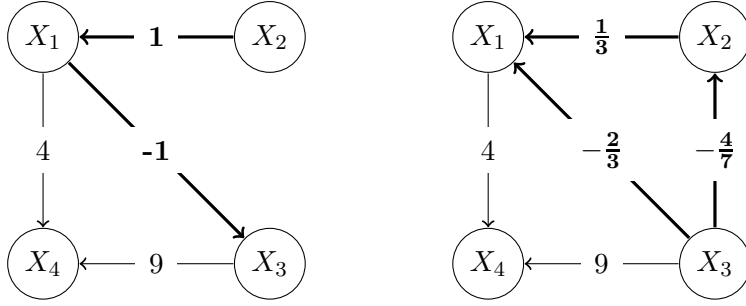
FIGURE 1. Two equivalent DAGs with different edge weights, as indicated by the edges highlighted in bold.

Example 2.1. Consider the following covariance matrix:

$$
(2.2) \qquad \Sigma = \begin{pmatrix} 6 & 4 & -6 & -30 \\ 4 & 4 & -4 & -20 \\ -6 & -4 & 7 & 39 \\ -30 & -20 & 39 & 234 \end{pmatrix}.
$$

Given Σ, we can construct matrices $(\widetilde{B}, \widetilde{\Omega})$ that satisfy (1.2) by the following procedure: First, project X_1 onto X_2, then project X_3 onto X_2 and X_1, and finally project X_4 onto X_2, X_1, and X_3. This induces an ordering $\prec$ on the variables given by $X_4 \prec X_3 \prec X_1 \prec X_2$, wherein each X_j is projected onto the nodes after it under $\prec$. This ordering induces a permutation π_1 defined by $\pi_1(1) = 4$, $\pi_1(2) = 3$, $\pi_1(3) = 1$, $\pi_1(4) = 2$. The coefficients of each linear projection and the resulting residual variances lead to a set of structural equations as in (1.2). Denoting the parameters obtained in this way by $\widetilde{B} = \widetilde{B}(\pi_1)$ and $\widetilde{\Omega} = \widetilde{\Omega}(\pi_1)$, we obtain the following parameters for (2.2):

$$
(2.3) \qquad \widetilde{B}(\pi_1) = \begin{pmatrix} 0 & 0 & -1 & 4 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 9 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad \widetilde{\Omega}(\pi_1) = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix}.
$$

This DAG is depicted on the left in Figure 1.

If we choose a different ordering, say $X_4 \prec X_1 \prec X_2 \prec X_3$, we obtain a different set of parameters:

$$
(2.4) \qquad \widetilde{B}(\pi_2) = \begin{pmatrix} 0 & 0 & 0 & 4 \\ \frac{1}{3} & 0 & 0 & 0 \\ -\frac{2}{3} & -\frac{4}{7} & 0 & 9 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad \widetilde{\Omega}(\pi_2) = \begin{pmatrix} \frac{2}{3} & 0 & 0 & 0 \\ 0 & \frac{12}{7} & 0 & 0 \\ 0 & 0 & 7 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix},
$$

where π_2 is the permutation induced by this ordering. This DAG is depicted on the right in Figure 1. The fact that the incoming edges for X_4 are the same in both cases is consistent with the observation that under both orderings, we are projecting X_4 onto $\{X_1, X_2, X_3\}$.

Both (2.3) and (2.4) lead to a linear model as in (1.2) with $\operatorname{cov}(X) = \Sigma$, but with distinct numbers of edges. Two other key points from this example are worth emphasizing:

- Both DAGs $\widetilde{B}(\pi_j)$ $(j = 1, 2)$ are sparse, as measured by the number of nonzero entries in $\widetilde{B}(\pi_j)$.
- The conditional variances given by $\widetilde{\Omega}(\pi_j)$ $(j = 1, 2)$ are different in general. Moreover, in this example there is no *equivariance* DAG; that is, one whose conditional variance matrix $\widetilde{\Omega}$ is a constant multiple of the identity (Section 5.2).

These observations motivate many of the developments in the sequel and lead us to make the following definition:

Definition 2.1. Let $\mathbb{R}_+^p$ denote the space of $p \times p$ diagonal matrices with positive entries on the diagonal. Given a covariance matrix $\Sigma \succ 0$, define the *equivalence class* of Σ to be

$$\mathfrak{D}(\Sigma) = \left\{ \widetilde{B} \in \mathbb{D}_p : \Sigma = \Sigma(\widetilde{B}, \widetilde{\Omega}) \text{ for some } \widetilde{\Omega} \in \mathbb{R}_+^p \right\},$$

where $\mathbb{D}_p \subset \mathbb{R}^{p \times p}$ is the space of DAGs and $\Sigma(\widetilde{B}, \widetilde{\Omega})$ is defined by (2.1). The DAGs in $\mathfrak{D}(\Sigma)$ will be called *equivalent*.

Equivalently, $\widetilde{B}$ is the set of DAGs that satisfy (1.2) for some $\widetilde{\Omega} \in \mathbb{R}_+^p$. For the reader familiar with graphical models, note that $\mathfrak{D}(\Sigma)$ is *not* the same as the Markov equivalence class, and in fact is quite different.

We will now show that $\mathfrak{D}(\Sigma)$ can be constructed explicitly. Let $\mathbb{S}_p$ denote the class of permutations on p elements and $\pi \in \mathbb{S}_p$ be a fixed permutation. Write $\Gamma := \Sigma^{-1}$ and use the (modified) Cholesky decomposition to write $P_\pi \Gamma$ uniquely as $P_\pi \Gamma = (I - L)D^{-1}(I - L)^T$ where L is strictly lower triangular and $D \in \mathbb{R}_+^p$. Define

$$(2.5) \qquad \widetilde{B}(\pi) := P_{\pi^{-1}} L, \quad \widetilde{\Omega}(\pi) := P_{\pi^{-1}} D.$$

Since an adjacency matrix represents a DAG if and only if it is permutation-similar to a strictly lower triangular matrix, $\widetilde{B}(\pi)$ is indeed a DAG. We refer to $\widetilde{B}(\pi)$ as the DAG associated with π.

Lemma 2.1. *For any $\Sigma \succ 0$, the equivalence class of Σ is given by $\mathfrak{D}(\Sigma) = \{\widetilde{B}(\pi) : \pi \in \mathbb{S}_p\}$.*

This characterization of the equivalence class confirms that $\mathfrak{D}(\Sigma)$ depends only on Σ and not on the sample $\mathbf{X}$, and will play an important role in the rest of the paper. Therefore, without further qualification, we shall always write an arbitrary element of $\mathfrak{D}(\Sigma)$ as $\widetilde{B}(\pi)$. The columns of $\widetilde{B}(\pi)$ will be denoted by $\widetilde{\beta}_j(\pi)$, and the jth diagonal element of $\widetilde{\Omega}(\pi)$ will be denoted by $\widetilde{\omega}_j^2(\pi)$. It follows from these definitions and (1.2) that

$$(2.6) \qquad X_j = \widetilde{\beta}_j(\pi)^T X + \widetilde{\varepsilon}_j(\pi), \quad \text{where} \quad \widetilde{\varepsilon}_j(\pi) \sim \mathcal{N}(0, \widetilde{\omega}_j^2(\pi)),$$

for $j = 1, \ldots, p$. As a consequence, the joint distribution factors according to the local conditional distributions implied by (2.6), which makes $\widetilde{B}(\pi)$ a *Bayesian network* for each $\pi \in \mathbb{S}_p$. Furthermore, each $\widetilde{B}(\pi)$ is a so-called minimal I-map for $\mathcal{N}_p(0, \Sigma)$ (Koller and Friedman, 2009).

2.2. **Global and restricted minimizers.** In presenting our results, we will take advantage of the permutation decomposition of the equivalence class $\mathfrak{D}(\Sigma)$, according to Lemma 2.1. In this section, we formalize these ideas further and provide some intuition behind the estimator (1.1). In particular, these ideas are purely theoretical and are not intended to provide practical guidance for computing $\widehat{B}$ (for more on computation, see Section 7.2).

Recall that $\mathbb{D}_p$ is the space of $p \times p$ real matrices that represent DAGs when interpreted as weighted adjacency matrices. For each permutation $\pi \in \mathbb{S}_p$, define a subset of $\mathbb{D}_p$ by

$$\mathbb{D}_p[\pi] = \{B \in \mathbb{D}_p : P_\pi B \text{ is lower triangular}\}.$$

A DAG $B = [\, \beta_1 \mid \cdots \mid \beta_p \,] \in \mathbb{D}_p$ is in $\mathbb{D}_p[\pi]$ if and only if $\mathrm{supp}(\beta_j) \subset S_j(\pi)$ for all $j = 1, \ldots, p$, where

$$(2.7) \qquad S_j(\pi) := \{k : \pi^{-1}(k) > \pi^{-1}(j)\}$$

consists of the nodes X_k that come after X_j under the ordering $X_{\pi(i)} \prec X_{\pi(i+1)}$ for $i = 1, \ldots, p - 1$. In other words, for each node X_j, the permutation π defines a unique set of candidate parents given by (2.7), and $B \in \mathbb{D}_p[\pi]$ if and only if the parent set of β_j comes from $S_j(\pi)$ for all j. By definition, $\widetilde{B}(\pi) \in \mathbb{D}_p[\pi]$ for every π and hence $\mathrm{supp}(\widetilde{\beta}_j(\pi)) \subset S_j(\pi)$ for all j.

Recall the estimator $\widehat{B}$ defined via (1.1) and (1.3). Using the decomposition of the feasible region in (1.1) as $\mathbb{D}_p = \bigcup_\pi \mathbb{D}_p[\pi]$, we can build some intuition about $\widehat{B}$: Suppose we had oracle knowledge of some *optimal* permutation π_0 in advance. For example, π_0 might be a permutation that minimizes the number of edges among all the equivalent DAGs. Then we can ignore any DAG that is not consistent with π_0, and hence restrict our search space to $\mathbb{D}_p[\pi_0]$. Thus, estimating the *oracle DAG* $\widetilde{B}(\pi_0)$ reduces to minimizing $Q(B)$ with the additional constraint that $B \in \mathbb{D}_p[\pi_0]$, which is exactly equivalent to estimating an autoregressive model. Of course, in practice we do not know π_0, so in addition to parameter estimation, an additional challenge that arises in DAG learning is adaptively finding an optimal permutation π_0, which is a challenging combinatorial problem.

These observations motivate the following definition:

Definition 2.2. For $\pi \in \mathbb{S}_p$, a *restricted minimizer* of Q is

$$(2.8) \qquad \widehat{B}(\pi) \in \underset{B \in \mathbb{D}_p[\pi]}{\arg\min} \, Q(B).$$

The columns of $\widehat{B}(\pi)$ will be denoted by $\widehat{\beta}_j(\pi)$, $j = 1, \ldots, p$.

Our strategy will be to show that $\widehat{B}(\pi) \approx \widetilde{B}(\pi)$ *for all* π simultaneously. In Section 5.1, we will exploit this decomposition into restricted minimizers to establish guarantees on the global minimizer $\widehat{B}$.

3. NEIGHBOURHOOD REGRESSION

The core of our analysis is the regression decomposition (2.6) which we will interpret as a neighbourhood regression problem. In the literature on undirected graphical models, "the" neighbourhood of a fixed node X_j is implicitly understood to be the set of all other variables, namely the collection X_{-j} (Meinshausen and Bühlmann, 2006). For DAGs, we need to generalize the notion of neighbourhood to an arbitrary subset $S \subset [p]_j$. At a high-level, the *neighbourhood regression problem* for the node X_j and a subset $S \subset [p]_j$ is the problem of estimating the linear projection of X_j onto X_S. These problems arise naturally from the iterative projection procedure described in Section 2.1 and made explicit by (2.6). Doing this for all possible permutations is tantamount to projecting X_j onto all possible subsets $S \subset [p]_j$. In this section we formalize these notions and introduce the concept of a *model selection exponent*, which quantifies the difficulty of a neighbourhood regression problem.

3.1. **Penalized least-squares estimators.** The following defines the population-level quantity that is our primary interest:

Definition 3.1. For any $S \subset [p]_j$, let

$$\beta_j(S) := \underset{\beta \in \mathbb{R}^p,\ \mathrm{supp}(\beta) \subset S}{\arg\min} \mathbb{E}\big[X_j - \beta^T X\big]^2.$$

We call $\beta_j(S)$ the SEM coefficients for variable j regressed on the variables S.

The SEM coefficients $\beta_j(S)$ are *population* level quantities that depend on Σ, but not on the sample $\mathbf{X}$. Furthermore, it is easy to verify that $\widetilde{\beta}_j(\pi) = \beta_j(S_j(\pi))$ from the definition of $\widetilde{\beta}_j(\pi)$ as the jth column of $\widetilde{B}(\pi)$.

Now we turn to the problem of estimating $\beta_j(S)$ via PLS. For now, ρ_λ is assumed to be a fixed, possibly nonconvex, regularizer. We start with some abstraction in order to emphasize the key underlying assumptions.

Definition 3.2. Suppose $y \in \mathbb{R}^n$ and $Z \in \mathbb{R}^{n \times m}$. Let $S \subset [m]$ and consider the set defined by

$$(3.1) \qquad \widehat{\Theta}_\lambda(y, Z; S) := \underset{\theta \in \mathbb{R}^m,\ \mathrm{supp}(\theta) \subset S}{\arg\min} \frac{1}{2n}\|y - Z\theta\|_2^2 + \rho_\lambda(\theta),$$

i.e., the set of global minimizers of the support-restricted PLS problem above. Let $\widehat{\Theta}_\lambda(y, Z) := \widehat{\Theta}_\lambda(y, Z; [m])$ correspond to the case where there is no support restriction.

In this abstract definition, y is considered a fixed quantity and may have no relation to Z. Of course, in practice we are interested the case where y and Z are linked via a linear model: If $\mathbf{y} = Z\theta^* + \mathbf{w}$, where $\mathbf{y} \in \mathbb{R}^n$, $Z \in \mathbb{R}^{n \times m}$, $\theta^* \in \mathbb{R}^m$ and $\mathbf{w} \sim \mathcal{N}_n(0, \sigma^2 I_n)$, then $\widehat{\Theta}_\lambda(\mathbf{y}, Z)$ is the collection of PLS estimators for θ^* in classical linear regression. The support-restricted version $\widehat{\Theta}_\lambda(\mathbf{y}, Z; S)$ then allows us to properly define a neighbourhood regression problem. Let $\mathbf{x}_j$ denote the jth column of $\mathbf{X}$.

Definition 3.3 (Neighbourhood regression). The *neighbourhood regression problem* for node X_j given a neighbourhood $S \subset [p]_j$ is defined to be the (possibly nonconvex) program solved by $\widehat{\Theta}_\lambda(\mathbf{x}_j, \mathbf{X}; S)$. An arbitrary solution to this program will be denoted by $\widehat{\beta}_j(S)$, i.e. $\widehat{\beta}_j(S) \in \widehat{\Theta}_\lambda(\mathbf{x}_j, \mathbf{X}; S)$.

For any π such that $S_j(\pi) = S$, a solution $\widehat{\beta}_j(S)$ estimates $\widetilde{\beta}_j(\pi) = \beta_j(S)$. Thus neighbourhood regression problems are the basic units in learning the parent set of a node and hence the DAG structure. This concept allows us to do two things:

(1) A key step in our proof is to show that the analysis of $\widehat{B}$ can be reduced to the study of a (very large) collection of neighbourhood regression problems;

(2) We will associate to each neighbourhood regression problem an "exponent", which is a fundamental quantity measuring how difficult the corresponding model selection problem is for a given regularizer. These exponents will then be used to write down explicit, non-asymptotic upper bounds on the model selection failure of $\widehat{B}$ in terms of the model selection failure of each neighbourhood regression.

The details of these reductions can be found in Section 8.

3.2. Model selection exponents.
Given some $n \times m$ matrix Z and m-vector θ^*, define a set of "bad" noise vectors as follows:

$$(3.2) \qquad A(Z, \theta^*; S) := \Big\{ w \in \mathbb{R}^n : \operatorname{supp}(\widehat{\theta}) \neq \operatorname{supp}(\theta^*)$$

$$\exists \, \widehat{\theta} \in \widehat{\Theta}_\lambda(Z\theta^* + w, Z; S) \Big\}.$$

For a random vector $\mathbf{w} \in \mathbb{R}^n$ (e.g. $\mathbf{w} \sim \mathcal{N}_n(0, \sigma^2 I_n)$), we then have the following model selection failure event:

$$(3.3) \qquad \mathcal{A}(\mathbf{w}, Z, \theta^*; S) := \Big\{ \mathbf{w} \in A(Z, \theta^*; S) \Big\}.$$

As usual we use the shorthand $\mathcal{A}(\mathbf{w}, Z, \theta^*) = \mathcal{A}(\mathbf{w}, Z, \theta^*; [m])$.

Definition 3.4. Given a regularizer ρ_λ, the *model selection exponent* for the regression problem $\mathbf{y} = Z\theta^* + \mathbf{w}$ is defined to be

$$\Phi_\lambda(Z, \theta^*, \sigma^2) := -\log \mathbb{P}\left[\mathcal{A}(\mathbf{w}, Z, \theta^*) \right],$$

where $\mathbb{P}$ is taken with respect to the distribution of $\mathbf{w} \sim \mathcal{N}_n(0, \sigma^2 I_n)$.

A larger exponent corresponds to better model selection performance. To characterize the exponent further, let us introduce two quantities that are familiar from the regression literature: The sparsity level and the signal strength. The difference now is that instead of quantifying these for a single, "true" parameter, we need to consider all $p \cdot 2^{p-1}$ SEM coefficients. Recall that we denote the jth column of $\widetilde{B}(\pi)$ by $\widetilde{\beta}_j(\pi)$. For any $u \in \mathbb{R}^p$, let $\tau_*(u) := \min\{|u_j| : u_j \neq 0\}$.

Definition 3.5. For any Σ, let

$$(3.4) \qquad d(\Sigma) := \sup_{\pi, j} \|\widetilde{\beta}_j(\pi)\|_0, \quad \tau_*(\Sigma) := \inf_{\pi, j} \tau_*(\widetilde{\beta}_j(\pi)).$$

In Definition 3.5, we could have equivalently used the SEM coefficients directly: $d(\Sigma) = \sup\{\|\beta_j(S)\|_0 : S \subset [p]_j, \ j \in [p]\}$ and similarly for $\tau_*(\Sigma)$.

Let $\sigma_{\max}^2 := \max_{1 \leq j \leq p} \operatorname{var}(X_j)$ and note that $\sigma_{\max}^2 \leq r_{\max}(\Sigma)$. Each of the quantities d, τ_*, and $\sigma_{\max}^2$ are allowed to depend on n, and in particular d may diverge and τ_* may vanish as $n \to \infty$. Now, let

$$(3.5) \qquad \Psi_\lambda = \Psi_\lambda(\mathbf{X}, \Sigma) := \inf_{\substack{0 < \sigma \leq \sigma_{\max}}} \inf_{\substack{\|\theta\|_0 \leq d(\Sigma) \\ \tau_*(\theta) \geq \tau_*(\Sigma)}} \Phi_\lambda(\mathbf{X}, \theta, \sigma^2),$$

which encodes what is usually proved in the regression literature: An upper bound on the probability of model selection failure given the maximum sparsity level $d(\Sigma)$, minimum signal strength $\tau_*(\Sigma)$, and the maximum variance $\sigma_{\max}^2$. This probability generally depends on λ, the regularization parameter, which in turn may depend on any of these quantities. We will provide an example of such a bound under the MCP in Section 4.2.

4. UNIFORM CONTROL

Our main results are divided into two separate theorems: (i) a bound on the probability of false selection, and (ii) an upper bound on the estimation error. The key idea is to show that uniformly for all permutations $\pi \in \mathbb{S}_p$, we have $\operatorname{supp}(\widehat{B}(\pi)) = \operatorname{supp}(\widetilde{B}(\pi))$ and control over the deviations $\|\widehat{B}(\pi) - \widetilde{B}(\pi)\|_r$ for $r = 1, 2$.

4.1. **Assumptions.** Let us collect the assumptions needed for our main results. The following is our only assumption on Σ, and simply ensures that $\mathfrak{D}(\Sigma)$ is well-defined (cf. Definition 2.1):

Condition 4.1. Σ is positive definite, i.e., $r_{\min}(\Sigma) > 0$.

In addition, we require some regularity conditions on the regularizer ρ_λ. We consider *coordinate-separable* regularizers, that is with some abuse of notation we have $\rho_\lambda(B) = \sum_{i,j} \rho_\lambda(|\beta_{ij}|)$, where $\rho_\lambda : [0, \infty) \to [0, \infty)$ is a univariate regularizer. A key feature of our analysis is to provide insights into how ρ_λ affects the solution to (1.1), so we will not specify a particular regularizer in advance. Instead, we make the following minimal assumptions on ρ_λ:

Condition 4.2. ρ_λ is concave, nondecreasing, right-differentiable at zero, and satisfies the following conditions:

(a) $\rho_\lambda(0) = 0$;
(b) $0 < \rho_\lambda'(0+) < \infty$;
(c) There are constants $\underline{\rho}_0, \underline{\rho}_1 \geq 0$, independent of λ, such that $\rho_\lambda(x) \geq \min\{\underline{\rho}_1 \lambda x, \underline{\rho}_0 \lambda^2\}$.

We are also interested in penalties that "approximate" the ℓ_0 penalty:

Definition 4.1. ρ_λ is called ℓ_0-compatible if there is a constant $\overline{\rho}_0 \geq 0$, independent of λ, such that $\rho_\lambda(x) \leq \overline{\rho}_0 \lambda^2$ for all $x \geq 0$.

An elementary consequence of Condition 4.2 is that ρ_λ is subadditive. Condition 4.2(c) says that ρ_λ can be bounded below by a capped-ℓ_1 penalty: It is always true that a concave, nondecreasing function can be bounded below by a capped-ℓ_1 penalty, and Condition 4.2(c) simply normalizes this capped-ℓ_1 penalty in terms of λ. These conditions allow for most concave penalties, including the MCP and SCAD, along with the familiar ℓ_1 penalty. When ρ_λ is ℓ_0-compatible, we have $\rho_\lambda(B) \leq \overline{\rho}_0 \lambda^2 \|B\|_0$ for any $B \in \mathbb{D}_p$.

The following is a list of some admissible regularizers under our conditions:

– The minimax concave penalty (MCP) proposed by Zhang (2010):

$$(4.1) \qquad \rho_\lambda(x; \gamma) := \lambda\Big(x - \frac{x^2}{2\lambda\gamma}\Big)1(x < \lambda\gamma) + \frac{\lambda^2\gamma}{2}1(x \geq \lambda\gamma).$$

This satisfies Condition 4.2 with $\rho_\lambda'(0+) = \lambda$, $\underline{\rho}_1 = 1/2$, and $\underline{\rho}_0 = \gamma/2$. Furthermore, it is ℓ_0-compatible with $\overline{\rho}_0 = \gamma/2$.

– The ℓ_1 penalty, $\rho_\lambda(x) = \lambda x$, also satisfies Condition 4.2 with $\rho_\lambda'(0+) = \lambda$, $\underline{\rho}_1 = 1$, and $\underline{\rho}_0 \in [0, \infty)$. The ℓ_1 penalty is not, however, ℓ_0-compatible.

– The ℓ_0 penalty, $\rho_\lambda(x) = (\lambda^2/2)1(x \neq 0)$, fails to satisfy Condition 4.2(b) since it is not right-differentiable at zero (although see Remark 4.1 and Section 7.1). Nonetheless, it satisfies all of the other assumptions, including Condition 4.2(c) with $\underline{\rho}_1 \in [0, \infty)$ and $\underline{\rho}_0 = 1/2$. Of course, the ℓ_0 penalty is trivially ℓ_0-compatible with $\overline{\rho}_0 = 1/2$.

Remark 4.1. Even though the ℓ_0 penalty does not satisfy Condition 4.2(b), all the results in this paper still apply to the ℓ_0 penalty under possibly different values of the constants involved. This requires a small modification to the overall argument, which is discussed in Section 7.1.

4.2. **Support recovery.** Our first result provides a guarantee for model selection consistency:

Theorem 4.1. *Suppose* $\mathbf{X} \overset{iid}{\sim} \mathcal{N}_p(0, \Sigma)$ *and assume Condition 4.1. Then*

$$\mathbb{P}\left(\bigcap_{j=1}^{p} \bigcap_{\pi \in \mathbb{S}_p} \big\{ \operatorname{supp}(\widehat{\beta}_j(\pi)) = \operatorname{supp}(\widetilde{\beta}_j(\pi)) \big\} \right)$$

$$= \mathbb{P}\left(\bigcap_{\pi \in \mathbb{S}_p} \big\{ \operatorname{supp}(\widehat{B}(\pi)) = \operatorname{supp}(\widetilde{B}(\pi)) \big\} \right)$$

$$\geq 1 - p\binom{p}{d} \mathbb{E} e^{-\Psi_\lambda(\mathbf{X}, \Sigma)}.$$

Theorem 4.1 explicitly relates the behaviour of the neighbourhood regression problems and the restricted minimizers to the model selection exponent for a random design regression problem. Since $p\binom{p}{d} \leq p^{d+1}$, it follows that

(4.2) $\Psi_\lambda(\mathbf{X}, \Sigma) \geq C(d+1)\log p, \quad$ for some $C > 1$

is sufficient for simultaneous, *uniform* model selection consistency for all $\widehat{B}(\pi)$ and $\widehat{\beta}_j(\pi)$. For example, a bound of the form (4.2) holds with high probability for the MCP as illustrated in Lemma 4.1.

Theorem 4.1 provides uniform control on the probability of false selection for all permutations $\pi \in \mathbb{S}_p$. Since there are $p!$ permutations, this is a fairly surprising result: A naïve union bound would have $p!$ in place of the $p\binom{p}{d}$ term. Even though existing bounds on false selection for linear regression decay exponentially as in (4.2), they are not strong enough to counteract $p!$. Theorem 4.1 combined with (4.2), on the other hand, leads to a similar exponential decay in controlling this *super*exponential collection of regression problems. The key is a novel monotonicity argument detailed in Section 8.1, more specifically Section 8.1.2.

To illustrate, let us give explicit conditions for (4.2) to hold in the case of MCP, defined in (4.1). Huang et al. (2012) consider PLS estimators $\widehat{\Theta}_\lambda(y, Z; S)$ as defined in (3.1), applied to the data from a linear regression model $y = Z\theta^* + \mathbf{w}$, and provides conditions for model selection consistency. Adapting their result to our setup and notation, we have the following bound on model selection exponent for the MCP:

Lemma 4.1. *Suppose* $\mathbf{X} \overset{iid}{\sim} \mathcal{N}_p(0, \Sigma)$. *Take* $\rho_\lambda = \rho_\lambda(\,\cdot\,; \gamma)$ *as in (4.1) and assume* Σ *is positive definite with bounded eigenvalues. Assume that*

(1) $d(\Sigma) \leq \kappa_4 \cdot \min\{p, n, n/\log p\}$,

(2) $\tau_*(\Sigma) > (1+\gamma)\lambda$ *for some* $\gamma > \kappa_5 > 0$.

Then for any $\lambda \geq \kappa_6 \cdot \sqrt{(d+1)\log p/n}$, *it follows that* $\mathbb{E}e^{-\Psi_\lambda(\mathbf{X}, \Sigma)} \leq 3\exp\left(-2\min\{d\log p, n\}\right)$. *Here,* $\kappa_j = \kappa_j(\Sigma)$ $(j = 4, 5, 6)$ *are constants depending only on* $\{r_{\min}(\Sigma), r_{\max}(\Sigma)\}$.

Taking λ as in Lemma 4.1, it follows from Theorem 4.1 that if $n \geq C(d + 1) \log p$ for a sufficiently large constant $C > 0$,

$$\mathbb{P}\left(\bigcap_{\pi \in \mathbb{S}_p} \left\{ \operatorname{supp}(\widehat{B}(\pi)) = \operatorname{supp}(\widetilde{B}(\pi)) \right\} \right) \to 1,$$

i.e., each restricted minimizer $\widehat{B}(\pi)$ is model selection consistent.

Remark 4.2. van de Geer and Bühlmann (2013) remark that a thresholded ℓ_0-penalized maximum likelihood estimator for (B, Ω) also attains exact support recovery with high probability if λ is of order $s_0 \log p/n$, where $s_0 := \arg\min_\pi \|\widetilde{B}(\pi)\|_0$. Requiring that $\lambda \to 0$, this result allows for p to grow with n, but is not truly high-dimensional: In general, unless the graph is extremely sparse (i.e. less than one parent per node), $s_0 > p$ in which case this scaling requires $p \ll n$. By contrast, our results only require λ on the order $d \log p/n$, which guarantees consistency when $p \gg n$ and allows s_0 as large as $O(pd)$ (i.e d parents per node).

4.3. **Deviation bounds.** In addition to support recovery, we can also bound the estimation errors for all columns $\|\widehat{\beta}_j(\pi) - \widetilde{\beta}_j(\pi)\|_r$ simultaneously:

Theorem 4.2. *Under Conditions 4.1 and 4.2, there are positive constants $\kappa_1 = \kappa_1(\Sigma)$ and $\kappa_2 = \kappa_2(\Sigma)$ and universal constants $c_3, c_4 > 0$ such that if*

$$n > \kappa_1 \, d \log p, \quad \lambda \geq \kappa_2 \sqrt{\frac{(d+1) \log p}{n}}$$

then for $r = 1, 2$, the following uniform deviation bounds hold:

$$(4.3) \qquad \mathbb{P}\left(\bigcap_{\pi \in \mathbb{S}_p} \left\{ \|\widehat{B}(\pi) - \widetilde{B}(\pi)\|_r \leq c_3 \frac{\rho'_\lambda(0+)}{r_{\min}(\Sigma)} \|\widetilde{B}(\pi)\|_0^{1/r} \right\} \right)$$

$$(4.4) \qquad \geq \mathbb{P}\left(\bigcap_{j=1}^p \bigcap_{\pi \in \mathbb{S}_p} \left\{ \|\widehat{\beta}_j(\pi) - \widetilde{\beta}_j(\pi)\|_r \leq c_4 \frac{\rho'_\lambda(0+)}{r_{\min}(\Sigma)} \|\widetilde{\beta}_j(\pi)\|_0^{1/r} \right\} \right)$$

$$\geq 1 - c_1 \exp(-c_2(d+1) \log p).$$

Thus, in addition to support recovery guarantees, we can bound the finite sample error in estimating the parameters $\widetilde{B}(\pi)$, uniformly over the entire equivalence class $\mathfrak{D}(\Sigma)$. Taking the MCP (4.1) as an example for which $\rho'_\lambda(0+) = \lambda$, we have an ℓ_2 error bound of the order $\lambda \|\widetilde{\beta}_j(\pi)\|_0^{1/2}$, uniformly for all π and j. Using standard interpolation bounds for ℓ_r norms, this result can easily be extended to bounds for all $1 \leq r \leq 2$. See, for example, the proof of Theorem 7.1 in Bickel et al. (2009).

4.4. **Discussion.** Let us pause to discuss Theorems 4.1 and 4.2. In order to obtain model selection consistency, it is sufficient to show that (4.2) holds: Of course, whether or not this holds for a particular regularizer depends on the regularity of the problem. For example, if the ℓ_1 penalty is used,

then it is well-known that the irrepresentable condition is required for a bound like (4.2) to hold (Zhao and Yu, 2006). By contrast, if nonconvex regularizers such as the MCP or SCAD are used, then this type of condition is not necessary (Loh and Wainwright, 2015; Huang et al., 2012; Zhang, 2010). Condition 2 on the signal strength in Lemma 4.1, also called a *beta-min* condition in the literature, is necessary in order to guarantee support recovery but is not necessary for the deviation bounds alone (Theorem 4.2). For more on this assumption, see the related discussion in Section 5.1.2.

Theorem 4.2 (along with Lemma 4.1) requires the regularization parameter to scale as $\lambda \sim \sqrt{d \log p / n}$, which is the familiar scaling from the literature up to a factor of $\sqrt{d}$. As our bounds are nonasymptotic, d may depend on n: For example, we may take $\lambda \to 0$ and $d \to \infty$ as long as $d = o(n / \log p)$. See also Remark 5.2. What's more, d can often be explicitly bounded (e.g. autoregressive models, see van de Geer and Bühlmann, 2013). The $\sqrt{d}$ term reflects the cost of requiring *uniform* control over all possible neighbourhood regression problems, which is a worthwhile trade-off as we shall see in the next section.

5. Applications

The results in Section 4 provide uniform control over the family of estimators $\{\widehat{B}(\pi) : \pi \in \mathbb{S}_p\}$. As stated, however, these theorems are fairly abstract. In this section we discuss three applications of these results to score-based learning, causal inference, and graphical models. Proofs are deferred to Section 8.

5.1. **Score-based learning.** As our first application, we use Theorems 4.1 and 4.2 to provide statistical guarantees for the score-based estimator $\widehat{B}$, defined by (1.1) and (1.3). Our results will come in three flavours: (i) consistency guarantees (Section 5.1.1), (ii) sparsity guarantees, (Section 5.1.2), and (iii) an oracle inequality (Section 5.1.3). The results in Section 5.1.1 will show that there is a DAG (denoted by $\widetilde{B}(\widehat{\pi})$) which $\widehat{B}$ is "close" to, and in Sections 5.1.2 and 5.1.3 we will carefully study the DAG $\widetilde{B}(\widehat{\pi})$ in order to show that the program (1.1) adaptively selects an equivalent DAG with appealing properties such as sparsity.

5.1.1. *Consistency guarantees.* So far, our results have been stated in terms of the *restricted* minimizers $\widehat{B}(\pi)$ of Definition 2.2. Obviously, $\widehat{B}$ is also a restricted minimizer for some permutation(s) π, although *a priori*, we do not know which one(s). The following definition formalizes the collection of such permutations:

Definition 5.1. The *collection of estimated permutations* is

$$\widehat{\mathbb{S}}_p = \{\pi \in \mathbb{S}_p : \widehat{B} \in \mathbb{D}_p[\pi]\}.$$

An arbitrary element of $\widehat{\mathbb{S}}_p$ will be denoted by $\widehat{\pi}$.

The following statements are equivalent: (i) $\pi \in \widehat{\mathbb{S}}_p$, (ii) $\widehat{B} \in \mathbb{D}_p[\pi]$, (iii) $P_\pi \widehat{B}$ is lower triangular, and (iv) $\widehat{B} = \widehat{B}(\pi)$. Since $\widehat{B}$ depends on the random matrix $\mathbf{X}$, the set of permutations $\widehat{\mathbb{S}}_p$ is also a random quantity. This randomness complicates any direct attempt at analyzing $\widehat{B}$, and is why we first established uniform control over all possible permutations in Section 4. Note that to solve (1.1), it is not necessary to compute $\widehat{\mathbb{S}}_p$—in fact, in practical applications, once a solution $\widehat{B}$ to (1.1) is found, $\widehat{\mathbb{S}}_p$ can be obtained as a byproduct by finding topological sorts of the DAG $\widehat{B}$.

Since a global minimizer is indeed a restricted minimizer, the following theorem is an immediate corollary of Theorems 4.1 and 4.2:

Theorem 5.1. *Take $\rho_\lambda = \rho_\lambda(\,\cdot\,;\gamma)$ as in (4.1) and assume Σ is positive definite with bounded eigenvalues. Suppose further that Conditions 4.1 and 4.2 hold. Then there exist positive constants $\kappa_j = \kappa_j(\Sigma)$ $(j = 1, 2, 3)$ such that if $n > \kappa_1 d \log p$ and $\lambda \geq \kappa_2 \sqrt{(d+1)\log p/n}$ then for $r = 1, 2$ and with probability at least $1 - c_1 \exp(-c_2(d+1)\log p)$,*

$$\|\widehat{B} - \widetilde{B}(\widehat{\pi})\|_r \lesssim \frac{\rho'_\lambda(0+)}{r_{\min}(\Sigma)} \|\widetilde{B}(\widehat{\pi})\|_0^{1/r} \quad \forall\, \widehat{\pi} \in \widehat{\mathbb{S}}_p, \text{ and}$$

$$\|\widehat{\beta}_j - \widetilde{\beta}_j(\widehat{\pi})\|_r \lesssim \frac{\rho'_\lambda(0+)}{r_{\min}(\Sigma)} \|\widetilde{\beta}_j(\widehat{\pi})\|_0^{1/r} \quad \forall\, 1 \leq j \leq p,\ \widehat{\pi} \in \widehat{\mathbb{S}}_p.$$

If also $\tau_*(\Sigma) > (1+\gamma)\lambda$ for some $\gamma > \kappa_3 > 0$, then with the same probability

$$\operatorname{supp}(\widehat{B}) = \operatorname{supp}(\widetilde{B}(\widehat{\pi})) \quad \forall\, \widehat{\pi} \in \widehat{\mathbb{S}}_p.$$

This theorem provides preliminary justification for the use of score-based estimators on high-dimensional datasets, providing practitioners with finite-sample guarantees for structure learning and parameter estimation. In addition to an upper bound on the overall deviation $\|\widehat{B} - \widetilde{B}(\widehat{\pi})\|_r$, Theorem 4.2 provides control over the column deviations $\|\widehat{\beta}_j - \widetilde{\beta}_j(\widehat{\pi})\|_r$, which highlights another advantage of our uniform control results. Furthermore, although our result is stated for the MCP, similar results are available for other regularizers using, e.g. Loh and Wainwright (2014).

5.1.2. *Sparsity.* Theorem 5.1 has a drawback: Although it guarantees that *there exists* a DAG $\widetilde{B}(\widehat{\pi}) \in \mathfrak{D}(\Sigma)$ which our estimator $\widehat{B}$ is "close" to, nothing is said about *which* member of the equivalence class this is. Due to the inclusion of the regularizer ρ_λ in the score function, one hopes the estimated permutation $\widehat{\pi}$ to be well-behaved, for example, in the sense that the corresponding population DAG $\widetilde{B}(\widehat{\pi})$ is sparse. Theorem 5.1, however, is silent about the properties of $\widetilde{B}(\widehat{\pi})$. In this section, we complete this picture by showing that $\widetilde{B}(\widehat{\pi})$ is sparse. In the next section, we will strengthen this result by proving an oracle-type inequality for $\widetilde{B}(\widehat{\pi})$ that holds under weaker assumptions than those in the current section.

We now show that the number of edges in $\widetilde{B}(\widehat{\pi})$ can be controlled under some assumptions on the *signal strength* and a *minimum-trace DAG* (Definition 5.2). Since $\widetilde{B}(\widehat{\pi})$ is approximated by $\widehat{B}$ according to Theorem 5.1, this implies sparsity of $\widehat{B}$ as well.

Condition 5.1 (Signal strength). The minimum signal τ_* satisfies

$$\rho_\lambda(\tau_*) \geq a_1 \frac{\rho'_\lambda(0+)^2}{r_{\min}(\Sigma)} \quad \text{for some } a_1 > 2.$$

If we assume that the eigenvalues of Σ are bounded, then for penalties satisfying Condition 4.2 and for which $\rho'_\lambda(0+) = O(\lambda)$ (e.g., both the MCP and ℓ_1), Condition 5.1 will hold if $\tau_* \gtrsim \lambda \gtrsim \sqrt{(d+1)\log p/n}$, which is the same scaling required by Lemma 4.1.

Our final condition involves a so-called *minimum-trace DAG*:

Definition 5.2. A *minimum-trace permutation* is any

$$\pi_0 \in \arg\min_{\pi \in \mathbb{S}_p} \operatorname{tr} \widetilde{\Omega}(\pi).$$

The corresponding DAG $\widetilde{B}(\pi_0)$ is called a *minimum-trace DAG*.

Minimum-trace DAGs are discussed in more detail in Section 5.2.

Condition 5.2 (Minimum-trace). There exists a minimum-trace permutation π_0 such that

$$\frac{\rho_\lambda(\widetilde{B}(\pi_0))}{\operatorname{tr} \widetilde{\Omega}(\pi_0)} \geq a_2 \sqrt{\frac{(d+1)\log p}{n}} \quad \text{for some } a_2 > 0.$$

In light of the lower bound on λ required by Theorem 4.2, Condition 5.2 can be rewritten as $\rho_\lambda(\widetilde{B}(\pi_0)) \gtrsim (\lambda/\kappa_2) \cdot \operatorname{tr} \widetilde{\Omega}(\pi_0)$ which puts a lower bound on the penalty, as measured by the minimum-trace DAG $\widetilde{B}(\pi_0)$.

The following result shows how the sparsity of $\widehat{B}$ is essentially sandwiched between that of $\widetilde{B}(\widehat{\pi})$ and $\widetilde{B}(\pi_0)$:

Theorem 5.2. *Assume the conditions of Theorem 4.2 hold along with Conditions 5.1 and 5.2. Then*

$$(5.1) \qquad \rho_\lambda(\widetilde{B}(\widehat{\pi})) \lesssim \rho_\lambda(\widehat{B}) \lesssim \rho_\lambda(\widetilde{B}(\pi_0))$$

with probability at least $1-c_1\exp(-c_2(d+1)\log p)$. *If ρ_λ is also ℓ_0-compatible, then with the same probability,*

$$(5.2) \qquad \|\widetilde{B}(\widehat{\pi})\|_0 \lesssim r_{\min}(\Sigma) \frac{\lambda^2}{\rho'_\lambda(0+)^2} \|\widetilde{B}(\pi_0)\|_0.$$

The constants in (5.1) *depend only on a_1 and a_2 in Conditions 5.1 and 5.2, while the constant in* (5.2) *also depends on $\overline{\rho}_0$ (from Definition 4.1).*

Combining Theorems 5.1 and 5.2, it follows that $\widehat{B}$ estimates a sparse DAG model for the distribution $\mathcal{N}_p(0,\Sigma)$. This is arguably the best one can

hope without assuming there is an identifiable DAG for the joint Gaussian distribution.

Remark 5.1. Theorem 5.2 applies to regularizers that are not ℓ_0-compatible, such as the ℓ_1 penalty. For example, we can achieve $\|\widetilde{B}(\widehat{\pi})\|_1 \lesssim \|\widehat{B}\|_1 \lesssim \|\widetilde{B}(\pi_0)\|_1$ with the ℓ_1 penalty. The idea is that the penalty ρ_λ itself may be interpreted as a "measure of sparsity" that is weaker than the ℓ_0 norm. Moreover, although the ℓ_0 penalty does not satisfy Condition 4.2, which is required in Theorem 4.2, (5.1) still applies to the ℓ_0 penalty (see Section 7.1). Compared to Theorem 3.1 in van de Geer and Bühlmann (2013), due to our use of the least-squares loss instead of the log-likelihood (see Section 7.2), our results do not guarantee that $\|\widetilde{B}(\widehat{\pi})\|_0$ is on the same order as the number of edges in the sparsest DAG. Furthermore, Theorem 5.2 stops short of providing a lower bound on $\|\widetilde{B}(\widehat{\pi})\|_0$, which allows for slightly stronger bounds compared to (5.1). If we assume that $\widetilde{B}(\pi_0)$ is also one of the sparsest DAGs, however, then the number of edges in $\widehat{B}$ is on the same order as the sparsest DAGs in $\mathfrak{D}(\Sigma)$.

Remark 5.2. Combining Theorems 5.1 and 5.2, with high probability,

$$(5.3) \qquad \|\widehat{B} - \widetilde{B}(\widehat{\pi})\|_2^2 \lesssim \frac{\lambda^2}{r_{\min}(\Sigma)} s_0,$$

where $s_0 = \|\widetilde{B}(\pi_0)\|_0$ is the number of edges in a minimum-trace DAG. Let us investigate under which scalings of (n, p, d, s_0) the ℓ_2 error vanishes asymptotically. Choose λ to be on the order of $\sqrt{d \log p / n}$ and assume that Σ has bounded eigenvalues for simplicity. Since the squared ℓ_2 error is essentially the sum over p regression problems, we normalize it by $1/p$. Thus, to achieve consistency in the normalized ℓ_2 error, it is sufficient to have $(s_0/p) d \log p \ll n$, where (s_0/p) is the average parent size of the minimum-trace DAG $\widetilde{B}(\pi_0)$. If $d \log p \leq \epsilon n$ with $\epsilon = o(1)$, then we can have $s_0/p \to \infty$ as long as $s_0/p \ll 1/\epsilon$, which imposes a quite weak sparsity assumption on $\widetilde{B}(\pi_0)$. Under this asymptotic scaling, Condition 5.1 allows $\tau_* \to 0$ for the MCP and ℓ_1, and Condition 5.2 allows $\rho_\lambda(\widetilde{B}(\pi_0))/\operatorname{tr} \widetilde{\Omega}(\pi_0) \to 0$. Furthermore, it is possible to allow $p \gg n$, which justifies our results for high-dimensional data. To establish normalized ℓ_2-consistency, van de Geer and Bühlmann (2013) assume that ϵ is sufficiently small but does not necessarily vanish asymptotically (their Condition 3.4). The fact that we need a slightly stronger assumption on d to obtain ℓ_2-consistency is probably the price we pay for obtaining uniform error control over all $\pi \in \mathbb{S}_p$.

Remark 5.3. For the degenerate case $\widetilde{B}(\pi_0) = 0$, i.e. a minimum-trace DAG is the empty graph, Condition 5.2 cannot hold. It is easy to check, however, that in this case it follows that $\widetilde{B}(\pi) = 0$ for all $\pi \in \mathbb{S}_p$, and hence Theorem 4.1 implies both $\rho_\lambda(\widetilde{B}(\widehat{\pi})) = \rho_\lambda(\widehat{B}) = \rho_\lambda(\widetilde{B}(\pi_0)) = 0$ and $\|\widetilde{B}(\widehat{\pi})\|_0 = \|\widehat{B}\|_0 = \|\widetilde{B}(\pi_0)\|_0 = 0$.

5.1.3. *An oracle inequality.* In order to control the sparsity of $\widehat{B}$ and $\widetilde{B}(\widehat{\pi})$, Theorem 5.2 requires a condition on a minimum-trace permutation π_0 (Condition 5.2). Here we provide a different guarantee on $\widetilde{B}(\widehat{\pi})$, in the form of an oracle inequality, which is true without any extra assumptions.

Observe that $\frac{1}{2} \operatorname{tr} \widetilde{\Omega}(\pi)$ is the same as the expected loss, i.e.

$$(5.4) \qquad \mathbb{E}\Big[\frac{1}{2n}\|\mathbf{X} - \mathbf{X}\widetilde{B}(\pi)\|_2^2\Big] = \frac{1}{2}\operatorname{tr}\widetilde{\Omega}(\pi) \quad \text{for any } \widetilde{B}(\pi) \in \mathfrak{D}(\Sigma).$$

This gives us an expression for the expected penalized loss, which we denote

$$\widetilde{Q}_\lambda(\pi) := \frac{1}{2}\operatorname{tr}\widetilde{\Omega}(\pi) + \rho_\lambda(\widetilde{B}(\pi)).$$

If we let $\pi^* \in \arg\min_\pi \widetilde{Q}_\lambda(\pi)$, it follows from Lagrange duality that the DAG $\widetilde{B}(\pi^*) \in \mathfrak{D}(\Sigma)$ minimizes the weak sparsity measure ρ_λ subject to some upper bound on the expected prediction error, i.e., $\operatorname{tr}\widetilde{\Omega}(\pi) \le c(\lambda)$. The next result shows that the expected penalized loss of an estimated permutation $\widehat{\pi}$ is (up to a vanishing term) on the same order as that of the population minimizer, thus giving another kind of control over the sparsity of $\widetilde{B}(\widehat{\pi})$.

Theorem 5.3. *Suppose Conditions 4.1, 4.2, and 5.1 hold. Then, there exist positive constants $\kappa_1 = \kappa_1(\Sigma)$ and $\kappa_2 = \kappa_2(\Sigma)$ such that if*

$$n > \kappa_1\, d \log p, \quad \lambda \ge \kappa_2 \sqrt{\frac{(d+1)\log p}{n}},$$

then

$$\widetilde{Q}_\lambda(\widehat{\pi}) \ \le\ \frac{3a_1 + 2}{a_1 - 2}\left(1 + 6\sqrt{\frac{(d+1)\log p}{n}}\right)\cdot \inf_\pi \widetilde{Q}_\lambda(\pi)$$

with probability at least $1 - c_1\exp(-c_2(d+1)\log p)$.

The proof of this inequality can be found in Appendix C.4. Note that the constant factor of $(3a_1 + 2)/(a_1 - 2)$ approaches 3 as the signal strength τ_* increases, and $\sqrt{(d+1)\log p/n} = o(1)$ as long as $n \gg (d+1)\log p$.

5.2. **Causal DAG learning.** So far we have avoided singling out any particular DAG for special treatment, instead opting to provide *uniform* control over the entire equivalence class $\mathfrak{D}(\Sigma)$. Of course, without additional information about Σ, the parameters $(\widetilde{B}, \widetilde{\Omega})$ are nonidentifiable. However, under additional assumptions it may be possible to uniquely identify a DAG $\widetilde{B} \in \mathfrak{D}(\Sigma)$. In this setting, one often refers to $\widetilde{B}$ as a *causal DAG*: It represents the unique (directed) graphical structure that describes the relationship between the random variables $(X_1, \ldots, X_p)$ via the SEM (1.2), and thus can be interpreted as a "causal" explanation of the underlying joint distribution as long as there are no unmeasured confounders (among other assumptions, see Spirtes et al., 2000, for details). In other words, one can regard $\widetilde{B}$ as the true DAG for the observed data. It is then an interesting

problem to study under what assumptions $\widehat{B}$ recovers $\widetilde{B}$. Note that $\widetilde{B}$ may be consistent with more than one permutation.

In the previous section, we defined the concept of a minimum-trace DAG (Definition 5.2), which by definition minimizes the sum of the error variances $\sum_j \widetilde{\omega}_j^2(\pi)$ over all π. Thus, minimum-trace DAGs minimize the total squared prediction error among all DAG parameterizations of the underlying joint Gaussian distribution. While minimum-trace DAGs are not unique in general, when they are we can guarantee their recovery via $\widehat{B}$.

Given $\widetilde{B} \in \mathfrak{D}(\Sigma)$, define $\widetilde{\Omega}$ to be such that $\Sigma(\widetilde{B}, \widetilde{\Omega}) = \Sigma$ (cf. (2.1)).

Theorem 5.4. *Suppose that the minimum-trace DAG $\widetilde{B}$ is unique and there exists $a_3 > 10$ such that*

$$\frac{\operatorname{tr}\widetilde{\Omega}}{\operatorname{tr}\widetilde{\Omega}(\pi)} \leq 1 - a_3 \sqrt{\frac{(d+1)\log p}{n}} \quad \forall\, \operatorname{tr}\widetilde{\Omega}(\pi) \neq \operatorname{tr}\widetilde{\Omega}.$$

Then there exist constants $\kappa_j = \kappa_j(\Sigma)$ such that if

(1) ρ_λ is ℓ_0-compatible,

(2) $n > \kappa_1 (d+1)^3 \log p$,

(3) $\lambda \geq \kappa_2 \sqrt{\frac{(d+1)\log p}{n}}$,

(4) $\tau_*(\Sigma) > (1+\gamma)\lambda$ for some $\gamma > \kappa_3 > 0$,

then with probability at least $1 - c_1 \exp(-c_2(d+1)\log p)$, it holds that

$$\operatorname{supp}(\widehat{B}) = \operatorname{supp}(\widetilde{B}) \quad and \quad \|\widehat{B} - \widetilde{B}\|_r \lesssim \frac{\rho'_\lambda(0+)}{r_{\min}(\Sigma)}\|\widetilde{B}\|_0^{1/r}.$$

Furthermore, if $\widetilde{B}$ is not unique, then all of the above statements are true with $\widetilde{B}$ replaced by some minimum-trace DAG.

Although the previous theorem remains valid when $\widetilde{B}$ is not unique, it is still of interest to provide conditions under which minimum-trace DAGs are indeed unique. Suppose that we know the error variances $\widetilde{\Omega}$ for the true DAG. Then by proper rescaling of X_j, we can always make $\widetilde{\Omega} = \omega_0^2 I$, which makes $\widetilde{B}$ an *equivariance DAG*. As it turns out, an equivariance DAG is automatically the unique minimum-trace DAG and therefore by Theorem 5.4, $\widehat{B}$ recovers the true causal DAG $\widetilde{B}$ accurately.

Lemma 5.1. *Suppose Σ is given and (1.2) holds for some $\widetilde{B} \in \mathbb{D}_p$ with $\widetilde{\Omega} = \omega_0^2 I$. Then $\widetilde{B}$ is the unique minimum-trace DAG.*

Thus, a practical use case of Theorem 5.4 is as follows: If one can estimate the error variances from data beforehand or if one has reasons to believe that the true error variances are approximately identical, then score-based learning (1.1) can be used to estimate the true DAG.

As a related result, Peters and Bühlmann (2013) have shown that equivariance DAGs are identifiable. Note that a minimum-trace DAG always

exists, whereas an equivariance DAG may not (Example 2.1). Therefore, the minimum-trace property provides a natural generalization of the equivariance property to general covariance matrices that are not equivariant. As a consequence of Lemma 5.1, we immediately obtain the following corollary:

Corollary 5.2. *Theorem 5.4 remains true if "minimum-trace" is replaced with "equivariance".*

Thus, if we are interested in recovering an equivariance DAG $\widetilde{B}$ then Corollary 5.2 shows that $\mathrm{supp}(\widehat{B}) = \mathrm{supp}(\widetilde{B})$. To the best of our knowledge, this is the first result that guarantees consistent structure learning of a Gaussian DAG when $p \gg n$.

Remark 5.4. A related but weaker result first appeared in van de Geer and Bühlmann (2013): Deviation, sparsity bounds, and permutation recovery are proved for equivariance DAGs under a low-dimensional setting with $p \lesssim n/\log n$. Theorem 5.4 thus provides a strengthening of this result to include support recovery for arbitrary covariance matrices and general nonconvex penalties under the high-dimensional scaling $n \gtrsim (d+1)^3 \log p$.

5.3. **Conditional independence learning.** Our final application concerns the problem of learning CI relations for a p-variate Gaussian distribution $\mathcal{N}_p(0, \Sigma)$. Throughout this section we assume the reader is familiar with concepts from the graphical modeling literature such as d-separation, Markov equivalence, faithfulness, Markov perfectness, etc. In particular, note that the equivalence class $\mathfrak{D}(\Sigma)$ is *not* the same as the Markov equivalence class.

In graphical modeling, the goal is to represent a distribution via a graph in which separation can be used to read off CI relations. In this way, a DAG $G = (V = [p], E)$ defines—via the notion of d-separation—a set of pairwise CI relations, which we denote by $\mathcal{I}(G) \subset [p] \times [p] \times 2^{[p]_{ij}}$. A triplet $(i, j, S) \in \mathcal{I}(G)$ indicates that that X_i is d-separated from X_j by X_S in the graph G, and the Markov property for Bayesian networks implies that $X_i \perp\!\!\!\perp X_j \mid X_S$. Specialized to our setting, we use the shorthand $\mathcal{I}(\pi) = \mathcal{I}(\widetilde{B}(\pi))$ and $\widehat{\mathcal{I}}(\pi) = \mathcal{I}(\widehat{B}(\pi))$. Finally, for the joint Gaussian $\mathcal{N}_p(0, \Sigma)$, we denote by $\mathcal{I}(\Sigma)$ the set of all pairwise CI relations among $X_1, \ldots, X_p$. Recall that an *I-map* is any graph such that $\mathcal{I}(G) \subset \mathcal{I}(\Sigma)$, and a *minimal I-map* is an I-map such that removing any edge from the graph violates this inclusion.

In general, the sets $\mathcal{I}(\pi)$, $\widehat{\mathcal{I}}(\pi)$, and $\mathcal{I}(\Sigma)$ can all be different, although $\widetilde{B}(\pi)$ is always a minimal I-map for $\mathcal{N}_p(0, \Sigma)$, and hence in particular $\mathcal{I}(\pi) \subset \mathcal{I}(\Sigma)$. On the other hand, it is not hard to see that $\bigcup_{\pi \in \mathbb{S}_p} \mathcal{I}(\pi) = \mathcal{I}(\Sigma)$. If there exists a single permutation π such that $\mathcal{I}(\pi) = \mathcal{I}(\Sigma)$, then we call $\widetilde{B}(\pi)$ *faithful*—i.e. all pairwise CI relations of the underying Gaussian distribution can be read off from the graph $\widetilde{B}(\pi)$ via the d-separation criterion. Under the assumptions of Theorem 5.4, if the joint Gaussian distribution $\mathcal{N}_p(0, \Sigma)$ is faithful to the DAG $\widetilde{B}(\pi_0)$, then the graphical structure of $\widehat{B}$ contains all CI relations among $X_1, \ldots, X_p$, which is implied by the exact support

recovery. That is, $\mathcal{I}(\widehat{B}) = \mathcal{I}(\widehat{\pi}) = \mathcal{I}(\pi_0) = \mathcal{I}(\Sigma)$. Unfortunately, faithfulness is a strong assumption that rarely holds in practice (Lin et al., 2014; Uhler et al., 2013). We illustrate this with the following simple examples:

Example 5.1. Consider the covariance matrix $\Sigma = \Gamma^{-1}$ where

(5.5)
$$\Gamma = \begin{pmatrix} 10 & 1 & 0 & 2 \\ 1 & 10 & 3 & 0 \\ 0 & 3 & 10 & 4 \\ 2 & 0 & 4 & 10 \end{pmatrix}.$$

This covariance matrix implies two conditional independence relations:

$$X_1 \perp\!\!\!\perp X_3 \,|\, \{X_2, X_4\}, \quad X_2 \perp\!\!\!\perp X_4 \,|\, \{X_1, X_3\}.$$

This example corresponds to the so-called undirected "diamond" graph, which does not admit a faithful DAG, as implied by the following argument. Notwithstanding, the equivalence class $\mathfrak{D}(\Sigma)$ is still well-defined and can be partitioned into two sets $\mathfrak{D}(\Sigma) = \mathfrak{D}_1 \cup \mathfrak{D}_2$, where $\mathfrak{D}_j$ is a Markov equivalence class of DAGs for $j = 1, 2$. That is, the DAGs in each $\mathfrak{D}_j$ encode the same set of CI relations (via d-separation). In this example, each DAG in $\mathfrak{D}_1$ implies exactly one CI relation, namely $X_1 \perp\!\!\!\perp X_3 \,|\, \{X_2, X_4\}$, and similarly $\mathfrak{D}_2$ implies exactly one CI relation, namely $X_2 \perp\!\!\!\perp X_4 \,|\, \{X_1, X_3\}$. Using the notation we introduced above, if $\widetilde{B}(\pi) \in \mathfrak{D}_1$ then $\mathcal{I}(\pi) = \{(1, 3, \{2, 4\})\}$ and if $\widetilde{B}(\pi) \in \mathfrak{D}_2$ then $\mathcal{I}(\pi) = \{(2, 4, \{1, 3\})\}$.

Example 5.2. The deficiency shown in Example 5.1 is not unique to directed graphs: There exist Σ that are faithful to a DAG but that are not perfect with respect to any Markov network. An example is given by the so-called *v-structure*, i.e. $X_1 \to X_2 \leftarrow X_3$. Moreover, by combining a v-structure with the diamond graph in the previous example, we can construct a Gaussian distribution that is neither faithful nor perfect. Let $\mathcal{N}(0, \Sigma_1)$ denote the distribution in Example 5.1, $\mathcal{N}(0, \Sigma_2)$ denote any trivariate Gaussian distribution parametrized by a faithful v-structure, and define

$$\Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix} \in \mathbb{R}^{7 \times 7}.$$

Then $\mathcal{N}(0, \Sigma)$ is neither faithful nor perfect.

Evidently, directed and undirected graphs encode different sets of CI relationships. Furthermore, there exist examples where a minimal DAG is much more parsimonious than a minimal Markov network and vice versa (Koller and Friedman, 2009). The key is that in practice one cannot determine in advance which representation is optimal, so it is important to be able to learn as much as possible from either type of model. Is it possible to learn $\mathcal{I}(\Sigma)$ without the existence of a faithful DAG or a perfect Markov network?

To answer this question, we consider the use of the restricted minimizers $\widehat{B}(\pi)$ for a potentially large set of permutations to learn CI relations from data without assuming that $\mathcal{N}_p(0, \Sigma)$ is faithful to some DAG. One hopes

that by taking the union over all (or a large collection of) the CI relations obtained in this way, we subsume the CI relations implied by any single DAG or Markov network. This heuristic is justified by the following theorem:

Theorem 5.5. *Under the conditions of Theorem 5.1, with probability at least $1 - c_1 \exp(-c_2(d+1) \log p)$:*

(1) $\widehat{B}(\pi)$ is a minimal I-map of $\mathcal{N}_p(0, \Sigma)$ for each π;
(2) $\bigcup_{\pi \in \mathbb{S}_p} \widehat{\mathcal{I}}(\pi) = \mathcal{I}(\Sigma)$.

This result is a consequence of uniform model selection consistency in Theorem 4.1 and is proved in Section 8.5. Conclusion 1 implies that each of the learned graphs $\widehat{B}(\pi)$ are as parsimonious as possible, and Conclusion 2 shows that we can detect all CI relations by enumerating all permutations π. Moreover, if we take any subset $K \subset \mathbb{S}_p$ of permutations and compute $\cup_{\pi \in K} \widehat{\mathcal{I}}(\pi)$, Conclusion 2 implies that there will be no false positive errors in this set. Applied to Example 5.1, this shows that although no single permutation is sufficient to learn the full CI structure, *two* permutations suffice.

Since $\widehat{B}(\pi)$ corresponds to an autoregressive model (Section 2.2), it can be estimated efficiently and consistently via vanilla sparse regression. Theorem 5.5 then justifies testing any subset of permutations (possibly *all* permutations) to discover CI relations in a distribution. Since the set K can be as large as $|K| = p!$, establishing the uniform validity of all possible subsets is clearly nontrivial. Furthermore, Theorem 5.5 has an interesting implication for the PC algorithm (Kalisch and Bühlmann, 2007): It implies that one can use regression as an independence oracle (e.g. in place of the Fisher z-test) to detect CI relations consistently from high-dimensional data. Since the PC algorithm runs in $O(p^d)$ time, this gives an efficient way to perform the search over permutations described above.

6. COMPARISON

Our results are most closely related to van de Geer and Bühlmann (2013), who study a thresholded ℓ_0-penalized MLE under the same statistical model. Our results generalize their results in several ways:

(1) We prove that $\widehat{B}$ attains exact support recovery when $p \gg n$, which has not been shown previously to the best of our knowledge. For a more detailed discussion, see Remark 4.2.
(2) We have shown how $\widetilde{B}(\widehat{\pi})$ mimics the behaviour of an oracle minimizer of the expected penalized loss (Section 5.1.3).
(3) We generalize existing results on recovery of an equivariance DAG to high-dimensions and general nonconvex regularizers (Section 5.2). In order to exploit equivariance, one must essentially assume that the noise variance is known up to a multiplicative constant. By relaxing this assumption to minimum-trace DAGs, our results provide

theoretical guarantees in the practical setting where we have no prior knowledge about the error variances $\widetilde{\Omega}$.

(4) We allow for a general class of tractable penalties ρ_λ beyond the ℓ_0 penalty, which is known to be intractable. Although still non-convex, with a concave penalty the program (1.1) is defined over a continuous parameter space and can be solved exactly with dynamic programming (Silander and Myllymaki, 2006; Xiang and Kim, 2013).

(5) We consider optimization over the full space of parameters $\mathbb{D}_p$, instead of a thresholded parameter space as in van de Geer and Bühlmann (2013) (see their Condition 3.3). This requires a more delicate analysis and makes our method more practical since the thresholded parameter space involves an unknown constant which is difficult to select.

(6) We apply our results to the problem of CI learning (Section 5.3) and formally justify the use of penalized regression as an independence oracle inside the PC algorithm. These results are made feasible by our novel results on support recovery for all permutations π.

Our theorems on deviation bounds (Theorem 4.2) and sparsity (Theorem 5.2) are similar in spirit to van de Geer and Bühlmann (2013), however, our sparsity bounds use a minimum-trace DAG in the upper bound, as opposed to a minimal-edge DAG (Remark 5.1).

The present work is also related to Raskutti and Uhler (2014), which seeks to find the sparsest Bayesian network for a distribution $\mathbb{P}$ in a nonparametric setting. They show that their algorithm is uniformly consistent (Zhang and Spirtes, 2002) under strictly weaker assumptions than the PC algorithm (Kalisch and Bühlmann, 2007), although it is not discussed if these results are applicable when $p \gg n$ or what the scaling between (n, p, d) would be in such a setting. Their method is a two-stage method, with the primary step involving a constraint-based search similar to the PC algorithm. By contrast, we have focused on score-based estimators, which are faster and more popular in applications (Section 1.2). For example, compared to Raskutti and Uhler (2014), whose numerical simulations only explore small graphs with $p \leq 8$, Aragam and Zhou (2015) provide examples with $p = 8000$.

Finally, Shojaie and Michailidis (2010) proved the model selection consistency of an ℓ_1-based PLS estimator for a single, *fixed* permutation π^* under an irrepresentability condition. By assuming that π^* is known, they circumvent all of the technical challenges associated with estimating an *unknown* permutation, in effect proving the consistency of $\widehat{B}(\pi^*)$ for a *single* permutation. This is equivalent to estimating a standard autoregressive model, which is well-studied.

7. DISCUSSIONS

To conclude, we discuss various issues related to extending the results derived herein to more general settings.

7.1. **The ℓ_0 case.** Condition 4.2 precludes certain regularizers including the ℓ_0 penalty, however, a simple modification to the proofs accounts for any regularizer that is ℓ_0-compatible as defined in Definition 4.1, including the ℓ_0 penalty (see Remark A.1). Thus everything derived in this work applies equally to the ℓ_0 penalty. In particular, Theorem 5.2 implies $\|\widetilde{B}(\widehat{\pi})\|_0 \lesssim \|\widehat{B}\|_0 \lesssim \|\widetilde{B}(\pi_0)\|_0$. If $\widetilde{B}(\pi_0)$ is one of the sparsest DAGs in $\mathfrak{D}(\Sigma)$, then the number of edges in $\widehat{B}$ will be on the same order as that of the sparsest ones, as in van de Geer and Bühlmann (2013). For example, this occurs when an equivariance DAG is amongst the sparsest.

7.2. **Computation and loss functions.** Since (1.1) is a nonconvex program, computation of $\widehat{B}$ is challenging and in fact NP-hard (Chickering et al., 2004). Fortunately, there are fast algorithms via dynamic programming for finding globally optimal Bayesian networks (Ott and Miyano, 2003; Singh and Moore, 2005; Silander and Myllymaki, 2006). For example, by combining dynamic programming with A* search, Xiang and Kim (2013) propose an exact algorithm to compute the ℓ_1-regularized version of $\widehat{B}$ (cf. (1.3)) that is tractable on problems with hundreds of nodes.

This highlights one reason for choosing the least-squares loss, i.e. it leads to more efficient computation in practice. Moreover, the least-squares loss has the potential for generalization to subgaussian distributions for which there is no closed-form likelihood. What is necessary in our analysis, however, is that (1.3) factors into p neighbourhood problems (for each fixed ordering of the variables), as detailed in Lemma 8.1, a property that holds for other losses as well. For example, we could have used the corresponding negative log-likelihood (NLL) of (1.2), as in van de Geer and Bühlmann (2013), instead of least-squares. When used with the ℓ_0 penalty, the NLL has the appealing property of *score equivalence* (Heckerman et al., 1995), which assigns the same score to DAGs that are Markov equivalent. While conceptually appealing, our results show that this property is not necessary to identify a sparse DAG.

7.3. **Undirected models and extensions.** The problem of learning DAGs from data is closely related to the problem of learning undirected *Markov networks*. In particular, our results must be contrasted with the much simpler problem of estimating an undirected Gaussian graphical model, which has received much more attention. Meinshausen and Bühlmann (2006) showed how to reduce the estimation of a Gaussian graphical model to neighbourhood regression by regressing each X_j onto the rest of the variables X_{-j}. This procedure works for other families of undirected graphical models, including Ising models (Ravikumar et al., 2010), discrete models (Loh and Wainwright, 2013), and exponential random families (Yang et al., 2015). In this setting, there are only p neighbourhoods, compared to $O(p!)$ for DAGs.

As an immediate corollary of the results derived in this work, we also obtain consistency of this popular neighbourhood-regression based estimator

of the Gaussian graphical model. This follows since, in particular, we have uniform control over the neighbourhood problems $\widehat{\Theta}_\lambda(\mathbf{x}_j, \mathbf{X}; [p]_j)$ for each $j = 1, \ldots, p$. In this case, there are only p neighbourhood problems to control, much smaller than $p\binom{p}{d}$ for DAGs. One advantage of our results is the relaxation of the strong *neighbourhood stability* assumption (also known as the *irrepresentability condition*) through the use of nonconvex penalties such as the MCP and SCAD.

It is natural to inquire how these results extend to more general settings with non-Gaussian structural equations. For example, the neighourhood-level concepts (e.g. Definition 3.4 and Theorem 8.1) easily extend to sub-gaussian errors, and as a consequence much of the machinery developed here applies to subgaussian classes. The difficulty, however, is that the *population-level* structure of such structural equations (i.e. $\mathfrak{D}(\Sigma)$) is not well understood in the non-Gaussian setting, and so while one could obtain formal results in this direction it is not clear how strong the assumptions would be. For example, for different permutations the errors $\widetilde{\varepsilon}_j(\pi)$ are no longer guaranteed to be independent of the parents $X_{S_j(\pi)}$ if the joint distribution is not Gaussian. It is an interesting future direction to consider such extensions.

8. Outline of proofs

In the present section we outline the main ideas of our proof technique; detailed proofs of the various technical results are postponed to Appendices A–C. Sections 8.1 and 8.2 outline the proofs of Theorems 4.1 and 4.2, respectively, which involve most of the technical heavy lifting. We then outline how to apply these two theorems to derive the results in Sections 5.1, 5.2, and 5.3 in Sections 8.3, 8.4, and 8.5, respectively.

The proofs of Theorems 4.1 and 4.2 will be broken down into several steps. First, we establish some basic properties of the objective function and the probability space in order to reduce the neighbourhood regression analysis to a family of maximal sets denoted by $M_j(S)$ (Definition 8.2). Then we introduce the notion of monotonicity (Lemma 8.3) that is central to our proofs, and exploit this to provide a uniform bound on the probability of false selection for any neighbourhood problem (Proposition 8.5). Then, we will derive an independent result that gives a deviation bound for a *fixed* design regression problem (Theorem 8.1). Using a similar monotonicity argument as in the first step, we then show that this result can be applied uniformly to each neighbourhood problem, which yields our claimed deviation bounds (see Proposition B.1).

8.1. **Support recovery.** We begin by controlling the event

$$(8.1) \qquad \mathcal{B} := \{\operatorname{supp}(\widehat{B}(\pi)) \neq \operatorname{supp}(\widetilde{B}(\pi)) \ \exists \, \pi \in \mathbb{S}_p\}.$$

We will do this by reducing the analysis of $\widehat{B}(\pi)$ to a series of neighbourhood regression problems. There are two key steps: (i) Showing that each estimator $\widehat{B}(\pi)$ is equivalent to solving a series of p regression problems given by $\widehat{\Theta}_\lambda(\mathbf{x}_j, \mathbf{X}; S_j(\pi))$ (cf. Definition 3.3), and (ii) Controlling the total number of sets S that need to be considered.

8.1.1. *Reduction to neighbourhood regression.* Recall that the jth column of $\widehat{B}(\pi)$ is denoted by $\widehat{\beta}_j(\pi)$ and as usual, denote the sample version of $\widetilde{\varepsilon}_j(\pi)$ by boldface, i.e. $\widetilde{\boldsymbol{\varepsilon}}_j(\pi) := \mathbf{x}_j - \mathbf{X}\widetilde{\beta}_j(\pi)$. The first step above is justified by the following result. The symbol $\perp\!\!\!\perp$ is used here to denote independence of random variables.

Lemma 8.1. *Suppose* $\mathbf{X} \stackrel{iid}{\sim} \mathcal{N}_p(0, \Sigma)$ *and* $\lambda \geq 0$. *Then the following statements are true:*

(a) *For any* $j \in [p]$ *and* $\pi \in \mathbb{S}_p$, $\widetilde{\boldsymbol{\varepsilon}}_j(\pi) \perp\!\!\!\perp \mathbf{X}_{S_j(\pi)}$.

(b) *A matrix* $\widehat{B}(\pi) \in \mathbb{D}_p$ *is a restricted minimizer (Definition 2.2) if and only if* $\widehat{\beta}_j(\pi) \in \widehat{\Theta}_\lambda(\mathbf{x}_j, \mathbf{X}; S_j(\pi))$ *for each* $j \in [p]$.

(c) $\widehat{B} = \widehat{B}(\widehat{\pi})$ *is a global minimizer of* $Q(B)$ *if and only if* $\widehat{\beta}_j(\widehat{\pi}) \in \widehat{\Theta}_\lambda(\mathbf{x}_j, \mathbf{X}; S_j(\widehat{\pi}))$ *for each* $j \in [p]$ *and* $\widehat{\pi} \in \widehat{\mathbb{S}}_p$.

The proof of this lemma, which is a simple consequence of how the least-squares loss and the regularizer factor, is found in Appendix A.4. This allows us to formally establish the equivalence between the DAG problem and neighbourhood regression: In order to construct $\widehat{B}(\pi)$, it suffices to solve a neighbourhood regression problem for each column of $\widehat{B}(\pi)$, given by $\widehat{\Theta}_\lambda(\mathbf{x}_j, \mathbf{X}; S_j(\pi))$. A key observation is that through the independence established in Lemma 8.1(a) and a conditioning argument, we can reduce the regression problem given by $\widehat{\Theta}_\lambda(\mathbf{x}_j, \mathbf{X}; S_j(\pi))$ to a fixed design problem. The details are outlined in the proof of Proposition 8.5.

8.1.2. *Invariant sets and monotonicity.* As a consequence of Lemma 8.1, we have (cf. (8.1))

$$\mathcal{B} = \bigcup_{j=1}^{p} \{\mathrm{supp}(\widehat{\beta}_j(\pi)) \neq \mathrm{supp}(\widetilde{\beta}_j(\pi)) \ \exists \, \pi \in \mathbb{S}_p\}.$$

Since there are $p!$ total permutations, in principle the event $\{\mathrm{supp}(\widehat{\beta}_j(\pi)) \neq \mathrm{supp}(\widetilde{\beta}_j(\pi)) \exists \pi \in \mathbb{S}_p\}$ involves controlling a superexponential number of estimators, which seems hopeless. In order to reduce the total number of estimators we must control, we will introduce the notion of an *invariant set.*

First, recall the definition of $\beta_j(S)$ (cf. Definition 3.1) and for any $j \in [p]$ and $S \subset [d]_j$ define the error (or noise) for the associated neighbourhood regression as the following residual:

$$\varepsilon_j(S) := X_j - \beta_j(S)^T X.$$

The support set of $\beta_j(S)$ is denoted by $m_j(S) := \text{supp}(\beta_j(S))$ and the error variance by $\omega_j^2(S) := \text{var}(\varepsilon_j(S))$.

Definition 8.1. For any $S \subset [p]_j$, define a collection of subsets by

$$\mathcal{T}_j(S) := \{T \subset [p]_j : \beta_j(T) = \beta_j(S)\} = \{T \subset [p]_j : m_j(T) = m_j(S)\},$$

where $\beta_j(S)$ and $m_j(S)$ are defined in Definition 3.1. If $T \in \mathcal{T}_j(S)$, we call T an *invariant set of S for j*, or *S-invariant* for short.

In other words, for any j, $\mathcal{T}_j(S)$ is the collection of candidate sets $T \subset [p]_j$ such that the projection of X_j onto $\{X_i, i \in T\}$ is invariant. With some abuse of terminology, let us refer to $m_j(T) = \text{supp}(\beta_j(T))$ as the *support of neighbourhood T* (for node j). An equivalent description of $\mathcal{T}_j(S)$ is the set of neighbourhoods T whose support (for node j) is the same and equals $m_j(S)$.

Example 8.1. Continuing with Example 2.1, one can verify that $\mathcal{T}_3(\{1,2\}) = \{\{1\}, \{1,2\}\}$ and $m_3(\{1,2\}) = \{1\}$. Note also that $\mathcal{T}_3(S) = \mathcal{T}_3(\{1,2\})$ for any $S \in \mathcal{T}_3(\{1,2\})$. In the same example,

$$\{\{1\}, \{1,2\}\} = \mathcal{T}_3(\{1,2\}) \neq \mathcal{T}_3(\{1,4\}) = \{\{1,4\}, \{1,2,4\}\}.$$

The following lemma illustrates a crucial property of invariant sets:

Lemma 8.2. $T_1, T_2 \in \mathcal{T}_j(S) \implies T_1 \cup T_2 \in \mathcal{T}_j(S)$.

In other words, if two neighbourhoods share the same support, the union of these neighbourhoods must also have the same support. This justifies the following definition:

Definition 8.2. The unique largest element of $\mathcal{T}_j(S)$ shall be denoted by $M_j(S)$. Formally,

$$M_j(S) := \bigcup \mathcal{T}_j(S) = \bigcup \{T \subset [p]_j : \beta_j(T) = \beta_j(S)\}.$$

For instance, in Example 8.1 we have $M_3(\{1\}) = \{1,2\}$. The name "S-invariant set" comes from the fact that for any $T \in \mathcal{T}_j(S)$, we have the following useful identities:

(8.2) $$\beta_j(m_j(S)) = \beta_j(S) = \beta_j(T) = \beta_j(M_j(S)),$$

(8.3) $$\varepsilon_j(m_j(S)) = \varepsilon_j(S) = \varepsilon_j(T) = \varepsilon_j(M_j(S)).$$

The reason for introducing invariant sets is that it is generally *sufficient* to study the neighbourhood problem for $M_j(S)$ in the sense that once we have model selection consistency for each estimator in $\widehat{\Theta}_\lambda(\mathbf{x}_j, \mathbf{X}; M_j(S))$, the same is *guaranteed* for estimators based on every other neighbourhood in $\mathcal{T}_j(S)$. In fact, we have the following result, which says that model selection properties of the S-restricted estimators are monotone with respect to those sets S that contain the true support.

Lemma 8.3. *Suppose that $Z \in \mathbb{R}^{n \times m}$ is fixed and consider the regression problem $\mathbf{y} = Z\theta^* + \mathbf{w}$ for some $\theta^* \in \mathbb{R}^m$. If $\mathrm{supp}(\theta^*) \subset S \subset U$, then we have the following inclusion: $A(Z, \theta^*; S) \subset A(Z, \theta^*; U)$. In particular,*

$$\mathcal{A}(\mathbf{w}, Z, \theta^*; S) \subset \mathcal{A}(\mathbf{w}, Z, \theta^*; U),$$

where $A(Z, \theta^*; S)$ and $\mathcal{A}(\mathbf{w}, Z, \theta^*; S)$ are defined in (3.2)–(3.3).

Intuitively, for a fixed support, the set of "bad" noise vectors for the larger problem involving U is at least as big as the set of "bad" noise vectors for the smaller problem involving S. We are interested in the model selection failure of $\widehat{\beta}_j(S)$ for $\beta_j(S)$, which can be stated as

$$
(8.4) \quad \left\{ \begin{aligned} &\mathrm{supp}(\widehat{\beta}_j(S)) \neq \mathrm{supp}(\beta_j(S)) \\ &\text{for some } \widehat{\beta}_j(S) \in \widehat{\Theta}_\lambda(\mathbf{x}_j, \mathbf{X}; S) \end{aligned} \right\} = \mathcal{A}(\widetilde{\varepsilon}_j(S), \mathbf{X}, \beta_j(S); S)
$$

in the notation introduced in (3.3). The next result encapsulates a notion of *monotonicity* that is used throughout the rest of the proof, establishing that the invariant sets for any fixed neighbourhood are monotone in the same sense as Lemma 8.3.

Corollary 8.4. *Suppose $\mathbf{X} \overset{iid}{\sim} \mathcal{N}_p(0, \Sigma)$. For any $S \subset [p]_j$, we have*

$$\mathcal{A}\Big(\widetilde{\varepsilon}_j(S), \mathbf{X}, \beta_j(S); S\Big) \subset \mathcal{A}\Big(\widetilde{\varepsilon}_j(M_j(S)), \mathbf{X}, \beta_j(M_j(S)); M_j(S)\Big).$$

This proves that in order to control the neighbourhood regression problem for some set $S \subset [p]_j$, it suffices to control the strictly harder problem given by $M_j(S)$. Note that Lemma 8.4 is a *deterministic* statement about the events defined in (8.4).

8.1.3. *A bound on false selection.* For any $\Sigma \succ 0$ and fixed node X_j, define the following collections of subsets:

$$(8.5) \qquad m_j(\Sigma) := \{m_j(S) : S \subset [p]_j\},$$

$$(8.6) \qquad M_j(\Sigma) := \{M_j(S) : S \subset [p]_j\}.$$

Note that $|m_j(\Sigma)| = |M_j(\Sigma)|$. As long as it is clear whether the argument is a set S or a matrix Σ, this should not cause any confusion with $m_j(S)$ and $M_j(S)$. In Example 2.1, $m_3(\Sigma) = \{\emptyset, \{2\}, \{4\}, \{1, 2\}, \{2, 4\}, \{1, 2, 4\}\}$. In contrast, one can verify that $m_2(\Sigma) = \{\emptyset, \{3\}, \{4\}, \{3, 4\}, \{1, 3, 4\}\}$ which illustrates how some nodes have a reduced set of parent sets.

For any neighbourhood $S \subset [p]_j$, recall that the associated error variance is given by $\omega_j^2(S) = \mathrm{var}(\varepsilon_j(S))$. With some more abuse of notation, let

$$(8.7) \qquad \Phi_j(S) := \Phi_\lambda(\mathbf{X}_S, (\beta_j(S))_S, \omega_j^2(S)).$$

Note that we must restrict the SEM coefficients $\beta_j(S)$ to the subset S in order for this exponent to be well-defined. Since $\mathrm{supp}(\beta_j(S)) \subset S$, this does not change anything.

The following general result gives a uniform upper bound on the probability of false selection for any neighbourhood problem in terms of the maximal sets $M_j(T)$, and is the main ingredient in proving Theorem 4.1.

Proposition 8.5. *Fix $j \in [p]$. Under Condition 4.1, we have*

$$\mathbb{P}\left(\operatorname{supp}(\widehat{\beta}_j(S)) \neq \operatorname{supp}(\beta_j(S)), \, \exists\, S \subset [p]_j\right) \leq \sum_{T \in m_j(\Sigma)} \mathbb{E}e^{-\Phi_j(M_j(T))},$$

where $m_j(\Sigma)$ is defined by (8.5) and $\Phi_j(\,\cdot\,)$ is defined by (8.7).

The proof of this result can be found in Appendix A.8.

Proposition 8.5 says that to control the probability of false selection uniformly for all 2^{p-1} neighbourhoods S of the node j, it suffices to control a much smaller class of problems given by the neighbhourhoods $M_j(T)$ for each support set $T \in m_j(\Sigma)$. By Definition 3.5, $|m_j(S)| \leq d$ for all j and S, which implies that $|m_j(\Sigma)| \leq \binom{p}{d} \leq p^d$. Since there are 2^{p-1} subsets of $[p]_j$, $|m_j(\Sigma)| \leq p^d \ll 2^{p-1}$ as long as $d \ll p$; i.e. as long as d is much smaller than p the cardinality of $m_j(\Sigma)$ is much smaller than that of $2^{[p]_j}$. All of a sudden, instead of a superexponential number of neighbourhood problems to control, we have a subexponential (i.e. polynomial) number to control.

8.1.4. *Proof of Theorem 4.1.* For any $T \in m_j(\Sigma)$, Lemma 8.3 applied with $S = M_j(T)$ and $U = [p]$ yields

$$\Phi_j(M_j(T)) \geq \Phi_\lambda(\mathbf{X}, \, \beta_j(T), \, \omega_j^2(T)).$$

Recalling $d(\Sigma)$ and $\tau_*(\Sigma)$ in Definition 3.5, we have $\|\beta_j(T)\|_0 \leq d(\Sigma)$ and $\tau_*(\beta_j(T)) \geq \tau_*(\Sigma)$, as well as $\omega_j^2(T) \leq \sigma_{\max}^2$. The previous expression combined with (3.5) implies:

$$(8.8) \qquad \Phi_j(M_j(T)) \geq \Psi_\lambda(\mathbf{X}, \Sigma) \quad \text{for all } T \in M_j(\Sigma).$$

Combining Proposition 8.5, (8.8) and a union bound over $j \in [p]$,

$$\mathbb{P}\left(\operatorname{supp}(\widehat{\beta}_j(S)) \neq \operatorname{supp}(\beta_j(S)), \, \exists\, j \in [p], \, S \subset [p]_j\right)$$

$$\leq \sum_{j=1}^{p} \sum_{T \in m_j(\Sigma)} \mathbb{E}\exp(-\Phi_j(M_j(T))$$

$$(8.9) \qquad \leq p\binom{p}{d} \mathbb{E}\exp(-\Psi_\lambda(\mathbf{X}, \Sigma)),$$

since there are at most $\binom{p}{d}$ subsets in $m_j(\Sigma)$. In order to complete the proof, note that Lemma 8.1 implies

$$\left\{\operatorname{supp}(\widehat{B}(\pi)) \neq \operatorname{supp}(\widetilde{B}(\pi)), \, \exists\, \pi \in \mathbb{S}_p\right\}$$

$$= \left\{\operatorname{supp}(\widehat{\beta}_j(\pi)) \neq \operatorname{supp}(\widetilde{\beta}_j(\pi)), \, \exists\, \pi \in \mathbb{S}_p, \, j \in [p]\right\}$$

$$= \left\{\operatorname{supp}(\widehat{\beta}_j(S)) \neq \operatorname{supp}(\beta_j(S)), \, \exists\, S \subset [p]_j, \, j \in [p]\right\}.$$

In the last line we used $\widetilde{\beta}_j(\pi) = \beta_j(S_j(\pi))$. Combined with (8.9), this gives the desired result.

8.2. **Deviation bounds.** In order to establish uniform control over the probability of false selection for all possible neighbourhood regression problems in the previous section, we relied on a monotonicity property (cf. Lemma 8.3) of model selection. Unfortunately, this may not hold for weaker properties such as deviation bounds. In order to bound the deviations $\|\widehat{B}(\pi) - \widetilde{B}(\pi)\|_r$ $(r = 1, 2)$, we will use a modified argument that invokes a different kind of monotone class.

8.2.1. *An upper bound for fixed design.* We start by establishing a general bound on the ℓ_r $(r = 1, 2)$ estimation errors for a fixed design regression problem with a general regularizer ρ_λ. The objective here is to derive conditions under which we can guarantee such bounds for a fixed design problem, and then show that these conditions hold uniformly for all neighbourhood problems. The conditions we will need are familiar from the literature: A *Gaussian width condition* and a *restricted eigenvalue condition*.

For the rest of this subsection, we let $Z \in \mathbb{R}^{n \times m}$ and $w \in \mathbb{R}^n$ be a fixed matrix and fixed vector, respectively.

Definition 8.3 (Gaussian width). We say that the *Gaussian width (GW) condition* holds for (w, Z) relative to ρ_λ if there is a numerical constant $\delta \in (0, 1)$ such that

$$\frac{1}{n}|\langle w, Zu \rangle| \le \delta\left[\frac{1}{2n}\|Zu\|_2^2 + \rho_\lambda(u)\right], \ \forall u \in \mathbb{R}^m,$$

in which case we write $(w, Z) \in \mathrm{GW}_{\rho_\lambda}(\delta)$. If this inequality is strict for all $u \ne 0$, we write $(w, Z) \in \mathrm{GW}_{\rho_\lambda}^{\circ}(\delta)$.

We will be interested in the case where both w and Z are allowed to be random but independent. In this setting, for Gaussian designs considered in this paper, the GW condition holds with high probability for the ℓ_1 penalty (this follows from a standard Hölder inequality argument), and has similarly been shown to hold for penalties induced by ℓ_q norms for $0 \le q \le 1$ (Raskutti et al., 2011). Zhang and Zhang (2012) provide a version of this condition that applies to general nonconvex regularizers.

Before we proceed, let us note the following key relation between model selection consistency and the GW condition:

Lemma 8.6. *Consider the setup of Lemma 8.3, namely, the regression problem* $\mathbf{y} = Z\theta^* + \mathbf{w}$ *but with* $\theta^* = 0$. *Then*

$$\mathcal{A}(\mathbf{w}/\delta, \, Z, \, 0)^c = \big\{(\mathbf{w}, Z) \in \mathrm{GW}_{\rho_\lambda}^{\circ}(\delta)\big\}.$$

Thus, in order to ensure the GW condition for $(\mathbf{w}, Z_S)$, it suffices to show that the corresponding regression problem is model selection consistent when the *true coefficients are all set to zero* and the noise variance is inflated by

a factor of $1/\delta^2$. Zhang and Zhang (2012) refer to this property as *null-consistency*.

For any set $A \subset [m]$ and $\xi > 0$, define the following "cone":

$$(8.10) \qquad C_{\rho_\lambda}(A, \xi) := \{u \in \mathbb{R}^m : \rho_\lambda(u_{A^c}) \le \xi \rho_\lambda(u_A)\}.$$

This definition also depends on the ambient dimension m; when we wish to emphasize this we will write $C_{\rho_\lambda}^m(A, \xi)$. The term "cone" here is used in an extended sense, in analogy with the ℓ_1 cone found in previous work.

Definition 8.4 (Generalized restricted eigenvalue). The *generalized restricted eigenvalue (RE) constant* of Z with respect to ρ_λ over a subset A is

$$(8.11) \qquad \phi_{\rho_\lambda}^2(Z, A; \xi) := \inf\left\{ \frac{\|Zu\|_2^2}{n\|u\|_2^2} : u \in C_{\rho_\lambda}(A, \xi),\, u \ne 0 \right\}.$$

In the sequel, we often suppress the dependence of the generalized RE constants on λ and ξ, writing $\phi_\rho^2(Z, A) = \phi_{\rho_\lambda}^2(Z, A; \xi)$. Note that the usual restricted eigenvalue is equivalent to the special case $\rho_\lambda = \lambda\|\cdot\|_1$ (Bickel et al., 2009).

Consider the usual linear regression set up, $y = Z\theta^* + w$, where $\theta^* \in \mathbb{R}^m$ and we define $S^* = \text{supp}(\theta^*)$. The following general result establishes that the two conditions $(w, Z) \in \text{GW}_\rho(\delta)$ and $\phi_\rho^2(Z, S^*) > 0$ are sufficient to bound the deviation $\widehat{\theta} - \theta^*$:

Theorem 8.1. *Assume* $(w, Z) \in \text{GW}_{\rho_\lambda}(\delta)$ *for some* ρ_λ *satisfying Condition 4.2 and* $\delta \in (0, 1)$. *Let* $\xi = \xi(\delta) := (1 + \delta)/(1 - \delta)$ *and assume* $\phi^2 := \phi_\rho^2(Z, S^*; \xi) > 0$. *Then any* $\widehat{\theta} \in \widehat{\Theta}_\lambda(Z\theta^* + w, Z)$ *satisfies*

$$(8.12) \qquad \|\widehat{\theta} - \theta^*\|_2 \le \frac{2\,\xi}{\phi^2}\, \rho_\lambda'(0+)\|\theta^*\|_0^{1/2},$$

$$(8.13) \qquad \|\widehat{\theta} - \theta^*\|_1 \le \frac{2\,\xi(1 + \xi)}{\phi^2}\, \rho_\lambda'(0+)\|\theta^*\|_0.$$

The proof of Theorem 8.1 is found in Appendix A.9. The GW condition is quantified by the constant $\delta \in (0, 1)$, and the restricted eigenvalue condition depends on the free parameter $\xi > 0$; these two are linked via the relation $\xi(\delta) = (1 + \delta)/(1 - \delta)$ and play subtle roles in the proof. A slightly modified version of this result first appeared in Zhang and Zhang (2012), under different assumptions. The particular version presented here is important to derive uniform bounds for all permutations—see Section 8.2.2.

8.2.2. *Uniform deviation bounds.* Our strategy from here will be to show that the two sufficient conditions in Theorem 8.1—namely, the GW condition and the generalized RE condition—hold uniformly for all S with $w = \widetilde{\varepsilon}_j(S)$ and $Z = \mathbf{X}_S$. First, we provide a uniform bound on the restricted eigenvalues $\phi_\rho^2(\mathbf{X}_S, m_j(S))$ in terms of the smallest eigenvalue of Σ. More precisely, we show in Proposition B.4 that with high probability $\phi_\rho^2(\mathbf{X}_S, A) \ge r_{\min}(\Sigma)$ for all pairs $A \subset S$ with $|A| \le d$.

The next step is to show that with high probability, the GW condition holds for $(\widetilde{\varepsilon}_j(S), \mathbf{X}_S)$ for all j and S. First, let us define

$$(8.14) \qquad \mathcal{E}_S(\delta, \lambda) = \mathcal{E}_S(\delta, \lambda; j) := \left\{ \left(\widetilde{\varepsilon}_j(S), \mathbf{X}_S \right) \in \mathrm{GW}^\circ_{\rho_\lambda}(\delta) \right\}.$$

According to Lemma 8.6, we have $\mathcal{E}_S(\delta, \lambda) = \mathcal{A}\left(\widetilde{\varepsilon}_j(S)/\delta, \mathbf{X}, 0; S \right)^c$. An immediate consequence is that the complements of the events $\mathcal{E}_S(\delta, \lambda)$ are monotonic in the sense that they obey Corollary 8.4 when $\beta_j(S)$ is replaced with the zero vector. Analogous to the arguments leading up to Proposition 8.5 and its proof, this allows us to provide uniform control on $\mathcal{E}_S(\delta, \lambda)$, in a sense made precise by Proposition B.3. Once again, the key is to reduce the total number of events to control—*a priori* superexponential in size—down to a tractable number which can be controlled.

Together, Propositions B.3 and B.4 show that we have uniform control over both the restricted eigenvalues and the Gaussian widths for the neighbourhood problems $\widehat{\Theta}_\lambda(\mathbf{x}_j, \mathbf{X}; S)$. Thus, we can apply Theorem 8.1 to each of these problems and obtain, with high probability, deviation bounds of the form

$$\|\widehat{\beta}_j(S) - \beta_j(S)\|_2 \leq \frac{2\,\xi}{r_{\min}(\Sigma)} \rho'_\lambda(0+) \cdot \|\beta_j(S)\|_0^{1/2}$$

for all j and $S \subset [p]_j$. The precise statement can be found in Proposition B.1. Theorem 4.2 is an immediate consequence of these two results, the details of which are discussed in the next section.

8.2.3. *Proof of Theorem 4.2.* Let us consider the ℓ_2 bound, noting that the ℓ_1 version follows similarly:

$$(8.15) \qquad \|\widehat{B} - \widetilde{B}(\widehat{\pi})\|_2 \leq \frac{2\xi}{r_{\min}(\Sigma)} \rho'_\lambda(0+) \|\widetilde{B}(\widehat{\pi})\|_0^{1/2}.$$

(8.15) can be seen by applying Proposition B.1 with $S = S_j(\widehat{\pi})$, and noting that $\widehat{\beta}_j(\widehat{\pi}) = \widehat{\beta}_j(S_j(\widehat{\pi}))$, $\widetilde{\beta}_j(\widehat{\pi}) = \widetilde{\beta}_j(S_j(\widehat{\pi}))$, and the fact that $\|u_j - v_j\|_2 \leq a\|v_j\|_0^{1/2}$ for all j, implies $\sum_j \|u_j - v_j\|_2^2 \leq a^2 \sum_j \|v_j\|_0$. The bound on the probability of (8.15) follows by combining Proposition B.1 with Proposition B.2.

8.3. **Score-based learning.** We now show how Theorems 4.1 and 4.2 can be applied to derive the results in Section 5.1. Theorem 5.1 is an immediate corollary of Theorems 4.1 and 4.2 applied to the special case $\pi = \widehat{\pi}$. Next, we control the sparsity of the estimate $\widehat{B}$ and the candidate DAG $\widetilde{B}(\widehat{\pi})$ via Theorems 5.2 and 5.3.

We start with a kind of basic inequality that is adapted to the present, nonidentifiable setting:

Lemma 8.7. *Let* $\mathbf{E}(\pi) := \mathbf{X} - \mathbf{X}\widetilde{B}(\pi)$. *For any* $\pi \in \mathbb{S}_p$ *and* $\widehat{\pi} \in \widehat{\mathbb{S}}_p$,

$$
\frac{1}{2n}\|\mathbf{X}(\widetilde{B}(\widehat{\pi}) - \widehat{B})\|_2^2 + \rho_\lambda(\widehat{B}) \ \leq \ \frac{1}{2n}\|\mathbf{E}(\pi)\|_2^2 - \frac{1}{2n}\|\mathbf{E}(\widehat{\pi})\|_2^2
$$

(8.16)
$$
+ \frac{1}{n}\operatorname{tr}\left(\mathbf{E}(\widehat{\pi})^T\mathbf{X}(\widetilde{B}(\widehat{\pi}) - \widehat{B})\right)
$$
$$
+ \rho_\lambda(\widetilde{B}(\pi)).
$$

In contrast to the basic inequality used in usual regression, (8.16) also leverages the many possible decompositions of $\mathbf{X}$, one for each permutation. Three terms in this inequality are particularly important:

(1) The difference in residuals $\|\mathbf{E}(\pi)\|_2^2/(2n) - \|\mathbf{E}(\widehat{\pi})\|_2^2/(2n)$ explains the origin of the minimum-trace permutation: We would like to make $\|\mathbf{E}(\pi)\|_2^2/(2n)$ as small as possible in order to minimize this difference. By standard concentration arguments, $\|\mathbf{E}(\pi)\|_2^2/n$ is close to its expectation, $\operatorname{tr}\widetilde{\Omega}(\pi)$ (cf. (5.4)). Hence, we choose π to minimize $\operatorname{tr}\widetilde{\Omega}(\pi)$. The details of this argument are in Appendix C.6; the explicit upper bound we use is detailed in Proposition C.5.

(2) The quantity $\operatorname{tr}(\mathbf{E}(\widehat{\pi})^T\mathbf{X}(\widetilde{B}(\widehat{\pi}) - \widehat{B}))/n$ can be bounded using the Gaussian width condition (Definition 8.3). There is a subtlety regarding whether to decompose this along rows or columns; see Lemma C.1.

(3) The penalty on $\widehat{B}$ can be replaced with $\rho_\lambda(\widetilde{B}(\widehat{\pi}))$ by showing that $\rho_\lambda(\widehat{B}) \gtrsim \rho_\lambda(\widetilde{B}(\widehat{\pi}))$ (Lemma C.2). This is true whenever the deviation $\widehat{B} - \widetilde{B}(\widehat{\pi})$ is not too large, which is guaranteed by Theorem 4.2.

Once we have established control of these three terms, the details of which are found in Appendix C, it is not hard to prove Theorems 5.2 and 5.3. More precisely, for Theorem 5.2, we have the following bound in terms of the constants δ (cf. Definition 8.3) and a_2 (cf. Condition 5.2):

$$
\frac{2\delta}{1-\delta}\rho_\lambda(\widetilde{B}(\widehat{\pi})) \leq \rho_\lambda(\widehat{B}) \leq \frac{2}{1-\delta}\left(1 + \frac{10}{a_2}\right)\rho_\lambda(\widetilde{B}(\pi_0)).
$$

The precise statement and proof are given by Proposition C.3, and Theorem 5.2 follows as a consequence of this bound and an ℓ_0-compatibility argument (see Appendix C.3). The oracle inequality Theorem 5.3 follows from a straightforward manipulation of (8.16) combined with a Gaussian concentration argument (Lemma C.7), see Appendix C.4 for details.

8.4. **Causal DAG learning.** The main result in Section 5.2 is Theorem 5.4, whose proof is similar to the proofs in the previous section and follows mostly from Lemma 8.7. The complete proof is in Appendix C.5.

8.5. **Conditional independence learning.** Theorem 5.5 from Section 5.3 follows almost immediately from Theorem 4.1. For completeness, we include a proof below.

8.5.1. *Proof of Theorem 5.5.* Recall that $\widetilde{B}(\pi)$ is always a minimal I-map for $\mathcal{N}_p(0, \Sigma)$ and $\bigcup_{\pi \in \mathbb{S}_p} \mathcal{I}(\pi) = \mathcal{I}(\Sigma)$ (e.g. Koller and Friedman, 2009). Theorem 4.1, combined with Lemma 4.1, implies that

$$\mathbb{P}\left(\bigcap_{\pi \in \mathbb{S}_p} \left\{ \mathrm{supp}(\widehat{B}(\pi)) = \mathrm{supp}(\widetilde{B}(\pi)) \right\} \right) \geq 1 - c_1 \exp(-c_2(d+1)\log p).$$

But $\mathrm{supp}(\widehat{B}(\pi)) = \mathrm{supp}(\widetilde{B}(\pi))$ implies $\widehat{\mathcal{I}}(\pi) = \mathcal{I}(\pi)$ (i.e. d-separation is a property of the graph alone and is independent of the exact values of the parameters). Since $\widetilde{B}(\pi)$ is a minimal I-map for each π, $\widehat{B}(\pi)$ must also be a minimal I-map. Furthermore, we deduce that

$$\mathbb{P}\left(\bigcup_{\pi \in \mathbb{S}_p} \widehat{\mathcal{I}}(\pi) = \bigcup_{\pi \in \mathbb{S}_p} \mathcal{I}(\pi) \right) \geq \mathbb{P}\left(\bigcap_{\pi \in \mathbb{S}_p} \left\{ \widehat{\mathcal{I}}(\pi) = \mathcal{I}(\pi) \right\} \right)$$

$$\geq \mathbb{P}\left(\bigcap_{\pi \in \mathbb{S}_p} \left\{ \mathrm{supp}(\widehat{B}(\pi)) = \mathrm{supp}(\widetilde{B}(\pi)) \right\} \right)$$

$$\geq 1 - c_1 \exp(-c_2(d+1)\log p).$$

Using $\bigcup_{\pi \in \mathbb{S}_p} \mathcal{I}(\pi) = \mathcal{I}(\Sigma)$ completes the proof.

APPENDIX A. PROOFS OF TECHNICAL RESULTS

A.1. **Proof of Lemma 2.1.** We need the following simple lemma, which follows since $P_\pi A = PAP^T$ for some permutation matrix P:

Lemma A.1. $A = MNM^T \iff P_\pi A = (P_\pi M)(P_\pi N)(P_\pi M)^T$.

Recall the modified Cholesky decomposition of A (also called the LDLT decomposition): $A = LDL^T$ for a lower triangular matrix L, with unit diagonal entries, and a diagonal matrix D. When A is positive definite, the pair (L, D) is unique and we refer to it as the *Cholesky decomposition of A*.

Let us denote the set of all pairs $(\widetilde{B}, \widetilde{\Omega})$ satisfying $\Sigma^{-1} = (I - \widetilde{B})\widetilde{\Omega}^{-1}(I - \widetilde{B}^T)$ (equivalently, (2.1)) as $\mathfrak{D}'$. Next, note that $\widetilde{B} \in \mathbb{D}_p$ if and only if $P_\pi \widetilde{B}$ is lower triangular for some permutation π. Lemma A.1 implies that $(\widetilde{B}, \widetilde{\Omega}) \in \mathfrak{D}'$ iff $(I - P_\pi \widetilde{B}, P_\pi \widetilde{\Omega}^{-1})$ is a Cholesky decomposition of $P_\pi \Sigma^{-1}$ for some π.

By the definition (2.5), $(I - P_\pi \widetilde{B}(\pi), P_\pi \widetilde{\Omega}(\pi)^{-1})$ is also a Cholesky decomposition of $P_\pi \Sigma^{-1}$. Since the Cholesky decomposition is unique for positive definite matrices, we have $(\widetilde{B}, \widetilde{\Omega}) \in \mathfrak{D}'$ iff $(\widetilde{B}, \widetilde{\Omega}) = (\widetilde{B}(\pi), \widetilde{\Omega}(\pi))$ for some π, which gives the desired result, since $\mathfrak{D}(\Sigma)$ is the projection of $\mathfrak{D}'$ onto its first coordinate.

A.2. **Proof of Lemma 5.1.** Consider the following program:

$$(A.1) \qquad \min \sum_{j=1}^{p} x_j^2 \text{ subject to } \sum_{j=1}^{p} \log x_j^2 = C.$$

The solution to this program is given by $x_j^2 = e^{C/p}$ for all $j = 1, \ldots, p$. In other words, the minimum is attained by a constant vector. It is straightforward to verify that $\log \det \widetilde{\Omega}(\pi) = \log \det \Sigma$ and hence $\log \det \widetilde{\Omega}(\pi) = \sum_j \log \widetilde{\omega}_j^2(\pi)$ is constant for all $\pi \in \mathbb{S}_p$. Thus for any $\pi \in \mathbb{S}_p$, the vector $(\widetilde{\omega}_1^2(\pi), \ldots, \widetilde{\omega}_p^2(\pi)) \in \mathbb{R}^p$ is feasible for (A.1), which implies that $\operatorname{tr} \widetilde{\Omega}(\pi)$ is minimized whenever $\widetilde{\omega}_1^2(\pi) = \cdots = \widetilde{\omega}_p^2(\pi)$. Finally, uniqueness of $\widetilde{B}(\pi_0)$ follows from Theorem 1 in Peters and Bühlmann (2013).

A.3. **Proof of Lemma 4.1.** This is a consequence of Theorem 4.2 in Huang et al. (2012) and Proposition 2 in Zhang and Huang (2008). Briefly, Huang et al. (2012) show that the least-squares MCP estimator correctly recovers the support of a linear model as long as the so-called *sparse Riesz condition* holds. We then use Zhang and Huang (2008) to bound the probability that **X** satisfies this condition. For the special case $\beta_j(S) = 0$ (which is not covered by Huang et al., 2012) we can invoke Proposition B.3.

A.4. **Proof of Lemma 8.1.** The first conclusion (a) follows from elementary properties of conditional expectation and the identity

$$\mathbb{E}(X_j \mid X_{S_j(\pi)}) = \widetilde{\beta}_j(\pi)^T X.$$

Since (c) is a special case of (b), it suffices to prove (b).

Fix $\pi \in \mathbb{S}_p$ and let $S_j = S_j(\pi)$. If $\widehat{\beta}_j(\pi) \in \widehat{\Theta}_\lambda(\mathbf{x}_j, \mathbf{X}; S_j)$ for each j, then evidently $\widehat{B}(\pi) = [\widehat{\beta}_1(\pi) \mid \cdots \mid \widehat{\beta}_p(\pi)]$ minimizes $Q(B)$ over $\mathbb{D}_p[\pi]$. For the reverse direction, recall that $\mathbf{X}_{S_j}$ is the $n \times |S_j|$ matrix formed by extracting the columns in S_j, and similarly for $(\beta_j)_{S_j}$. For any $B \in \mathbb{D}_p[\pi]$ we have $(\beta_j)_{S_j^c} = 0$ for each j, so we can write

$$\frac{1}{2n}\|\mathbf{X} - \mathbf{X}B\|_2^2 + \rho_\lambda(B) = \sum_{j=1}^p \left\{ \frac{1}{2n}\|\mathbf{x}_j - \mathbf{X}\beta_j\|_2^2 + \rho_\lambda(\beta_j) \right\}$$

$$= \sum_{j=1}^p \left\{ \frac{1}{2n}\|\mathbf{x}_j - \mathbf{X}_{S_j}(\beta_j)_{S_j}\|_2^2 + \rho_\lambda((\beta_j)_{S_j}) \right\}.$$

Then $\widehat{B}(\pi) = [\widehat{\beta}_1(\pi) \mid \cdots \mid \widehat{\beta}_p(\pi)] \in \min_{\mathbb{D}_p[\pi]} Q(B)$ if and only if

$$\widehat{\beta}_j(\pi) \in \arg\min_\beta \frac{1}{2n}\|\mathbf{x}_j - \mathbf{X}\beta\|_2^2 + \rho_\lambda(\beta) \quad \text{subject to } \beta_{S_j^c} = 0.$$

In other words, $\widehat{\beta}_j(\pi) \in \widehat{\Theta}_\lambda(\mathbf{x}_j, \mathbf{X}; S_j)$ for each j. Since π was arbitrary, the desired claim follows.

A.5. **Proof of Lemma 8.2.** The proof relies on the following property of L^2 projections: For any two sets $S, R \subset [p]_j$, we have

(A.2) $$\beta_j(S \cup R) = \beta_j(S) \iff \varepsilon_j(S) \perp\!\!\!\perp X_i, \ \forall i \in R.$$

To lighten the notation, let $S^* = m_j(S)$. Note that $\beta_j(S) = \beta_j(S^*)$ since $\operatorname{supp}(\beta_j(S)) = S^*$. It follows from (A.2) that $\varepsilon_j(S^*) \perp\!\!\!\perp X_i$ for $i \in S \setminus S^*$.

Similarly, since $\mathrm{supp}(\beta_j(T_k)) = S^*$, we have $\varepsilon_j(S^*) \perp\!\!\!\perp X_i$ for $i \in T_k \setminus S^*$ and $k = 1, 2$. It follows that

$$\varepsilon_j(S^*) \perp\!\!\!\perp X_i, \ \forall i \in (T_1 \setminus S^*) \cup (T_2 \setminus S^*)$$

hence the application of (A.2) in the reverse direction yields

$$\beta_j(T_1 \cup T_2) = \beta_j\big(S^* \cup (T_1 \setminus S^*) \cup (T_2 \setminus S^*)\big) = \beta_j(S^*) = \beta_j(S).$$

A.6. **Proof of Lemma 8.3.** It suffices to show

$$A(Z, \theta^*; U)^c \subset A(Z, \theta^*; S)^c.$$

Suppose $w \in A(Z, \theta^*; U)^c$, i.e., $\mathrm{supp}(\widetilde{\theta}) = \mathrm{supp}(\theta^*) := S^*$ for any $\widetilde{\theta} \in \widehat{\Theta}_\lambda(Z\theta^* + w, Z; U)$. We wish to show that for any $\widehat{\theta} \in \widehat{\Theta}_\lambda(Z\theta^* + w, Z; S)$, it must also be true that $\mathrm{supp}(\widehat{\theta}) = S^*$. Let

$$F(\theta) = \frac{1}{2n}\|Z(\theta^* - \theta) + w\|_2^2 + \rho_\lambda(\theta)$$

denote the objective function in Definition 3.2 of $\widehat{\Theta}_\lambda(y, Z; S)$ with $y = Z\theta^* + w$. Since $\mathrm{supp}(\widehat{\theta}) \subset S \subset U$, $\widehat{\theta}$ is feasible for the U-restricted problem, whence

$$F(\widetilde{\theta}) \leq F(\widehat{\theta})$$

for any $\widetilde{\theta} \in \widehat{\Theta}_\lambda(Z\theta^* + w, Z; U)$. But $\widetilde{\theta}$ is also feasible for the S-restricted problem since $\mathrm{supp}(\widetilde{\theta}) = S^* \subset S$, so that

$$F(\widetilde{\theta}) \geq F(\widehat{\theta}) \implies F(\widetilde{\theta}) = F(\widehat{\theta}).$$

Since the value $F(\widetilde{\theta})$ is by definition the global minimum of F for the U-restricted problem and $\mathrm{supp}(\widehat{\theta}) \subset U$, $\widehat{\theta}$ must be a global minimizer of F for the U-restricted problem, i.e., $\widehat{\theta} \in \widehat{\Theta}_\lambda(Z\theta^* + w, Z; U)$, whence $\mathrm{supp}(\widehat{\theta}) = S^*$ as desired.

A.7. **Proof of Corollary 8.4.** By Lemma 8.3 and the fact that $S \subset M_j(S)$, we have

(A.3) $\quad \mathcal{A}\Big(\widetilde{\varepsilon}_j(S), \mathbf{X}, \beta_j(S); S\Big) \subset \mathcal{A}\Big(\widetilde{\varepsilon}_j(S), \mathbf{X}, \beta_j(S); M_j(S)\Big).$

Using (8.2) and (8.3), we have the following identity:

$$\mathcal{A}\Big(\widetilde{\varepsilon}_j(S), \mathbf{X}, \beta_j(S); M_j(S)\Big) = \mathcal{A}\Big(\widetilde{\varepsilon}_j(M_j(S)), \mathbf{X}, \beta_j(M_j(S)); M_j(S)\Big).$$

Plugging this into (A.3) yields the desired result.

A.8. **Proof of Proposition 8.5.** Throughout, for simplicity, let

$$\mathcal{A}_S := \mathcal{A}(\widetilde{\varepsilon}_j(S), \mathbf{X}, \beta_j(S); S).$$

Fix $S \subset [p]_j$ and let $\theta^* = \beta_j(S)$, $s^* = |m_j(S)| = \|\theta^*\|_0$ and $\varepsilon^* = \widetilde{\varepsilon}_j(S)$ so that $\mathcal{A}_S = \mathcal{A}(\varepsilon^*, \mathbf{X}, \theta^*; S)$. Note that $\mathcal{A}(\varepsilon^*, \mathbf{X}, \theta^*; S)$ represents the following model selection failure:

$$\operatorname{supp}(\widehat{\theta}) \neq \operatorname{supp}(\theta^*) \quad \exists \widehat{\theta} \in \widehat{\Theta}_\lambda(\mathbf{X}\theta^* + \varepsilon^*, \mathbf{X}; S).$$

Since $\operatorname{supp}(\theta^*) \subset S$, we can restrict $\mathbf{X}$ and θ^* to S, so that the above is equivalent to

$$\operatorname{supp}(\widehat{\theta}) \neq \operatorname{supp}(\theta^*_S) \quad \exists \widehat{\theta} \in \widehat{\Theta}_\lambda(\mathbf{X}_S\theta^*_S + \varepsilon^*, \mathbf{X}_S).$$

which is the same event as $\mathcal{A}(\varepsilon^*, \mathbf{X}_S, \theta^*_S)$. To summarize, $\mathcal{A}_S = \mathcal{A}(\varepsilon^*, \mathbf{X}_S, \theta^*_S)$.

Since ε^* is independent of $\mathbf{X}_S$ by Lemma 8.1(a), by conditioning on $\mathbf{X}_S$ we are dealing with a fixed design regression problem with Gaussian noise $\varepsilon^* = \widetilde{\varepsilon}_j(S) \sim \mathcal{N}_n(0, \omega_j^2(S)I_n)$. We obtain

$$\begin{aligned}
\mathbb{P}(\mathcal{A}_S) &= \mathbb{E}\Big[\, \mathbb{P}\big(\mathcal{A}(\varepsilon^*, \mathbf{X}_S, \theta^*_S)\big) \mid \mathbf{X}_S\Big] \\
&\leq \mathbb{E}\exp[-\Phi_\lambda(\mathbf{X}_S, \theta^*_S, \omega_j^2(S))] \\
&= \mathbb{E}\exp(-\Phi_j(S)),
\end{aligned}$$

where the last line uses (8.7). Now we have

$$\{\operatorname{supp}(\widehat{\beta}_j(S)) \neq \operatorname{supp}(\beta_j(S)),\ \exists\, S \subset [p]_j\} = \bigcup_{S \subset [p]_j} \mathcal{A}_S \subset \bigcup_{T \in m_j(\Sigma)} \mathcal{A}_{M_j(T)},$$

where the equality is by (8.4) and the inclusion follows from Corollary 8.4. Note that this is the key step where the reduction occurs. Hence, we have

$$\begin{aligned}
\mathbb{P}\Big(\bigcup_{S \subset [p]_j} \mathcal{A}_S\Big) &\leq \mathbb{P}\Big(\bigcup_{T \in m_j(\Sigma)} \mathcal{A}_{M_j(T)}\Big) \\
&\leq \sum_{T \in m_j(\Sigma)} \mathbb{P}(\mathcal{A}_{M_j(T)}) \leq \sum_{T \in m_j(\Sigma)} \mathbb{E}\exp(-\Phi_j(M_j(T))),
\end{aligned}$$

which is the desired probability bound.

A.9. **Proof of Theorem 8.1.** Recall that $S^* := \operatorname{supp}(\theta^*)$. To lighten notation, for any vector u let $u_1 := u_{S^*}$, $u_2 := u_{(S^*)^c}$, and also $\Delta := \widehat{\theta} - \theta^*$. Then invoking the subadditivity of ρ_λ (this is a consequence of Condition 4.2),

$$\begin{aligned}
\rho_\lambda(\widehat{\theta}) - \rho_\lambda(\theta^*) &= \rho_\lambda(\Delta + \theta^*) - \rho_\lambda(\theta^*) \\
&= \rho_\lambda(\Delta_1 + \theta^*_1) + \rho_\lambda(\Delta_2) - \rho_\lambda(\theta^*_1)
\end{aligned}$$

$$\text{(A.4)} \qquad\qquad\qquad \geq -\rho_\lambda(\Delta_1) + \rho_\lambda(\Delta_2).$$

It is straightforward to derive

$$\text{(A.5)} \qquad \frac{1}{2n}\|y - Z\widehat{\theta}\|_2^2 - \frac{1}{2n}\|y - Z\theta^*\|_2^2 = \frac{1}{2n}\|Z\Delta\|^2 - \frac{1}{n}\langle w, Z\Delta\rangle.$$

Since $(w, Z) \in \mathrm{GW}_{\rho_\lambda}(\delta)$, we can invoke the GW condition with $u = \Delta$,

$$(A.6) \qquad -\frac{1}{n}\langle w, Z\Delta\rangle \geq -\frac{1}{n}|\langle w, Z\Delta\rangle| \geq -\delta\frac{1}{2n}\|Z\Delta\|^2 - \delta\rho_\lambda(\Delta).$$

It follows that

$$
\begin{aligned}
0 &\geq \frac{1}{2n}\|y - Z\widehat{\theta}\|_2^2 - \frac{1}{2n}\|y - Z\theta^*\|_2^2 + \rho_\lambda(\widehat{\theta}) - \rho_\lambda(\theta^*) \\
&\geq \frac{1}{2n}\|Z\Delta\|^2 - \frac{1}{n}\langle w, Z\Delta\rangle - \rho_\lambda(\Delta_1) + \rho_\lambda(\Delta_2) \\
&\geq \frac{1-\delta}{2n}\|Z\Delta\|^2 - \delta\rho_\lambda(\Delta) - \rho_\lambda(\Delta_1) + \rho_\lambda(\Delta_2) \\
&= \frac{1-\delta}{2n}\|Z\Delta\|^2 - (1+\delta)\rho_\lambda(\Delta_1) + (1-\delta)\rho_\lambda(\Delta_2)
\end{aligned}
$$

$$(A.7) \qquad = (1-\delta)\Big[\frac{1}{2n}\|Z\Delta\|^2 + \rho_\lambda(\Delta_2) - \xi\rho_\lambda(\Delta_1)\Big],$$

where the first inequality by optimality of $\widehat{\theta}$, the second by (A.5), and the third by (A.6). The next line follows from an an application of $\rho_\lambda(\Delta) = \rho_\lambda(\Delta_1) + \rho_\lambda(\Delta_2)$. Since $\delta < 1$ by assumption, it follows that $\rho_\lambda(\Delta_2) \leq \xi\rho_\lambda(\Delta_1)$ which implies $\Delta \in C_\rho(S^*, \xi(\delta))$.

Recalling the definition (8.11) of $\phi_\rho^2(Z, S^*)$, we conclude that $\frac{1}{2n}\|Z\Delta\|^2 \geq \frac{\phi^2}{2}\|\Delta\|_2^2$ which combined with (A.7), dropping $\rho_\lambda(\Delta_2)$, gives

$$0 \geq \frac{\phi^2}{2}\|\Delta\|_2^2 - \xi\rho_\lambda(\Delta_1).$$

Combining with the following (note $\|\Delta_1\|_0 \leq \|\theta^*\|_0$),

$$(A.8) \qquad \rho_\lambda(\Delta_1) \leq \rho_\lambda'(0+)\|\Delta_1\|_1 \leq \rho_\lambda'(0+)\|\theta^*\|_0^{1/2}\|\Delta\|_2$$

and re-arranging proves (8.12). For (8.13), since $\Delta \in C_\rho(S^*, \xi(\delta))$, we can use Lemma B.7 to construct a set $M \subset [p]$ with $|M| = |S^*| = \|\theta^*\|_0$ such that $\Delta \in C_1(M, \xi(\delta))$. Then

$$
\begin{aligned}
\|\Delta\|_1 = \|\Delta_M\|_1 + \|\Delta_{M^c}\|_1 &\leq (1+\xi)\|\Delta_M\|_1 \\
&\leq (1+\xi)\|\theta^*\|_0^{1/2}\|\Delta_M\|_2 \\
&\leq \frac{2\xi(1+\xi)}{\phi^2} \cdot \rho_\lambda'(0+)\|\theta^*\|_0.
\end{aligned}
$$

Remark A.1. Theorem 8.1 also applies to the ℓ_0 penalty, even though Condition 4.2 does not hold. Simply replace (A.8) with

$$\rho_\lambda(\Delta_1) \leq \overline{\rho}_0\lambda^2\|\Delta_1\|_0 \leq \overline{\rho}_0\lambda^2\|\theta^*\|_0.$$

The rest of the proof requires no changes. The resulting ℓ_2 bound (8.12) becomes

$$\|\widehat{\theta} - \theta^*\|_2 \leq \sqrt{\frac{2\overline{\rho}_0\xi}{\phi^2}} \cdot \lambda\|\theta^*\|_0^{1/2}.$$

The proof of Lemma C.2 in Appendix C also simplifies significantly under the ℓ_0 penalty.

A.10. **Proof of Lemma 8.6.** If $(\mathbf{w}, Z) \in \mathrm{GW}^\circ_{\rho_\lambda}(\delta)$, then for any $u \neq 0$,

$$\frac{\delta}{2n}\|Zu\|_2^2 - \frac{1}{n}\mathbf{w}^T Zu + \delta\rho_\lambda(u) > 0$$

$$\iff \frac{1}{2n}\|\mathbf{w}/\delta - Zu\|_2^2 + \rho_\lambda(u) > \frac{1}{2n}\|\mathbf{w}/\delta\|_2^2.$$

The latter inequality implies

$$\{0\} = \arg\min_u \|\mathbf{w}/\delta - Zu\|_2^2/(2n) + \rho_\lambda(u),$$

that is, 0 is the unique global minimizer of the right hand side. Recalling the definition of $\mathcal{A}(\mathbf{w}/\delta, Z, 0)$ in (3.3), we obtain the desired result.

A.11. **Proof of Lemma 8.7.** Observe that for any $\pi \in \mathbb{S}_p$,

$$(A.9) \qquad\qquad Q(\widehat{B}) \leq Q(\widehat{B}(\pi)) \leq Q(\widetilde{B}(\pi)),$$

where $\widehat{B}(\pi)$ is a restricted minimizer as in (2.8). Moreover, we have the following alternative expression for Q:

$$(A.10) \quad Q(B) = \frac{1}{2n}\|\mathbf{X}(\widetilde{B}(\widehat{\pi}) - B) + \mathbf{E}(\widehat{\pi})\|_2^2 + \rho_\lambda(B), \quad \text{for any } \widehat{\pi} \in \widehat{\mathbb{S}}_p.$$

Thus, using (A.9) and (A.10),

$$0 \leq Q(\widetilde{B}(\pi)) - Q(\widehat{B})$$

$$= \frac{1}{2n}\|\mathbf{E}(\pi)\|_2^2 - \frac{1}{2n}\|\mathbf{X}(\widetilde{B}(\widehat{\pi}) - \widehat{B}) - \mathbf{E}(\widehat{\pi})\|_2^2 + \rho_\lambda(\widetilde{B}(\pi)) - \rho_\lambda(\widehat{B})$$

$$= \frac{1}{2n}\|\mathbf{E}(\pi)\|_2^2 - \frac{1}{2n}\|\mathbf{E}(\widehat{\pi})\|_2^2 - \frac{1}{2n}\|\mathbf{X}(\widetilde{B}(\widehat{\pi}) - \widehat{B})\|_2^2$$

$$\qquad + \frac{1}{n}\operatorname{tr}\left(\mathbf{E}(\widehat{\pi})^T\mathbf{X}(\widetilde{B}(\widehat{\pi}) - \widehat{B})\right) + \rho_\lambda(\widetilde{B}(\pi)) - \rho_\lambda(\widehat{B}).$$

Since (A.9) holds for any π, this completes the proof.

APPENDIX B. AUXILIARY RESULTS FOR DEVIATION BOUNDS

This section contains the bulk of our technical results on controlling the deviations $\widehat{\beta}_j(S) - \beta_j(S)$ for a neighbourhood problem. These constitute the main ingredients used in proving Theorem 4.2, which we also do in this section.

For any $\delta \in (0,1)$ and $\lambda \geq 0$, define the following events:

$$(B.1) \qquad \mathcal{E}(\delta, \lambda) = \left\{ \left(\widetilde{\boldsymbol{\varepsilon}}_j(S), \mathbf{X}_S\right) \in \mathrm{GW}^\circ_{\rho_\lambda}(\delta), \ \forall j \in [p], S \subset [p]_j \right\},$$

$$(B.2) \qquad \mathcal{R}(\delta) = \left\{ \phi^2_\rho(\mathbf{X}_S, m_j(S)) \geq r_{\min}(\Sigma) > 0, \ \forall j \in [p], S \subset [p]_j \right\}.$$

Note that $\mathcal{E}(\delta, \lambda) = \bigcap_{j=1}^p \bigcap_{S \subset [p]_j} \mathcal{E}_S(\delta, \lambda; j)$, where $\mathcal{E}_S(\delta, \lambda; j)$ was defined in (8.14).

The structure of the proofs will be to show that on one or both of these events, the desired conclusions follow. Explicit bounds on the probabilities of these events are established in Section B.1.

B.1. **Uniform deviation bounds.** As discussed in Section 8.2.2, one of the main ingredients in proving Theorem 4.2 is a general result regarding deviation bounds for neighbourhood problems, given by Proposition B.1 below. The other ingredient is Proposition B.2 below, regarding the behavior of certain model selection exponents defined in analogy with (3.5):

$$\text{(B.3)} \qquad \psi_\lambda(\mathbf{X}, \sigma_{\max}^2; \delta) := \inf_{0 \le \sigma \le \sigma_{\max}} \Phi_\lambda(\mathbf{X}, 0, \sigma^2/\delta^2).$$

We often suppress the dependence on δ and write $\psi_\lambda(\mathbf{X}, \sigma_{\max}^2)$. Note that, in view of Lemma 8.6, $\psi_\lambda(\mathbf{X}, \sigma_{\max}^2)$ describes the conditional probability, given $\mathbf{X}$, that $(\sigma \mathbf{w}, \mathbf{X})$ violates a GW condition, where $\mathbf{w} \sim \mathcal{N}_n(0, I_n)$ is independent of $\mathbf{X}$. More precisely,

$$\sup_{0 \le \sigma \le \sigma_{\max}} \mathbb{P}\left[(\sigma\mathbf{w}, \mathbf{X}) \notin \mathrm{GW}_{\rho_\lambda}^\circ(\delta) \mid \mathbf{X} \right] = \sup_{0 \le \sigma \le \sigma_{\max}} \exp[-\Phi_\lambda(\mathbf{X}, 0, \sigma^2/\delta^2)]$$
$$= \exp[-\psi_\lambda(\mathbf{X}, \sigma_{\max}^2)].$$

We also recall the relation

$$\text{(B.4)} \qquad \xi = \xi(\delta) = \frac{1+\delta}{1-\delta}.$$

Proposition B.1. *Assume that Σ satisfies Condition 4.1 and ρ_λ satisfies Condition 4.2. Suppose $\mathbf{X} \overset{iid}{\sim} \mathcal{N}_p(0, \Sigma)$, $\delta \in (0,1)$, and define ξ by (B.4). Then there exist constants $c_0, c_1, c_2 > 0$ such that the following holds: If*

$$n > c_0 \frac{\sigma_{\max}^2 (1+\xi)^2}{r_{\min}(\Sigma)} d \log p,$$

then with probability at least $1 - c_1 \exp(-c_2 n) - p\binom{p}{d}\mathbb{E}\exp(-\psi_\lambda(\mathbf{X}, \sigma_{\max}^2; \delta))$,

$$\text{(B.5)} \qquad \|\widehat{\beta}_j(S) - \beta_j(S)\|_2 \le \frac{2\xi}{r_{\min}(\Sigma)} \rho_\lambda'(0+) \cdot \|\beta_j(S)\|_0^{1/2},$$

$$\text{(B.6)} \qquad \|\widehat{\beta}_j(S) - \beta_j(S)\|_1 \le \frac{2\xi(1+\xi)}{r_{\min}(\Sigma)} \rho_\lambda'(0+) \cdot \|\beta_j(S)\|_0,$$

uniformly over all $j \in [p]$ and $S \subset [p]_j$.

For future reference, inspection of the proof shows that the conclusion of Proposition B.1 holds on $\mathcal{E}(\delta, \lambda) \cap \mathcal{R}(\delta)$. For regularizers that satisfy the lower bound in Condition 4.2(c) we have the following control on the exponent $\psi_\lambda(\mathbf{X}, \sigma_{\max}^2)$:

Proposition B.2. *Assume that $\mathbf{X} \overset{iid}{\sim} \mathcal{N}_p(0, \Sigma)$, and that ρ_λ satisfies Condition 4.2(c). Then there exist constants $c > 0$ and $C = C(\underline{\rho}_1, \underline{\rho}_0)$ such that*

for any $\delta \in (0,1)$, if

(B.7)
$$\lambda \geq C\delta^{-1}\sigma_{\max}\|\Sigma\|^{1/4}\sqrt{\frac{(d+1)\log p}{n}}$$

then $\mathbb{E}\exp(-\psi_\lambda(\mathbf{X}, \sigma_{\max}^2; \delta)) \leq c\exp(-\min\{2(d+1)\log p, n\})$.

The proof of Proposition B.2 follows from an argument similar to that in Zhang and Zhang (2012) and is omitted for brevity. In order to prove Proposition B.1, we need the following two intermediate results, providing uniform control on RE constants and GW conditions. Recall $\mathcal{E}(\delta, \lambda)$ as defined in (B.1).

Proposition B.3 (Uniform GW control). *For any $\delta \in (0,1)$ and $\lambda > 0$,*

$$\mathbb{P}[\mathcal{E}(\delta, \lambda)] \geq 1 - p\binom{p}{d}\mathbb{E}\exp\left[-\psi_\lambda(\mathbf{X}, \sigma_{\max}^2; \delta)\right].$$

Proof. Fix $\delta \in (0,1)$. By analogy with (8.7), for any neighbourhood $S \subset [p]_j$, let

(B.8)
$$\xi_j(S) := \Phi_\lambda(\mathbf{X}_S, 0, \omega_j^2(S)/\delta^2) \geq \psi_\lambda(\mathbf{X}, \sigma_{\max}^2; \delta),$$

where the inequality follows from (B.3) and $\omega_j^2(S) \leq \sigma_{\max}^2$. We follow the proof of Proposition 8.5, but with $\beta_j(S)$ replaced with 0, and $\widetilde{\varepsilon}_j(S)$ replaced with $\widetilde{\varepsilon}_j(S)/\delta$. To simplify, let $\mathcal{F}_S^j = \mathcal{E}_S(\delta, \lambda; j)^c$ and $\mathcal{E} = \mathcal{E}(\delta, \lambda)$. By the comment following (8.14),

$$\mathcal{F}_S^j = \mathcal{A}\big(\widetilde{\varepsilon}_j(S)/\delta, \mathbf{X}, 0; S\big) = \mathcal{A}\big(\widetilde{\varepsilon}_j(S)/\delta, \mathbf{X}_S, 0\big)$$

where the second equality is by the same argument in the proof of Proposition 8.5. Since $\widetilde{\varepsilon}_j(S)/\delta \sim \mathcal{N}\big(0, [\omega_j^2(S)/\delta^2]I_n\big)$ independent of $\mathbf{X}_S$, we conclude, using Definition 3.4, that

$$\mathbb{P}\left(\mathcal{F}_S^j \mid \mathbf{X}_S\right) = \exp[-\xi_j(S)],$$

hence $\mathbb{P}(\mathcal{F}_S^j) \leq \mathbb{E}\exp[-\psi_\lambda(\mathbf{X}, \sigma_{\max}^2)]$, $\forall S \subset [p]_j$, using the inequality in (B.8). The events $\mathcal{F}_S^j$ are monotonic in S according to Corollary 8.4. (The division of $\varepsilon_j(S)$ by δ does not change anything in that proof.) It follows that

$$\mathcal{E}^c = \bigcup_{j=1}^{p}\bigcup_{S \subset [p]_j}\mathcal{F}_S^j \subset \bigcup_{j=1}^{p}\bigcup_{T \in m_j(\Sigma)}\mathcal{F}_{M_j(T)}^j.$$

Taking the union bound, and using $|m_j(\Sigma)| \leq \binom{p}{d}$ and

$$\mathbb{P}\left[\mathcal{F}_{M_j(T)}^j\right] \leq \mathbb{E}\exp[-\psi_\lambda(\mathbf{X}, \sigma_{\max}^2)], \quad \forall T \in m_j(\Sigma),$$

finishes the proof. $\qquad\square$

Proposition B.4 (Uniform RE control). *Assume* $\mathbf{X} \overset{iid}{\sim} \mathcal{N}_p(0, \Sigma)$, Σ *satisfies Condition 4.1 and ρ_λ satisfies Condition 4.2. There exist universal constants $c_0, c_1, c_2 > 0$, such that if*

$$n > c_0 \, \frac{\sigma_{\max}^2 (1 + \xi)^2}{r_{\min}(\Sigma)} d(\Sigma) \log p$$

then with probability at least $1 - c_1 \exp(-c_2 n)$,

$$\inf_{1 \leq j \leq p} \inf_{S \subset [p]_j} \inf_{\substack{A \subset S \\ |A| \leq d}} \phi_\rho^2(\mathbf{X}_S, A; \xi) \geq r_{\min}(\Sigma).$$

The proof of this proposition appears in Section B.2 below. Recalling the definition of $\mathcal{R}(\delta)$ in (B.2), combined with $m_j(S) = \|\beta_j(S)\|_0 \leq d$ (cf. Definition 3.5), Proposition B.4 implies that $\mathcal{R}(\delta)$ holds with probability at least $1 - c_1 \exp(-c_2 n)$. Let us show how Proposition B.1 follows.

Proof of Proposition B.1. Recall the definitions of $\mathcal{E}(\delta, \lambda)$ in (B.1) and $\mathcal{R}(\delta)$ in (B.2). Propositions B.3 and B.4 guarantee that

$$\mathbb{P}\left(\mathcal{R}(\delta) \cap \mathcal{E}(\delta, \lambda)\right) \geq 1 - c_1 \exp(-c_2 n) - p \binom{p}{d} \mathbb{E} \exp(-\psi_\lambda(\mathbf{X}, \sigma_{\max}^2; \delta)).$$

Thus, it suffices to deduce (B.5) and (B.6) whenever we are on the event $\mathcal{R}(\delta) \cap \mathcal{E}(\delta, \lambda)$. The case $\beta_j(S) = 0$ follows from (8.14) and Lemma 8.6, and the case $\beta_j(S) \neq 0$ follows from Theorem 8.1 applied to the corresponding neighbourhood regression problems. □

B.2. **Uniform control of restricted eigenvalues.** In this section, we collect the necessary results that lead to the proof of Proposition B.4. We begin with a definition which generalizes the familiar (ℓ_1) restricted eigenvalue (Bickel et al., 2009; Raskutti et al., 2011):

Definition B.1. $Z \in \mathbb{R}^{n \times m}$ satisfies a *generalized restricted eigenvalue condition of order k w.r.t.* ρ_λ with parameters $\alpha, \xi > 0$, denoted as $Z \in \text{RE}_{\rho_\lambda}(k, \alpha; \xi)$, if

$$\frac{1}{n}\|Zu\|_2^2 \geq \alpha^2 \|u\|_2^2 \quad \forall u \in C_{\rho_\lambda}(A, \xi),$$

uniformly for all $A \subset [m]$ with $|A| = k$. Equivalently, recalling Definition 8.4,

$$Z \in \text{RE}_{\rho_\lambda}(k, \alpha; \xi) \iff \inf_{\substack{A \subset [m] \\ |A| = k}} \phi_{\rho_\lambda}^2(Z, A, \xi) \geq \alpha^2.$$

In the sequel, we will suppress the dependence of various quantities on λ, ξ and m, when no confusion arises. For example, we will write $\rho_\lambda = \rho$, $\text{RE}_\rho(k, \alpha) = \text{RE}_{\rho_\lambda}(k, \alpha; \xi)$ or $C_\rho(A) = C_{\rho_\lambda}^m(A, \xi)$. The following lemma collects some simple consequences of these definitions.

Lemma B.5. *The following hold:*

(a) *When ρ is nondecreasing,*
  − $A' \subset A \implies C_\rho(A') \subset C_\rho(A)$,

$$- A' \subset A \implies \phi_\rho^2(Z, A') \geq \phi_\rho^2(Z, A),$$
$$- Z \in \mathrm{RE}_\rho(k, \alpha) \implies \phi_\rho^2(Z, A) \geq \alpha^2, \ \forall A : |A| \leq k,$$
$$- k' \leq k \implies \mathrm{RE}_\rho(k, \alpha) \subset \mathrm{RE}_\rho(k', \alpha).$$

(b) $Z \in \mathrm{RE}_\rho(k, \alpha) \implies Z_S \in \mathrm{RE}_\rho(k \wedge |S|, \alpha)$.

The next result shows that we can control the generalized RE constants for Z_S uniformly by a suitable generalized RE constant for Z:

Lemma B.6. *If $Z \in \mathrm{RE}_\rho(d, \alpha)$, then*

$$\text{(B.9)} \qquad \inf_{1 \leq j \leq m} \ \inf_{S \subset [m]_j} \ \inf_{\substack{A \subset S \\ |A| \leq d}} \phi_\rho^2(Z_S, A) \geq \alpha^2.$$

Proof. Fix j, $S \subset [m]_j$ and $A \subset S$ with $|A| \leq d$. By Lemma B.5(b), the assumption implies $Z_S \in \mathrm{RE}_\rho(d \wedge |S|, \alpha)$. Then, the last assertion in Lemma B.5(a) implies $Z_S \in \mathrm{RE}_\rho(|A|, \alpha)$, hence $\phi_\rho^2(Z_S, A) \geq \alpha^2$. Since the lower bound does not depend on A, S or j, we get the desired result. $\square$

Next, we show that we can control RE constants for ρ by those for the ℓ_1 norm. Let us write $C_1^m(A, \xi)$ for the cone corresponding to $\rho = \|\cdot\|_1$, and similarly for the RE constants. We have the following lemma:

Lemma B.7. *Under Condition 4.2 on ρ,*

$$C_\rho^m(A, \xi) \subset \bigcup_{\substack{A' \subset [m] \\ |A'| = |A|}} C_1^m(A', \xi).$$

Proof of Lemma B.7. Fix nonzero $u \in C_\rho(A, \xi)$ and assume, without loss of generality that $|u_i| > 0$ for all $i \in [m]$; otherwise, we can inflate all the zero entries by $\epsilon > 0$, change ξ to $\xi + \rho(\epsilon)|A^c|/\rho(u_A)$, and let $\epsilon \to 0$ at the end.

Let $M = M(u)$ be the index set of the $|A|$ largest $|u_i|$. Then $\rho(u_{A^c}) \leq \xi \rho(u_A)$ implies $\rho(u_{M^c}) \leq \xi \rho(u_M)$ since ρ is nondecreasing. We note that $M = \{i : |u_i| \geq \tau\}$, for some $\tau > 0$, assuming $M^c \neq \emptyset$ without loss of generality. As a consequence of Condition 4.2, $x \mapsto \rho_\lambda(x)/x$ is nonincreasing. Then,

$$\tau \rho(|u_i|) \leq |u_i| \rho(\tau), \ i \in M,$$

with the reverse inequality for $i \in M^c$. Summing over M and M^c and rearranging, we have

$$\frac{\rho(u_M)}{\|u_M\|_1} \leq \frac{\rho(\tau)}{\tau} \leq \frac{\rho(u_{M^c})}{\|u_{M^c}\|_1},$$

or $\|u_{M^c}\|_1 / \|u_M\| \leq \rho(u_{M^c})/\rho(u_M) \leq \xi$. Hence, $u \in C_1^m(M(u), \xi)$ where $|M(u)| = |A|$, which gives the desired result. $\square$

As a consequence of Lemma B.7, $Z \in \mathrm{RE}_1(|A|, \alpha)$ implies $\phi_\rho^2(Z, A) \geq \alpha^2$, from which we get:

Lemma B.8. *Under Condition 4.2, $\mathrm{RE}_1(d, \alpha) \subset \mathrm{RE}_\rho(d, \alpha)$.*

In particular, the conclusion of Lemma B.6 holds under the (stronger) assumption $Z \in \mathrm{RE}_1(d, \alpha)$. In other words, to obtain uniform control over generalized restricted eigenvalues for all possible neighbourhood regression problems, it suffices to show that $\mathbf{X} \in \mathrm{RE}_1(d, \alpha)$ for some constant $\alpha > 0$. This is guaranteed by the following lemma, which is essentially a restatement of Corollary 1 in Raskutti et al. (2010):

Lemma B.9. *Assume* $\mathbf{X} \overset{iid}{\sim} \mathcal{N}_p(0, \Sigma)$ *for some* Σ *satisfying Condition 4.1. There exist universal constants* $c_0, c_1, c_2 > 0$, *such that if*

$$n > c_0 \frac{\sigma_{\max}^2 (1 + \xi)^2}{r_{\min}(\Sigma)} d(\Sigma) \log p$$

then with probability at least $1 - c_1 \exp(-c_2 n)$,

$$\mathbf{X} \in \mathrm{RE}_1\big(d(\Sigma), \sqrt{r_{\min}(\Sigma)}; \xi\big).$$

Proposition B.4 now follows as a straightforward consequence of Lemmas B.6, B.8 and B.9.

Appendix C. Auxiliary results for score-based learning

This section provides some additional results which are needed to prove Theorems 5.2, 5.3, and 5.4, which are also proved in this section.

For any $\delta \in (0, 1)$, $\lambda \geq 0$, $\delta_0 > 0$, and $\pi \in \mathbb{S}_p$, define the following event:

$$(\text{C.1}) \qquad \mathcal{G}(\delta_0, \lambda; \pi) = \left\{ \frac{1}{2n} \|\mathbf{E}(\pi)\|_2^2 - \frac{1}{2n} \|\mathbf{E}(\widehat{\pi})\|_2^2 \leq \delta_0 \rho_\lambda(\widetilde{B}(\pi)) \right\}.$$

As in Appendix B, the idea will be to show that on this event (along with (B.1) and (B.2)), the desired conclusions hold. In Appendix C.6, we provide an explicit bound on the probability of $\mathcal{G}(\delta_0, \lambda; \pi)$.

C.1. **Some intermediate lemmas.** Recall the definitions of $\mathcal{E}(\delta, \lambda)$ and $\mathcal{R}(\delta)$ in (B.1)–(B.2). We start with the following extension of GW bounds:

Lemma C.1. *Let* $\widehat{\Delta} := \widehat{B} - \widetilde{B}(\widehat{\pi})$. *On* $\mathcal{E}(\delta, \lambda)$, *we have*

$$(\text{C.2}) \qquad \frac{1}{n} \left| \mathrm{tr}\left(\mathbf{E}(\widehat{\pi})^T \mathbf{X} \widehat{\Delta} \right) \right| < \delta \left[\frac{1}{2n} \|\mathbf{X} \widehat{\Delta}\|_2^2 + \rho_\lambda(\widehat{\Delta}) \right].$$

Proof. Let $\widehat{\Delta}_j := \widehat{\beta}_j - \widetilde{\beta}_j(\widehat{\pi})$ be the jth column of $\widehat{\Delta}$. Then

$$(\text{C.3}) \qquad \frac{1}{n} \left| \mathrm{tr}\left(\mathbf{E}(\widehat{\pi})^T \mathbf{X} \widehat{\Delta} \right) \right| \leq \frac{1}{n} \sum_{j=1}^{p} |\langle \widetilde{\varepsilon}_j(\widehat{\pi}), \mathbf{X} \widehat{\Delta}_j \rangle|.$$

According to (B.1), on $\mathcal{E}(\delta, \lambda)$, we have $(\widetilde{\varepsilon}_j(S), \mathbf{X}_S) \in \mathrm{GW}_{\rho_\lambda}^\circ(\delta)$ for all $S \subset [p]_j$. In particular, applying with $S = S_j(\widehat{\pi})$ and using $u = \widehat{\Delta}_j$ in the Definition 8.3 of GW, we have

$$\frac{1}{n} |\langle \widetilde{\varepsilon}_j(\widehat{\pi}), \mathbf{X} \widehat{\Delta}_j \rangle| < \delta \left[\frac{1}{2n} \|\mathbf{X} \widehat{\Delta}_j\|_2^2 + \rho_\lambda(\widehat{\Delta}_j) \right], \quad \forall j$$

Summing over j and plugging into (C.3) yields (C.2).   □

For any matrix $A = (a_{ij}) \in \mathbb{R}^{p \times p}$ and $S \subset [p] \times [p]$, let $A_{\langle S \rangle}$ denote the $p \times p$ matrix formed by zero-ing the elements outside of S, i.e.

$$(A_{\langle S \rangle})_{ij} = \begin{cases} a_{ij}, & (i,j) \in S, \\ 0, & (i,j) \notin S. \end{cases}$$

In analogy with Condition 5.1 on signal strength, let us define

$$(\text{C.4}) \qquad \tau_\lambda(\alpha; \Sigma) := \inf \left\{ \tau : \ \frac{\rho_\lambda'(0+)^2}{\rho_\lambda(\tau)} \leq \frac{r_{\min}(\Sigma)}{\alpha} \right\}$$

where we often suppress the dependence on Σ. Note that we can write Condition 5.1 equivalently as $\tau_* \geq \tau_\lambda(a_1)$.

Lemma C.2. *Assume that ρ_λ satisfies Condition 4.2 and*

$$(\text{C.5}) \qquad \tau_* \geq \tau_\lambda \Big(\frac{2\xi}{1 - \delta_1} \Big), \quad \text{for some } \delta_1 \in (0,1)$$

where $\xi = \xi(\delta)$ is defined by (B.4). Then, on $\mathcal{R}(\delta) \cap \mathcal{E}(\delta, \lambda)$,

$$(\text{C.6}) \qquad \rho_\lambda(\widehat{B}) \ \geq \ \delta_1 \rho_\lambda(\widetilde{B}(\widehat{\pi})) + \rho_\lambda \Big((\widehat{B} - \widetilde{B}(\widehat{\pi}))_{\langle \operatorname{supp}(\widetilde{B}(\widehat{\pi}))^c \rangle} \Big).$$

Proof. To lighten the notation, let $\Delta = \widehat{B} - \widetilde{B}(\widehat{\pi})$, $S_1 = \operatorname{supp}(\widetilde{B}(\widehat{\pi}))$, $\Delta_1 = \Delta_{\langle S_1 \rangle}$, and $\Delta_2 = \Delta_{\langle S_1^c \rangle}$. We have

$$(\text{C.7}) \qquad \rho_\lambda(\Delta_1) \leq \rho_\lambda'(0+) \|\Delta_1\|_1 \leq \rho_\lambda'(0+) \|\widetilde{B}(\widehat{\pi})\|_0^{1/2} \|\Delta_1\|_2.$$

Since we are on $\mathcal{R}(\delta) \cap \mathcal{E}(\delta, \lambda)$, Proposition B.1 yields the ℓ_2 deviation bound (B.5), which we use with $S = S_j(\widehat{\pi})$. Plugging into (C.7) and using $\|\Delta_1\|_2 \leq \|\Delta\|_2$,

$$(\text{C.8}) \qquad \rho_\lambda(\Delta_1) \leq \big[\rho_\lambda'(0+) \big]^2 \frac{2\xi}{r_{\min}(\Sigma)} \|\widetilde{B}(\widehat{\pi})\|_0.$$

Trivially, we have $\rho_\lambda(\widetilde{B}(\widehat{\pi})) \geq \rho_\lambda(\tau_*) \|\widetilde{B}(\widehat{\pi})\|_0$, so that by (C.8)

$$(\text{C.9}) \qquad \rho_\lambda(\Delta_1) \leq \Big[\frac{\rho_\lambda'(0+)^2}{\rho_\lambda(\tau_*)} \frac{2\xi}{r_{\min}(\Sigma)} \Big] \rho_\lambda(\widetilde{B}(\widehat{\pi})) \ \leq \ (1 - \delta_1) \rho_\lambda(\widetilde{B}(\widehat{\pi})),$$

where the last inequality follows from (C.5). Finally, note that

$$\rho_\lambda(\widehat{B}) \ \geq \ \rho_\lambda(\widetilde{B}(\widehat{\pi})) + \rho_\lambda(\Delta_2) - \rho_\lambda(\Delta_1)$$
$$\geq \ \delta_1 \rho_\lambda(\widetilde{B}(\widehat{\pi})) + \rho_\lambda(\Delta_2).$$

where the first inequality is by arguments similar to those leading to (A.4) and the second is by (C.9).   □

Remark C.1. van de Geer and Bühlmann (2013) use a slightly weaker beta-min condition in which only a constant fraction of the edges of each DAG are assumed to be sufficiently large. Lemma C.2 and the ensuing arguments

carry through under such an assumption: Under Condition 3.5 in van de Geer and Bühlmann (2013), we can use

$$\rho_\lambda(\widetilde{B}(\widehat{\pi})) \geq (1 - \eta_1)\rho_\lambda(\tau_*)\|\widetilde{B}(\widehat{\pi})\|_0,$$

between (C.8) and (C.9) and obtain a bound similar to (C.6), with only the constants modified.

The conclusion of Lemma C.2 is stronger than what we need in the sequel. We only use the weaker inequality $\rho_\lambda(\widehat{B}) \geq \delta_1\rho_\lambda(\widetilde{B}(\widehat{\pi}))$ implied by (C.6).

C.2. **A sparsity bound.** Proposition C.3 below is the main ingredient used in the proof of Theorem 5.2. This result provides an explicit, nonasymptotic relationship between the "weak" sparsity of $\{\widehat{B}, \widetilde{B}(\widehat{\pi}), \widetilde{B}(\pi_0)\}$ as measured by the regularizer ρ_λ.

Proposition C.3. *Assume $n > 8\,(d + 1)\log p$. Under Condition 4.2 on ρ_λ, further assume*

$$\tau_* \geq \tau_\lambda\Big(\frac{2(1 + \delta)}{1 - 3\delta}\Big) \quad \text{for some } \delta \in (0, 1/3).$$

Then, given π_0 satisfying Condition 5.2 for some $a_2 > 0$, we have

$$(C.10) \qquad \frac{2\delta}{1 - \delta}\rho_\lambda(\widetilde{B}(\widehat{\pi})) \overset{(i)}{\leq} \rho_\lambda(\widehat{B}) \overset{(ii)}{\leq} \frac{2}{1 - \delta}\Big(1 + \frac{10}{a_2}\Big)\rho_\lambda(\widetilde{B}(\pi_0)),$$

with probability at least $1 - c_1 e^{-c_2 \min\{n,\,(d+1)\log p\}} - p\binom{p}{d}\mathbb{E}e^{-\psi_\lambda(\mathbf{X}, \sigma_{\max}^2; \delta)}$.

Proof. Recall the definition of $\mathcal{G}(\delta_0, \lambda; \pi)$ in (C.1). Fix some π_0 satisfying Condition 5.2 with $a_2 > 0$. Taking (arbitrarily) $C = 1$ and $\delta_0 = 10/a_2$ in Proposition C.5, we have

$$\mathbb{P}\left[\mathcal{G}(\delta_0, \lambda; \pi_0)^c\right] \leq 2e^{-(d+1)\log p}.$$

Combined with Propositions B.4 and B.3, we obtain

$$\mathbb{P}\left(\mathcal{G}(\delta_0, \lambda; \pi_0) \cap \mathcal{E}(\delta, \lambda) \cap \mathcal{R}(\delta)\right)$$

$$\geq 1 - c_1 \exp(-c_2 \min\{n, (d + 1)\log p\}) - p\binom{p}{d}\mathbb{E}\exp(-\psi_\lambda(\mathbf{X}, \sigma_{\max}^2; \delta)).$$

Thus, we may assume we are on $\mathcal{G}(\delta_0, \lambda; \pi_0) \cap \mathcal{E}(\delta, \lambda) \cap \mathcal{R}(\delta)$. Since we are on $\mathcal{E}(\delta, \lambda)$, we can combine Lemma C.1 with Lemma 8.7 (applied with $\pi = \pi_0$) to deduce (recall $\widehat{\Delta} := \widehat{B} - \widetilde{B}(\widehat{\pi})$)

$$\frac{1}{2n}\|\mathbf{X}\widehat{\Delta}\|_2^2 + \rho_\lambda(\widehat{B}) \leq \frac{\delta}{2n}\|\mathbf{X}\widehat{\Delta}\|_2^2 + \delta\rho_\lambda(\widehat{\Delta})$$
$$+ \frac{1}{2n}\|\mathbf{E}(\pi_0)\|_2^2 - \frac{1}{2n}\|\mathbf{E}(\widehat{\pi})\|_2^2 + \rho_\lambda(\widetilde{B}(\pi_0)).$$

Dropping the prediction loss terms (those involving $\|\mathbf{X}\widehat{\Delta}\|_2^2$), and using that we are on $\mathcal{G}(\delta_0, \lambda; \pi_0)$ to bound $\frac{1}{2n}\|\mathbf{E}(\pi_0)\|_2^2 - \frac{1}{2n}\|\mathbf{E}(\widehat{\pi})\|_2^2$, we have after

rearranging,

$$(C.11) \qquad \rho_\lambda(\widehat{B}) \leq (1+\delta_0)\rho_\lambda(\widetilde{B}(\pi_0)) + \delta\rho_\lambda(\widehat{B} - \widetilde{B}(\widehat{\pi}))$$
$$\leq (1+\delta_0)\rho_\lambda(\widetilde{B}(\pi_0)) + \delta\rho_\lambda(\widetilde{B}(\widehat{\pi})) + \delta\rho_\lambda(\widehat{B}).$$

Let $\delta_1 = 2\delta/(1-\delta)$, so that $\xi/(1-\delta_1) = (1+\delta)/(1-3\delta)$ (cf. (B.4)). Furthermore, since $\delta < 1/3$ by assumption, $\delta_1 < 1$, so that Lemma C.2 implies $\rho_\lambda(\widehat{B}) \geq \delta_1\rho_\lambda(\widetilde{B}(\widehat{\pi}))$ which gives (i) in (C.10).

Since $\rho_\lambda(\widetilde{B}(\widehat{\pi})) \leq (1/\delta_1)\rho_\lambda(\widehat{B})$, the bounds in (C.11) imply that

$$\rho_\lambda(\widehat{B}) \leq (1+\delta_0)\rho_\lambda(\widetilde{B}(\pi_0)) + \frac{\delta}{\delta_1}\rho_\lambda(\widehat{B}) + \delta\rho_\lambda(\widehat{B}).$$

Rearranging we get

$$\rho_\lambda(\widehat{B}) \leq \left[1 - \delta(1+\delta_1)/\delta_1\right]^{-1}(1+\delta_0)\rho_\lambda(\widetilde{B}(\pi_0)).$$

We have $[1 - \delta(1+\delta_1)/\delta_1]^{-1}(1+\delta_0) = \frac{2}{1-\delta}(1+\frac{10}{a_2})$, using $\delta_0 = 10/a_2$ and $\delta_1 = 2\delta/(1-\delta)$ as before. This proves (ii) in (C.10). $\qquad\square$

C.3. **Proof of Theorem 5.2.** For regularizers satisfying Condition 4.2, the desired bound follows by taking $\delta = (a_1-2)/(3a_1+2) \in (0,1/3)$ in Proposition C.3, and using Proposition B.2 to complete the probability bound.

To deduce the ℓ_0 bound, consider the case where ρ_λ is also ℓ_0-compatible. Condition 5.1 implies

$$\rho_\lambda(\widetilde{B}(\widehat{\pi})) \geq a_1 \frac{\rho_\lambda'(0+)^2}{r_{\min}(\Sigma)}\|\widetilde{B}(\widehat{\pi})\|_0,$$

while on the other hand, ℓ_0-compatibility (Definition 4.1) gives $\rho_\lambda(\widetilde{B}(\pi_0)) \leq \overline{\rho}_0\lambda^2\|\widetilde{B}(\pi_0)\|_0$. Combining these with (C.10) yields

$$a_1\frac{\rho_\lambda'(0+)^2}{r_{\min}(\Sigma)}\|\widetilde{B}(\widehat{\pi})\|_0 \leq \frac{1}{\delta}\Big(1+\frac{10}{a_2}\Big)\overline{\rho}_0\,\lambda^2\|\widetilde{B}(\pi_0)\|_0$$
$$\implies \|\widetilde{B}(\widehat{\pi})\|_0 \leq \kappa_3 \cdot \Big[\frac{\lambda^2}{\rho_\lambda'(0+)^2}\Big]\|\widetilde{B}(\pi_0)\|_0,$$

as desired. Here, $\kappa_3 = \kappa_3(\Sigma) = \frac{1}{a_1}\big(\frac{3a_1+2}{a_1-2}\big)\big(1+\frac{10}{a_2}\big)\overline{\rho}_0 r_{\min}(\Sigma)$, using our earlier choice of δ.

C.4. **Proof of Theorem 5.3.** Assume that we are on the event $\mathcal{R}(\delta) \cap \mathcal{E}(\delta,\lambda)$. As in the proof of Theorem 5.2, choose $\delta = (a_1-2)/(3a_1+2) \in (0,1/3)$. Applying (C.2) to Lemma 8.7 yields

$$(C.12) \quad \frac{1}{2n}\|\mathbf{E}(\widehat{\pi})\|_2^2 + \rho_\lambda(\widehat{B}) \leq \frac{1}{2n}\|\mathbf{E}(\pi)\|_2^2 + \rho_\lambda(\widetilde{B}(\pi)) + \delta\rho_\lambda(\widehat{B} - \widetilde{B}(\widehat{\pi}))$$

for any $\pi \in \mathbb{S}_p$. Lemma C.1 combined with and Propositions B.2 and B.3 imply that (C.12) holds with probability at least $1-c_1\exp(-c_2(d+1)\log p)$.

Invoking the subadditivity of ρ_λ and re-arranging,

(C.13)
$$\frac{1}{2n}\|\mathbf{E}(\widehat{\pi})\|_2^2 + (1-\delta)\rho_\lambda(\widehat{B}) \leq \frac{1}{2n}\|\mathbf{E}(\pi)\|_2^2 + \rho_\lambda(\widetilde{B}(\pi)) + \delta\rho_\lambda(\widetilde{B}(\widehat{\pi})).$$

Now apply Lemma C.2 with $\delta_1 = 2\delta/(1-\delta)$ to deduce that

(C.14)
$$\frac{1}{2n}\|\mathbf{E}(\widehat{\pi})\|_2^2 + \delta\rho_\lambda(\widetilde{B}(\widehat{\pi})) \leq \frac{1}{2n}\|\mathbf{E}(\pi)\|_2^2 + \rho_\lambda(\widetilde{B}(\pi)).$$

Taking $u = \sqrt{4(d+1)\log p}$ in Lemma C.7 and observing that $1 - h_n(u) > 1/2 > \delta$ (cf. (C.21) for the definitions of $h_n(u)$, $H_n(u)$) for $n > 5(d+1)\log p$, we have

$$\frac{1}{2}\operatorname{tr}\widetilde{\Omega}(\widehat{\pi}) + \rho_\lambda(\widetilde{B}(\widehat{\pi})) \leq \frac{1 + H_n(u)}{\delta} \cdot \left[\frac{1}{2}\operatorname{tr}\widetilde{\Omega}(\pi) + \rho_\lambda(\widetilde{B}(\pi))\right]$$

$$\leq \frac{1}{\delta}\left(1 + 6\sqrt{\frac{(d+1)\log p}{n}}\right) \cdot \left[\frac{1}{2}\operatorname{tr}\widetilde{\Omega}(\pi) + \rho_\lambda(\widetilde{B}(\pi))\right].$$

Combined with the probability bound implied by Propositions B.2 and B.3, this holds with probability at least $1-c_1\exp(-c_2(d+1)\log p)$ (c_1, c_2 possibly different from above). Since π was arbitrary, substituting $\delta = (a_1-2)/(3a_1+2)$ completes the proof.

C.5. **Proof of Theorem 5.4.** The desired result follows from combining Theorem 5.1 with the following (slightly) more general result:

Proposition C.4. *Assume that ρ_λ is ℓ_0-compatible and there exists $a_3 > 10$ such that*

(C.15)
$$\frac{\operatorname{tr}\widetilde{\Omega}(\pi_0)}{\operatorname{tr}\widetilde{\Omega}(\pi)} \leq 1 - a_3\sqrt{\frac{(d+1)\log p}{n}} \quad \forall \pi \in \mathbb{S}_p, \quad and$$

(C.16)
$$d \leq \frac{a_3 - 10}{2(1+\delta)\overline{p}_0}\sqrt{\frac{(d+1)\log p}{n}} \cdot \frac{r_{\min}(\Sigma)}{\lambda^2}.$$

If

(C.17)
$$\lambda \geq C\delta^{-1}\sigma_{\max}\|\Sigma\|^{1/4}\sqrt{\frac{(d+1)\log p}{n}}$$

then with probability at least $1 - c_1\exp(-c_2(d+1)\log p)$, it holds that $\widehat{\pi} \in \arg\min_\pi \operatorname{tr}\widetilde{\Omega}(\pi)$.

Proof. Define $\mathbb{S}_0 := \arg\min_\pi \operatorname{tr}\widetilde{\Omega}(\pi)$ and let $\pi_0 \in \mathbb{S}_0$ be arbitrary. Assuming that we are on $\mathcal{E}(\delta, \lambda)$, we can combine Lemma C.1 with Lemma 8.7 (applied with $\pi = \pi_0$) to deduce (recall $\widehat{\Delta} := \widehat{B} - \widetilde{B}(\widehat{\pi})$)

$$\frac{1}{2n}\|\mathbf{X}\widehat{\Delta}\|_2^2 + \rho_\lambda(\widehat{B}) \leq \frac{\delta}{2n}\|\mathbf{X}\widehat{\Delta}\|_2^2 + \delta\rho_\lambda(\widehat{\Delta})$$
$$+ \frac{1}{2n}\|\mathbf{E}(\pi_0)\|_2^2 - \frac{1}{2n}\|\mathbf{E}(\widehat{\pi})\|_2^2 + \rho_\lambda(\widetilde{B}(\pi_0)).$$

This implies, in particular, that

$$(C.18) \qquad \rho_\lambda(\widehat{B}) + \frac{1}{2n}\|\mathbf{E}(\widehat{\pi})\|_2^2 - \frac{1}{2n}\|\mathbf{E}(\pi_0)\|_2^2 \le \delta\rho_\lambda(\widehat{\Delta}) + \rho_\lambda(\widetilde{B}(\pi_0)).$$

Suppose $\widehat{\pi} \notin \mathbb{S}_0$, so that $\operatorname{tr}\widetilde{\Omega}(\widehat{\pi}) > \operatorname{tr}\widetilde{\Omega}(\pi_0)$. Then Lemma C.7 combined with (C.15) implies that for any $0 < u < n/\sqrt{n+1}$,

$$\frac{1}{2n}\|\mathbf{E}(\widehat{\pi})\|_2^2 - \frac{1}{2n}\|\mathbf{E}(\pi_0)\|_2^2$$

$$\ge \frac{1}{2}\operatorname{tr}\widetilde{\Omega}(\widehat{\pi})\Big[1 - h_n(u)\Big] - \frac{1}{2}\operatorname{tr}\widetilde{\Omega}(\pi_0)\Big[1 + H_n(u)\Big]$$

$$= \left[\frac{1}{2}\operatorname{tr}\widetilde{\Omega}(\widehat{\pi}) - \frac{1}{2}\operatorname{tr}\widetilde{\Omega}(\pi_0)\right] - \frac{1}{2}\operatorname{tr}\widetilde{\Omega}(\widehat{\pi})\, h_n(u) - \frac{1}{2}\operatorname{tr}\widetilde{\Omega}(\pi_0)\, H_n(u)$$

$$> \left[\frac{1}{2}\operatorname{tr}\widetilde{\Omega}(\widehat{\pi}) - \frac{1}{2}\operatorname{tr}\widetilde{\Omega}(\pi_0)\right] - \frac{1}{2}\operatorname{tr}\widetilde{\Omega}(\widehat{\pi})\Big(h_n(u) + H_n(u)\Big)$$

$$> \frac{a_3 - 10}{2}\sqrt{\frac{(d+1)\log p}{n}}\operatorname{tr}\widetilde{\Omega}(\widehat{\pi})$$

$$(C.19) \qquad \ge \frac{a_3 - 10}{2}\sqrt{\frac{(d+1)\log p}{n}}p\, r_{\min}(\Sigma).$$

Note that here we have used $h_n(u) + H_n(u) \le 5u/\sqrt{n}$ and the (arbitrary) choice of $u = \sqrt{4(d+1)\log p}$.

Combining (C.18) and (C.19), we have

$$\rho_\lambda(\widehat{B}) + \frac{a_3 - 10}{2}\sqrt{\frac{(d+1)\log p}{n}}p\, r_{\min}(\Sigma) < \delta\rho_\lambda(\widehat{\Delta}) + \rho_\lambda(\widetilde{B}(\pi_0))$$

$$\le \delta\rho_\lambda(\widehat{B}) + \delta\rho_\lambda(\widetilde{B}(\widehat{\pi})) + \rho_\lambda(\widetilde{B}(\pi_0)).$$

Thus

$$\frac{a_3 - 10}{2}\sqrt{\frac{(d+1)\log p}{n}}p\, r_{\min}(\Sigma) < \delta\rho_\lambda(\widetilde{B}(\widehat{\pi})) + \rho_\lambda(\widetilde{B}(\pi_0)) \le (1+\delta)\overline{\rho}_0\lambda^2\, dp$$

$$\implies \frac{a_3 - 10}{2}\sqrt{\frac{(d+1)\log p}{n}} \cdot r_{\min}(\Sigma) < (1+\delta)\overline{\rho}_0\lambda^2\, d.$$

In the second inequality we have invoked ℓ_0-compatibility of ρ_λ (Definition 4.1). But this contradicts (C.16), whence $\widehat{\pi} \in \mathbb{S}_0$, as desired. The probability bound on $\mathbb{P}[\mathcal{E}(\delta, \lambda)]$ follows by combining Propositions B.2 and B.3 along with (C.17). $\qquad\square$

C.6. **A bound on the sample residuals.** In this section, we prove the following result, which is used in the proof of Proposition C.3:

Proposition C.5. *Assume $n > 4(C+1)(d+1)\log p$ for some $C > 0$ and let π_0 be a minimum-trace permutation such that*

$$(\text{C.20}) \qquad \frac{\rho_\lambda(\widetilde{B}(\pi_0))}{\operatorname{tr}\widetilde{\Omega}(\pi_0)} \geq \frac{1}{\delta_0}\sqrt{\frac{50(C+1)(d+1)\log p}{n}}.$$

Then for any $\delta_0 > 0$, $\mathbb{P}(\mathcal{G}(\delta_0, \lambda; \pi_0)) \geq 1 - 2e^{-C(d+1)\log p}$, i.e.

$$\mathbb{P}\left(\frac{1}{2n}\|\mathbf{E}(\pi_0)\|_2^2 - \frac{1}{2n}\|\mathbf{E}(\widehat{\pi})\|_2^2 > \delta_0\rho_\lambda(\widetilde{B}(\pi_0))\right) \leq 2e^{-C(d+1)\log p}.$$

Define two functions by

$$(\text{C.21}) \qquad h_n(u) := -\frac{u^2}{n} + \frac{2u}{\sqrt{n+1}} + \frac{1}{n+1}, \quad H_n(u) := \frac{u^2}{n} + \frac{2u}{\sqrt{n}}.$$

These functions bound the deviations in the normed residuals $\widetilde{\varepsilon}_j(\pi)$, and will be used repeatedly in the sequel. We note that

$$(\text{C.22}) \qquad H_n(u) + h_n(u) \leq \frac{5u}{\sqrt{n}}, \quad u \geq n^{-1/2}.$$

Lemma C.6. *Suppose $\mathbf{w} \sim \mathcal{N}_n(0, \sigma^2 I_n)$. Then for any $0 < u < n/\sqrt{n+1}$,*

$$(\text{C.23}) \qquad \sigma^2\left(1 - h_n(u)\right) \leq \frac{1}{n}\|\mathbf{w}\|_2^2 \leq \sigma^2\left(1 + H_n(u)\right)$$

with probability at least $1 - 2e^{-u^2/2}$.

Proof. For $z \sim \mathcal{N}_n(0, I_n)$, we have the following useful bounds (see, e.g., Gordon, 1988, Corollary 1.2):

$$\frac{n}{\sqrt{n+1}} \leq \mathbb{E}\|z\|_2 = \sqrt{2}\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \leq \sqrt{n}.$$

Gaussian concentration implies that for any $u > 0$, both

$$\left\{\|\mathbf{w}\|_2 \leq \sigma(n/\sqrt{n+1} - u)\right\}, \quad \text{and} \quad \left\{\|\mathbf{w}\|_2 \geq \sigma(\sqrt{n} + u)\right\}$$

hold with probability at most $e^{-u^2/2}$. Thus,

$$(\text{C.24}) \quad \mathbb{P}\left(\sigma^2\left(\frac{n}{\sqrt{n+1}} - u\right)^2 \leq \|\mathbf{w}\|_2^2 \leq \sigma^2\left(\sqrt{n} + u\right)^2\right) \geq 1 - 2e^{-u^2/2}.$$

Re-writing (C.24) using (C.21) yields the desired result. $\qquad\square$

Lemma C.7. *Suppose $\mathbf{X} \overset{iid}{\sim} \mathcal{N}_p(0, \Sigma)$. Then for any $\pi \in \mathbb{S}_p$ and $0 < u < n/\sqrt{n+1}$,*

$$(\text{C.25}) \qquad \frac{1}{2}\operatorname{tr}\widetilde{\Omega}(\pi)\left(1 - h_n(u)\right) \leq \frac{1}{2n}\|\mathbf{E}(\pi)\|_2^2 \leq \frac{1}{2}\operatorname{tr}\widetilde{\Omega}(\pi)\left(1 + H_n(u)\right)$$

with probability at least $1 - 2p\binom{p}{d}e^{-u^2/2}$.

Proof. Note that for any $\pi \in \mathbb{S}_p$,

$$(\text{C.26}) \qquad \frac{1}{2n}\|\mathbf{E}(\pi)\|_2^2 = \frac{1}{2n}\sum_{j=1}^p \|\widetilde{\varepsilon}_j(\pi)\|_2^2 = \frac{1}{2n}\sum_{j=1}^p \|\widetilde{\varepsilon}_j(S_j(\pi))\|_2^2.$$

Thus it suffices to bound the deviations in $\|\widetilde{\varepsilon}_j(S)\|_2$ for $S \subset [p]_j$. Consider the following events

$$\mathcal{G}_j(S) := \left\{\frac{\omega_j^2(S)}{2}\Big(1 - h_n(u)\Big) \le \frac{1}{2n}\|\widetilde{\varepsilon}_j(S)\|_2^2 \le \frac{\omega_j^2(S)}{2}\Big(1 + H_n(u)\Big)\right\}$$

and let $\mathcal{G} := \bigcap_{j=1}^p \bigcap_{S \subset [p]_j} \mathcal{G}_j(S)$. By Lemma C.6, we have $\mathbb{P}(\mathcal{G}_j(S)) \ge 1 - 2e^{-u^2/2}$, for all $S \in [p]_j$. By a monotonicity argument (cf. (8.3)), we have $\mathcal{G} = \bigcap_{j=1}^p \bigcap_{S \in m_j(\Sigma)} \mathcal{G}_j(M_j(S))$. Applying the union bound and using (3.4),

$$(\text{C.27}) \qquad \mathbb{P}(\mathcal{G}^c) \le 2p\binom{p}{d}e^{-u^2/2}.$$

Summing the inequalities defining $\mathcal{G}_j(S_j(\pi))$, over j, we conclude that (C.25) holds on $\mathcal{G}$. The proof is complete. $\qquad\square$

Consider the (random) collection of permutations

$$\mathbb{S}_p^0 = \mathbb{S}_p^0(\delta_0; u) := \left\{\pi \in \mathbb{S}_p : \frac{1}{2}\operatorname{tr}\widetilde{\Omega}(\pi)\Big[1 + H_n(u)\Big]\right.$$
$$\left. - \frac{1}{2}\operatorname{tr}\widetilde{\Omega}(\widehat{\pi})\Big[1 - h_n(u)\Big] \le \delta_0\rho_\lambda(\widetilde{B}(\pi))\right\}.$$

Lemma C.8. *For any $\pi \in \mathbb{S}_p^0(\delta_0; u)$ and $0 < u < n/\sqrt{n+1}$, we have*

$$\mathbb{P}\left(\frac{1}{2n}\|\mathbf{E}(\pi)\|_2^2 - \frac{1}{2n}\|\mathbf{E}(\widehat{\pi})\|_2^2 > \delta_0\rho_\lambda(\widetilde{B}(\pi))\right) \le 2p\binom{p}{d}e^{-u^2/2}.$$

Proof. Lemma C.7 implies that

$$\frac{1}{2n}\|\mathbf{E}(\pi)\|_2^2 - \frac{1}{2n}\|\mathbf{E}(\widehat{\pi})\|_2^2 \le \frac{1}{2}\operatorname{tr}\widetilde{\Omega}(\pi)\Big[1 + H_n(u)\Big] - \frac{1}{2}\operatorname{tr}\widetilde{\Omega}(\widehat{\pi})\Big[1 - h_n(u)\Big]$$

with probability at least $1 - 2p\binom{p}{d}e^{-u^2/2}$. Since $\pi \in \mathbb{S}_p^0$, the right-side is bounded above by $\delta_0\rho_\lambda\widetilde{B}(\pi)$ by definition, which establishes the claim. $\quad\square$

Lemma C.9. $1 - h_n(u) > 0$ *for all* $u \ne 0$, $n > 0$.

Proof. Since $(u + \sqrt{n})^2 + 1 > 0$, re-writing this inequality yields

$$\frac{u^2}{n} + 1 > \frac{2u}{\sqrt{n}} + \frac{1}{n} > \frac{2u}{\sqrt{n+1}} + \frac{1}{n+1}$$

$$\implies 1 + \frac{u^2}{n} - \frac{2u}{\sqrt{n+1}} - \frac{1}{n+1} > 0$$

Comparing with (C.21) yields the claim. $\qquad\square$

Proof of Proposition C.5. Lemma C.8 implies that for a choice of $u = \sqrt{2(C+1)(d+1)\log p}$, we have

$$\mathbb{P}\left(\frac{1}{2n}\|\mathbf{E}(\widehat{\pi})\|_2^2 - \frac{1}{2n}\|\mathbf{E}(\widehat{\pi})\|_2^2 > \delta_0\rho_\lambda(\widetilde{B}(\pi))\right) \le 2p\binom{p}{d}e^{-(C+1)(d+1)\log p}$$
$$\le 2e^{-C(d+1)\log p}$$

for any $\pi \in \mathbb{S}_p^0$. Thus the claim will follow if we can show that $\pi_0 \in \mathbb{S}_p^0$. Note that

$$\operatorname{tr}\widetilde{\Omega}(\pi_0)\Big[1 + H_n(u)\Big] - \operatorname{tr}\widetilde{\Omega}(\widehat{\pi})\Big[1 - h_n(u)\Big]$$
$$\stackrel{(i)}{\le} \operatorname{tr}\widetilde{\Omega}(\pi_0)\Big[H_n(u) + h_n(u)\Big]$$
$$\stackrel{(ii)}{\le} \operatorname{tr}\widetilde{\Omega}(\pi_0)\sqrt{\frac{50(C+1)(d+1)\log p}{n}}$$
$$\stackrel{(iii)}{\le} \delta_0\rho_\lambda(\widetilde{B}(\pi_0)),$$

where (i) follows from $\operatorname{tr}\widetilde{\Omega}(\pi_0) \le \operatorname{tr}\widetilde{\Omega}(\widehat{\pi})$ and Lemma C.9, (ii) follows by using (C.22) with $u = \sqrt{2(C+1)(d+1)\log p}$, and (iii) follows from assumption (C.20). Hence, $\pi_0 \in \mathbb{S}_p^0$ and the proof is complete. $\qquad\square$

References

B. Aragam and Q. Zhou. Concave penalized estimation of sparse Gaussian Bayesian networks. *Journal of Machine Learning Research*, 16:2273–2328, 2015.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

P. Bühlmann, J. Peters, J. Ernest, et al. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.

X. Cai, J. A. Bazerque, and G. B. Giannakis. Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. *PLoS Comput Biol*, 9(5):e1003068, 2013.

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

D. M. Chickering, D. Heckerman, and C. Meek. Large-sample learning of Bayesian networks is NP-hard. *The Journal of Machine Learning Research*, 5:1287–1330, 2004.

M. Drton, R. Foygel, and S. Sullivant. Global identifiability of linear structural equation models. *The Annals of Statistics*, 39(2):865–886, 2011.

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.

Y. Gordon. On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$. *Geometric Aspects of Functional Analysis*, pages 84–106, 1988.

J. Gu, F. Fu, and Q. Zhou. Adaptive penalized estimation of directed acyclic graphs from categorical data. *arXiv preprint arXiv:1403.2310*, 2016.

F. Guay, A. J. Morin, D. Litalien, P. Valois, and R. J. Vallerand. Application of exploratory structural equation modeling to evaluate the academic motivation scale. *The Journal of Experimental Education*, 83(1):51–82, 2015.

S. W. Han, G. Chen, M.-S. Cheon, and H. Zhong. Estimation of directed acyclic graphs through two-stage adaptive lasso for gene network inference. *Journal of the American Statistical Association*, 111(515):1004–1019, 2016.

D. Heckerman, D. Geiger, and D. M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243, 1995.

S. Horvath. *Weighted network analysis: applications in genomics and systems biology*. Springer Science & Business Media, 2011.

P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696, 2009.

J. Huang, P. Breheny, and S. Ma. A selective review of group selection in high-dimensional models. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 27(4), 2012.

M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *The Journal of Machine Learning Research*, 8:613–636, 2007.

D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

S. Lin, C. Uhler, B. Sturmfels, and P. Bühlmann. Hypersurfaces and their singularities in partial correlation testing. *Foundations of Computational Mathematics*, 14(5):1079–1116, 2014.

P.-L. Loh and P. Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *Journal of Machine Learning Research*, 15:3065–3105, 2014.

P.-L. Loh and M. J. Wainwright. Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *The Annals of Statistics*, 41(6):3022–3049, 2013.

P.-L. Loh and M. J. Wainwright. Support recovery without incoherence: A case for nonconvex regularization. *arXiv preprint arXiv:1412.5632*, 2014.

P.-L. Loh and M. J. Wainwright. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16:559–616, 2015.

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

J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *arXiv preprint arXiv:1412.3773*, 2014.

A. J. Morin, A. K. Arens, and H. W. Marsh. A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric multidimensionality. *Structural Equation Modeling: A Multidisciplinary Journal*, pages 1–24, 2015.

S. Ott and S. Miyano. Finding optimal gene networks using biological constraints. *Genome Informatics*, 14:124–133, 2003.

J. Peters and P. Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2013.

J. Peters, J. Mooij, D. Janzing, and B. Schölkopf. Identifiability of causal graphs using functional models. *arXiv preprint arXiv:1202.3757*, 2012.

J. Peters, J. M. Mooij, D. Janzing, B. Schölkopf, et al. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(1):2009–2053, 2014.

M. Pourahmadi. *High-dimensional covariance estimation*. John Wiley & Sons, 2013.

J. Ramsey, M. Glymour, R. Sanchez-Romero, and C. Glymour. A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International Journal of Data Science and Analytics*, pages 1–9, 2016.

G. Raskutti and C. Uhler. Learning directed acyclic graphs based on sparsest permutations. *arXiv preprint arXiv:1307.0366*, 2014.

G. Raskutti, M. J. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated Gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.

G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *Information Theory, IEEE Transactions on*, 57(10):6976–6994, 2011.

P. Ravikumar, M. J. Wainwright, J. D. Lafferty, et al. High-dimensional ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

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

S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *The Journal of Machine Learning Research*, 7:2003–2030, 2006.

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

T. Silander and P. Myllymaki. A simple approach for finding the globally optimal bayesian network structure. In *Proceedings of the 22nd Conference*

on Uncertainty in Artificial Intelligence, 2006.

A. P. Singh and A. W. Moore. Finding optimal bayesian networks by dynamic programming. 2005.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search*, volume 81. The MIT Press, 2000.

M. Teyssier and D. Koller. Ordering-based search: A simple and effective algorithm for learning bayesian networks. *arXiv preprint arXiv:1207.1429*, 2012.

C. Uhler, G. Raskutti, P. Bühlmann, and B. Yu. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, 41(2):436–463, 2013.

S. van de Geer and P. Bühlmann. ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *The Annals of Statistics*, 41(2):536–567, 2013.

N. Wermuth. Linear recursive equations, covariance selection, and path analysis. *Journal of the American Statistical Association*, 75(372):963–972, 1980.

S. Wright. Correlation and causation. *Journal of agricultural research*, 20 (7):557–585, 1921.

S. Wright. The method of path coefficients. *The Annals of Mathematical Statistics*, 5(3):161–215, 1934.

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.

E. Yang, P. Ravikumar, G. I. Allen, and Z. Liu. Graphical models via univariate exponential family distributions. *Journal of Machine Learning Research*, 16:3813–3847, 2015.

K.-H. Yuan and P. M. Bentler. *Structural Equation Modeling*, volume 26. Handbook of Statistics, 2007.

C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

C.-H. Zhang and J. Huang. The sparsity and bias of the Lasso selection in high-dimensional linear regression. *The Annals of Statistics*, pages 1567–1594, 2008.

C.-H. Zhang and T. Zhang. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27(4): 576–593, 2012.

J. Zhang and P. Spirtes. Strong faithfulness and uniform consistency in causal inference. In *Proceedings of the nineteenth conference on uncertainty in artificial intelligence*, pages 632–639. Morgan Kaufmann Publishers Inc., 2002.

K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 647–655. AUAI Press, 2009.

P. Zhao and B. Yu. On model selection consistency of Lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.