THE UNIVERSITY OF CHICAGO

STATISTICAL OPTIMALITY IN CAUSAL STRUCTURE LEARNING

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE UNIVERSITY OF CHICAGO
BOOTH SCHOOL OF BUSINESS
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY
MING GAO

CHICAGO, ILLINOIS
JUNE 2026

© 2026 by Ming Gao

ORCID iD: https://orcid.org/0000-0002-3794-2881

To those who made this disenchanting journey an enchanting one.

“Desire is the desire of the Other.” — Jacques Lacan

CONTENTS

LIST OF FIGURES viii

LIST OF TABLES x

ACKNOWLEDGMENTS xi

ABSTRACT xv

1 INTRODUCTION 1
1.1 Related Work 8
1.2 Summary of Contributions 13
1.3 Preliminaries 15
1.3.1 Causal Graphical Model 15
1.3.2 Model Selection and Optimal Sample Complexity 18
1.3.3 Notations 20

2 OPTIMAL STRUCTURE LEARNING IN PARAMETRIC MODELS 22
2.1 Problem Setup 22
2.2 Optimality in Equal Variance DAG 24
2.2.1 Upper Bound and Algorithm 24
2.2.2 Lower Bound 27
2.3 Optimality in Faithful Poly-tree 30
2.3.1 Upper Bound and Algorithm 30
2.3.2 Lower Bound 32

3 NONPARAMETRIC MODELS IN TREES: REDUCTION FROM CONDITIONAL INDEPENDENCE TESTING 34
3.1 Problem Setup 34
3.2 Reduction from CI Testing to Structure Learning 38
3.3 Applications 40
3.3.1 Application to Bernoulli Distribution 40
3.3.2 Application to Gaussian Distribution 42
3.3.3 Extension to Optimal Poly-tree Learning 43
3.3.4 Application to Nonparametric Continuous Distribution 44

4 TOWARD GENERAL STRUCTURES: RETHINKING VARIABLE SELECTION IN CAUSAL GRAPHICAL MODELS 47
4.1 Problem Setup 47
4.2 Optimality and Computational Barriers in Variable Selection 51
4.2.1 Optimality of BSS for Variable Selection When Sparsity Is Known . . 52
4.2.2 Optimality of BSS for Variable Selection When Sparsity Is Unknown 56
4.2.3 Polynomial-Efficient Sample Complexity Lower Bound 59

4.3 KL-BSS: A Better Choice for Neighbourhood Selection . . . 64
4.3.1 KL-BSS Estimator . . . 64
4.3.2 Analysis of KL-BSS . . . 68
4.3.3 Practical Considerations . . . 73
4.3.4 Experiments . . . 78

5 CONCLUSION . . . 84

APPENDIX A PROOFS IN CHAPTER 2 . . . 86
A.1 Proof of Theorem 2.2.1 (Upper Bound for EqVar DAG) . . . 87
A.2 Proof of Theorem 2.2.2 (Upper Bound for EqVar DAG without tuning parameter) . . . 93
A.3 Proof of Theorem 2.2.3 (Lower Bound for EqVar DAG) . . . 93
A.3.1 Sample Conditional Correlation Coefficient as CI Tester . . . 97
A.4 Proof of Theorem 2.3.1 (Upper Bound for Faithful Poly-tree) . . . 102
A.5 Proof of Proposition 2.3.2 . . . 105
A.6 Proof of Theorem 2.3.3 (Lower Bound for Faithful Poly-tree) . . . 109

APPENDIX B PROOFS IN CHAPTER 3 . . . 117
B.1 Details of Applications . . . 117
B.1.1 Bernoulli Distribution (Section 3.3.1) . . . 117
B.1.2 Gaussian Distribution (Section 3.3.2) . . . 118
B.1.3 Nonparametric Continuous Distribution (Section 3.3.4) . . . 119
B.2 Proof of Theorem 3.2.1 . . . 120
B.3 Proof of Theorem 3.3.1 (Bernoulli Distribution) . . . 125
B.3.1 Proof of Upper Bound for Bernoulli Distribution . . . 125
B.3.2 Proof of Lower Bound for Bernoulli Distribution . . . 129
B.4 Proof of Theorem 3.3.2 (Gaussian distribution) . . . 130
B.5 Proof of Theorem 3.3.3 (Poly-tree Learning) . . . 131
B.6 Proof of Theorem 3.3.4 (Nonparametric Continuous Distribution) . . . 135

APPENDIX C PROOFS IN CHAPTER 4 . . . 142
C.1 Proof of Lemma 4.2.1 (Variable selection Upper Bound) . . . 142
C.2 Proof of Theorem 4.2.3 (Variable Selection Lower Bound) . . . 147
C.3 Proof of Theorem 4.2.5 (Variable Selection Upper Bound) . . . 150
C.4 Proof of Theorem 4.2.6 (Variable Selection Lower Bound) . . . 159
C.5 Neighbourhood Selection and Variable Selection . . . 161
C.6 Proof of Lemma 4.2.7 . . . 165
C.7 Discussion of KL-BSS . . . 167
C.7.1 Signal and Analysis of KL-BSS in General Variable Selection . . . 167
C.7.2 Vanilla KL-BSS . . . 170
C.7.3 Beyond Gaussian design . . . 173
C.8 Proof of Lemma 4.3.1 . . . 174
C.9 Proof of Theorem 4.3.2 . . . 175

C.10 Proof of Theorem 4.3.3 . . . 176
C.11 Proof of Example 1 . . . 179
C.12 Proof of Theorem C.7.1 . . . 181
C.13 Proof of Lemma C.12.1 . . . 183
C.14 Proof of Lemma C.7.2 . . . 192

APPENDIX D AUXILIARY LEMMAS . . . 194
D.1 Concentration of χ^2 Random Variable . . . 194
D.2 Lower Bound Techniques . . . 195

APPENDIX E EXPERIMENT DETAILS . . . 198
E.1 Simulation Setup . . . 198
E.2 Real Data Application . . . 200
E.2.1 Selection of Genes . . . 200
E.2.2 Prediction Performance . . . 201
E.3 Structure Learning . . . 201
E.4 Additional Metrics . . . 202
E.5 Overall Comparison . . . 203
E.6 Vanilla KL-BSS and KL-BSS . . . 203

BIBLIOGRAPHY . . . 207

LIST OF FIGURES

4.1 A graphical model over $X = (X_1, \ldots, X_d)$ with one more target node X_{d+1} appended to it. The corresponding G will refer to a DAG over X (ignoring X_{d+1}). The neighbourhood of X_{d+1} is $\{X_2, X_{d-2}, X_{d-1}\}$ in this example. 50

4.2 Comparison on variable selection performance of BSS, KL-BSS and Lasso under different types of graphs and dimensions $(d, s, \bar{s})$ averaged over 200 replications. The horizontal axis is sample size, the vertical axis is probability of exact recovery. The first/middle/last two columns are for ER graph, SF graph, and complete graph. The left/right columns of each graph type indicate known and unknown sparsity cases. The solid/dashed lines indicate known and unknown $\beta_{\min}$ (by CV). There is a notable performance gap between KL-BSS and BSS. Lasso is never consistent. 79

4.3 Comparison on structure learning performance of BSS and KL-BSS on ER and SF graphs: Structural Hamming Distance (SHD) vs. sample size. The first and third panels are input with oracle valid topological ordering of the graph, the second and forth apply EqVar algorithm for ordering estimation, then conduct neighbourhood selection via KL-BSS or BSS. KL-BSS gives better structure learning performance than BSS. KL-BSS and Vanilla KL-BSS perform similarly thus lines are overlapped. 81

4.4 The percentage of randomly sampled SEM covariances that satisfy the constraint in Ω_Δ (4.20) for ER/SF graphs with various expected number of edges $(s_0 \times d)$ and sparsity s and fixed $d = 12$. Significant proportion of random SEM covariances meet Ω_Δ. 82

4.5 (Left) Recovery performance comparison on RNA-Seq gene expression data with $s = 10$. The solid lines are for KL-BSS and dashed lines are for BSS. The barplots indicate the gap between the performances of KL-BSS and BSS, which is more significant for larger d and smaller sample size. (Right) Prediction performance of $\widehat{S}$ given by KL-BSS with CV choice of $\beta_{\min}$ and BSS on gene expression data. Each point is one random sampling of $d = 50$ variables from $20,531$ genes to choose $s = 10$ in training set for prediction in test set. The grey dashed line is the $y = x$ line. The genes selected by KL-BSS result in better prediction error than BSS (above the $y = x$ line). 83

A.1 Four cases of ℓ to verify for c-strong Tree-faithfulness, indicated by the superscript of X_ℓ. The first case is when $\ell = \emptyset$. The second, third and fourth are when ℓ is the ancestor of j, descendant of j and descendant of k. 106

E.1 Detailed comparison of KL-BSS vs. BSS. Overall evaluation by percentage of improvement over BSS on a per-dataset basis (i.e. not averaged for all replications in each setting). The solid bars indicate percentage of strict improvement. The transparent bars above the solid bars indicate the percentage of tied performance. For exact variable selection metric (the second panel from left), if representing the outcomes of KL-BSS and BSS as a tuple, we count $(1, 0)$ as strict improvement, $(1, 1)$ as tied performance, and the most transparent bars indicate percentage of $(0, 0)$. 202
E.2 Comparison between KL-BSS and Vanilla KL-BSS on covariance matrix with large maximum eigenvalue. KL-BSS gives better performance when the covariance (sub)matrix is relatively harder to estimate. 204
E.3 Experiment results of Hamming distance for the same setups with Figure 4.2. . 205
E.4 Experiment results of FDR for the same setups with Figure 4.2. 205
E.5 Experiment results of TPR for the same setups with Figure 4.2. 206

LIST OF TABLES

4.1 Summary of comparison of KL-BSS with existing methods under model (4.24). The first column is the names of methods under comparison; the second column is the sample complexity upper bound of each method specified under (4.24) or the conclusion about the critical condition of the method; the last column is further explanation about the second column. 73

ACKNOWLEDGMENTS

I am deeply grateful to the people I have shared this PhD journey with, they are indispensable to this thesis and dissertation. I always believe that human language or symbolic order does not fully capture our reality. What I feel in gratitude lies far beyond the pale and feeble words I can put down here. Therefore, I offer them in the most honest and sincere way I can.

First and foremost, my deepest gratitude to my advisor, Bryon Aragam, for his support and guidance over the past seven years. I learned how to recognize meaningful research questions, conduct rigorous and high-quality research, and communicate ideas effectively. I still remember our very first meeting. Bryon asked why I chose to pursue research in statistics, and I replied that I wanted to discover or develop something that does not change over time, unlike social constructs created by human beings. He agreed with me then, though I am still unsure whether it was out of politeness. Later on, I came to realize that, the only things that do not change (within the scope of academic research) are: (i) the ever-changing focus/taste/trends of the research community, and (ii) the ability to recognize and adapt to these changes. Bryon also holds annual one-on-one meetings with his students. I remember there was one question on his questionnaire "How do you imagine yourself in 10 years", and my answer was to raise sheep on a pasture in Scotland. We then had a serious conversation about how to achieve that.

Secondly, I am grateful to Veronika Ročková, Panagiotis Toulis, Shuangning Li for generously serving on my dissertation committee, and contributing their time and expertise when I was on job market. I further thank all the faculty in Booth E&S and Statistics Department for their guidance and education. In particular, I would like to thank Jingshu Wang and Dacheng Xiu for their support and help.

Five years is long and short. Long enough for AI to reshape the world; short enough for someone in their 20s to suddenly feel like 60s. People often say that pursuing a PhD is

a waste of time, at least from a utilitarian perspective. I would say for myself it has been a perfect usage of time, to pause at the most restless period of my life, to reflect on deeper questions, to figure out what I want and want to be, and to simply do nothing.

Pessimistically, I started to believe human life is completely meaningless in the very first year of my PhD, I struggled for a while. But over time, I come to recognize the beauty (and difficulty) of interactions between individuals. We often find ourselves missing places we once stayed. Yet when we return, the feeling is no longer there. I come to realize what we miss is not the place itself, either a building or an address, but the moments we shared with the people who were there with us. The place remains; the moment does not. So I now have considered the small, memorable moments shared with others as the motivation to proceed with such "misery".

Next, I would also like to express my gratitude to my wonderful cohorts across Booth, Stats, CAM, CS departments at the University of Chicago. They have been not only my academic peers but also companions in my daily life. I thank them in alphabetical order (last name, first name), along with a few memorable moments and impressions:

Shirui Chen, for her alien intelligence and for her simple coherence in this winding world; Yifei Chen, for her differentiation from a stem cell and for her mysterious left hair tip; Chang Deng, for his WeChat QR code in the first email and the bench press in the last academic year; Wenxuan Guo, for always listening carefully to my naive philosophical reflections and for being my last belief in academia; Shunzhuang Huang, for karaoke together, for introducing me to Girls' Last Tour (2D), and for offering me a luosifen; Sowon Jeong, for asking me if I considered doing AI research which I tried to avoid; Yibo Jiang, for his manual transmission car and for the discussion on topology; Takuya Koriyama, for his flying-in-the-air Mario Kart and for sharing a similarly tragic perspective on life; Jiguang Li, for developing and maintaining an encrypted language system and for discussions on philosophical and artistic works; Jizhou Liu, for the wonderful business trip

in Spain and for accusing/praising me of being a saint; Yanlong Liu, for his endless storytelling and for his unlimited capacity for tea; Edoardo Marcelli, for educating me away from steroids and for teaching me special Italian gestures; Zhouyu Shen, for the ridge and Lasso presentation; Jiabi Wang, for her montage-like (sometimes behavior scientific) stories and for the cute Kirby Perler beads; Yinjie Wang, for the small kitty magnets and for his youthful aspirations about AI; Tong Xie, for her clear understanding of academia and for her resting position while gazing out the window in the office; Boxin Zhao, for the upfront discussions in the ML reading group and for always calling me a "master"; Shengjun (Percy) Zhai, for his spacious and delicate apartment(s) and for always being surprised by–and responsive to–what I say; Cong Zhang, for sharing details/thoughts in all aspects of life and for showing me Roosevelt Park as a hidden gem; Lyu Zhao, for always calling us to care for her, while being the one who cared for me when I was ill in Spain.

I have also been fortunate to work with many postdocs and collaborators. I thank Wai Ming Tai, for describing everything as either "good" or "no good" and for not doing much work in our collaboration(s); Inbeom (Daniel) Lee, for swimming, sauna, tea, Jo Malone, Cho Sun Ok, and reviewing memes I sent him; Yuhao Wang, for many years of collaboration while being able to drive me mad all the time; Jinzhou Li, for being my new collaborator and having a wonderful new collaboration.

Finally, I thank my friends in Chicago, across the United States, and around the world. Again, in alphabetical order: Tianyi Chen, for wishing me to find the inner freedom; Xinkai Chen, for being my roommate for five years during the PhD and two years in high school and for sharing all aspects of life; Yu-Wei Chen, for talking nonsense and playing guitar in my apartment; Tianrui (Dora) Hou for visiting Chicago with the highest frequency and for the attempt to convert me into a libertarian; Zewei Lin, for hosting six people in an 1B1B and for hosting the best marriage I ever attended; Puyuan Peng,

for his honesty and confidence and default brotherhood; Xiaoyu Sun, for drag racing in every trip while keeping a negative traveling cost; Mufeng Tang, for lakeside walks and conversations about Dota, even though I no longer play; Benjamin Wang, for the constant greeting at the gym and for forcing me to run outside; Yingzhuo (Dessi) Wang, for her energetic entertainment schedule and for keeping me updated on RedNote; Zhiyuan (Zephyr) Xu, for being a considerate host during every visit and for fluently reciting Empresses in the Palace; and everyone I met during my internship at Berkeley.

Lastly, I thank my family for their support.

ABSTRACT

Causal graphical models are fundamental tools for representing complex dependency structures in probabilistic machine learning. When the underlying structure is unknown, a central task is structure learning, which aims to recover the causal graph from data. A key open problem is to characterize its statistical optimality and identify procedures that achieve it. While the statistical limits are well understood for undirected models, the directed setting remains much less explored.

We study this problem through the lens of minimax sample complexity, focusing on both structure learning and its core subroutines. For parametric Gaussian models, we establish sharp optimal sample complexity results for (i) DAGs under the equal-variance assumption and (ii) tree-structured models under faithfulness. Moving beyond parametric settings, we develop a general reduction from structure learning to conditional independence (CI) testing, showing that the optimal rate for structure learning is governed by that of CI testing. This provides a unified framework applicable to a broad class of models, including fully nonparametric settings.

Moving towards general DAGs, we investigate neighborhood selection as a key subroutine. We first study variable selection in linear regression under general design and show that best subset selection (BSS) achieves optimal sample complexity, while no polynomial time algorithm can attain this optimality under standard complexity conjectures, revealing a fundamental statistical–computational trade-off. We then demonstrate that, under structured designs induced by linear structural equation models, BSS can be improved. We propose a new estimator, KL-BSS, which exploits the unknown structure to achieve better sample complexity with weaker eigenvalue conditions, and establish both pointwise and minimax optimal guarantees. Together, these results provide a comprehensive perspective on statistical optimality in causal structure learning, highlighting the roles of the model assumptions, problem parameters, and basic subroutines.

CHAPTER 1

INTRODUCTION

Causal graphical models are important tools in probabilistic machine learning for representing complex dependency structures that arise in a wide range of applications in causality, artificial intelligence, and statistics [Spirtes et al., 2000, Pearl, 2010, Murphy, 2012]. When these structures are known in advance, it is straightforward to exploit this structure for inference tasks. When the structure is unknown, a crucial preliminary step of using causal graphical models is *structure learning*, which seeks to recover the underlying causal graph from data. The accuracy of structure learning directly impacts the performance of downstream tasks. A significant open problem in causal structure learning is to characterize the statistical optimality, especially, the optimal sample complexity, i.e. the minimum number of samples required to reliably recover the causal graph from data, and what approach achieves such an optimal performance. Despite all the methodological development, a tight characterization of the statistical optimality remains largely unresolved. This stands in contrast to structure learning of undirected graphical models (Markov random fields), the information-theoretic limits of structure learning have been widely investigated [Wang et al., 2010a, Santhanam and Wainwright, 2012, Drton and Maathuis, 2017, Misra et al., 2020]. In comparison, much less is known about causal (directed) structure learning, which presents several distinctive technical challenges, e.g. causal ordering, Markov equivalence class, and the faithfulness assumption. Therefore, a comprehensive characterization remains an important open research direction. Motivated by these challenges, this thesis aims to investigate the following central question:

What is a right characterization of statistical optimality in causal structure learning?

In particular, the goal is to understand the optimal sample complexity required for learning causal graphical models and to identify algorithms that achieve this optimality. A

key perspective of this work is to study the subroutines that arise in structure learning procedures. For instance, conditional independence (CI) testing plays a central role in constraint-based methods, while variable selection techniques are commonly used for estimating the neighborhood of a node in the graph.

The problem of deriving upper bounds on the sample complexity for causal structure learning goes back decades [Zuk et al., 2006, Friedman and Yakhini, 1996], and in recent years there has been significant progress [Ghoshal and Honorio, 2017b, Chen et al., 2019, Park and Raskutti, 2017, Park, 2018, Park and Park, 2019, Park, 2020, Wang and Drton, 2020, Gao et al., 2020, Gao and Aragam, 2021], though a tight characterization of the optimal sample complexity is missing. Causal structures are often modeled using directed acyclic graphs (DAG). Observational data only reveal the Markov equivalence class, captured by a completed partially directed acyclic graph (CPDAG) [Andersson et al., 1997]. Classical approaches to learning a CPDAG from data include the PC algorithm [Spirtes and Glymour, 1991, Kalisch and Bühlman, 2007] and GES [Chickering, 2002, Nandy et al., 2018]. Moreover, it is also known that the general problem of learning DAGs from observational data is an NP-complete problem [Chickering, 1996, Chickering et al., 2004], although a few polynomial-time algorithms have been proposed for special cases [Ghoshal and Honorio, 2017b, Chen et al., 2019, Park, 2020, Gao et al., 2020].

We start by studying the simplest DAG model with parametric assumptions such as Gaussianity. Perhaps surprisingly, despite its simplicity, and unlike in the undirected case, the optimal sample complexity of learning Gaussian DAGs has remained an open problem. We consider two identifiable cases. The first one is linear Gaussian DAG model with equal error variances. The identifiability of these models was established in Peters and Bühlmann [2013], and eventually led to the development of several polynomial-time algorithms under the equal variance assumption [Ghoshal and Honorio, 2017b, Chen et al., 2019, Gao et al., 2020]. Nonetheless, it was not known whether or not any of these

algorithms were optimal for this precise statistical setting. We will show that a variant of the EqVar algorithm from Chen et al. [2019] is indeed optimal. This involves the derivation of new lower bounds and a novel analysis of the EqVar algorithm that sharpens the existing sample complexity upper bound. The second case is the faithful Gaussian polytree. Although faithfulness can be a strong assumption in practice, we emphasize that to the best of our knowledge, no optimality results under this assumption are known. Thus, our analysis presents a possible first foray in this direction. Previous work has shown that faithfulness is notoriously challenging to analyze [e.g. Uhler et al., 2013].

Next, we would like to move further and avoid specific distributional assumptions are needed in the analysis (e.g. linear SEM with equal error variances, Gaussianity). It is well-known that structure learning is closely related to conditional independence (CI) testing, which itself is a fundamental topic that has attracted significant attention in recent years. Owing to its foundational role, CI testing has been extensively studied both methodologically and theoretically, leading to a rich body of literature on its statistical properties [Zhang et al., 2011, Shah and Peters, 2020, Canonne et al., 2018, Neykov et al., 2021]. In particular, structure learning methods often rely on CI testing as a subroutine for inferring graphical dependencies [Spirtes and Glymour, 1991, Maathuis et al., 2018].

The relationship between CI testing and structure learning has long been understood from the algorithmic perspective, which takes CI tests as black-box oracles and often assumes perfect CI information to guarantee graph recovery. However, the relationship between their finite-sample complexity and statistical hardness has yet to be formalized. Specifically, given that graphical models are intrinsically defined by the Markov property, a deeper understanding of how the statistical difficulty of CI testing governs the learning of graphical structures is both natural and fundamental. By contrast, CI testing in its own right has been studied more broadly, with well-established minimax results across various distributional settings [Canonne et al., 2018, Neykov et al., 2021]. Moreover, existing

optimality results do not establish a general (information-theoretic) reduction between structure learning and CI testing, which would be useful both in its own right, as well as for future investigations into the hardness of structure learning broadly. These gaps suggest an opportunity to leverage statistical results in CI testing to better understand the sample complexity of structure learning for DAGs in general.

We aim to establish a general reduction between the minimax optimality of structure learning in DAG models and the minimax optimality of CI testing. In addition to establishing this reduction in a general setting that allows us to analyze non-Gaussian models, we also show how Gaussian results can be deduced as a special case. We focus on poly-forests, a tractable yet rich subclass of DAGs that captures structured dependencies [Chow and Liu, 1968, Dasgupta, 1999], serving as a principled starting point toward fully general DAGs. We develop results under generic conditions that apply to both parametric and nonparametric distributions. Our work provides a framework connecting the statistical complexity of CI testing with that of structure learning, offering new insights into the understanding of sample-efficient DAG recovery in diverse settings.

Finally, a basic operation to learn the structure of a graphical model is the estimation of the neighbourhood of a given node. Under fairly general assumptions, this problem reduces to the familiar problem of variable selection, a.k.a. support recovery, and has been extensively studied as a prototypical model selection problem [Shibata, 1981, Nishii, 1984, Foster and George, 1994, Shao, 1997a, Meinshausen and Yu, 2009, Wainwright, 2009a, Wang et al., 2010b, Aksoylar et al., 2016]. Variable selection itself is a classical problem in statistical learning theory. It aims to select the most effective subset of variables for predicting a target variable. Variable selection has drawn more attention in the high-dimensional era, and many computationally efficient algorithms have been proposed and studied [Tibshirani, 1996, Fan and Lv, 2008].

We consider the variable selection task for linear regression with Gaussian noise:

$$Y = X^\top \beta + \epsilon\,, \qquad X \sim \mathcal{N}(0, \Sigma)\,, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2)\,.$$

In this setup, variable selection is also known as support recovery of β vector. In the view of minimax optimality, existing work primarily considers standard design [Aeron et al., 2010, Reeves et al., 2019], i.e. the covariance $\Sigma = I_d$ where d is the dimension, and conclude the optimal sample complexity. While standard design is reasonable in certain experimental settings, it is more practical to study general design covariance beyond I_d in modern applications. In particular, the potential presence of strong dependence among variables (such as in graphical models) imposes difficulties for existing computationally efficient methods, making it of special theoretical interest. Although finite sample analysis in general design has been explored in prior work [Wainwright, 2009a, Shen et al., 2012, 2013], a gap remains between the obtained upper and lower bounds with respect to the problem parameters of β and Σ. On the other hand, the computational cost of variable selection is known to be prohibitively high [Natarajan, 1995]. Therefore, simultaneous statistical and computational optimality are unresolved. This poses the natural question of whether or not there exists a simultaneously computationally efficient and sample optimal algorithm under general dependence.

To approach this, it is necessary to first establish matching upper and lower bounds on the sample complexity while setting aside computational considerations. Only then can we attempt characterize the (potential) statistical-computational trade-off in variable selection. We focus on problems with general design covariance, which exhibits strong dependence between variables. Hence, naive approaches based on thresholding or marginal independence testing between response and covariates typically fails, and ℓ_1-penalty based methods or other computationally efficient alternatives potentially require additional assumptions to perform, further highlighting the computational challenges of

this problem. Moreover, existing studies on sample complexity typically assume the exact knowledge of sparsity level s, which is the number of nonzero entries of β. While convenient to simplify the analysis, it sidesteps important technical challenges that arise when the sparsity is unknown and requires special treatment. Naturally, it is more practical and realistic to assume only an upper bound $\bar{s} \geq s$ is provided. Under this setup, while consistency results are available for information-criterion type estimators, the finite sample analysis is still an open line of research. Therefore, the question of potential extension of optimality from known sparsity case to this more general setting has yet to be addressed.

Furthermore, despite this long line of work, existing results are insufficient for understanding the nuances of neighbourhood selection in graphical models with structured dependencies. There is an exception for *undirected*, Markov random fields, for which much is now known [for an overview, see Drton and Maathuis, 2017, Loh, 2018]. In the setting of *directed*, structural equation models (SEM), however, although regression is widely used for neighbourhood selection in practice, a detailed understanding of the tradeoffs—both practical and theoretical—in neighbourhood selection is missing. This leads to a simple, fundamental question: *Are existing variable selection techniques adequate for neighbourhood selection in SEM?* The obvious candidates are BSS and the Lasso [Tibshirani, 1996]. BSS is known to be effective for variable selection with general random design matrices [Wainwright, 2009a], whereas the Lasso requires nearly orthogonal designs [i.e. the irrepresentability condition; Zhao and Yu, 2006, Wainwright, 2009b]. By imposing structural assumptions through the (directed) Markov property, SEM represent a potential middle ground between nearly orthogonal designs (required by the Lasso) and general design (allowed by BSS). Thus, the question is whether the neighbourhood selection problem for SEM is essentially equivalent to general design regression or—if not—how to leverage *unknown* structure to design better estimators of the structure itself. Crucially, we emphasize that we *will not* assume the structure of the SEM itself is known.

We continue to answer this fundamental question and in doing so, propose a new method for variable selection in SEM that illustrates the deficiencies with existing methods and shows how they can be improved. We quantify these deficiencies via an analysis of the minimax rate of variable selection in SEM, which is achieved by our approach, as well as its pointwise rates. A major takeaway is that it is often *easier* to recover the neighbourhood in linear SEM, even when its structure is unknown. This shows that worst-case analyses under general random designs are overly pessimistic, and the mere existence of structure can simplify the recovery problem. Our method, called KL-BSS, locates the hidden signature left by the unknown SEM structure, obtaining significant improvements in accuracy, and is easily implemented. In fact, in the worst-case, our method performs no worse than BSS on average.

From a broad statistical perspective, this dissertation lies at the intersection of graphical models, model selection, and minimax theory. Our primary motivation comes from graphical models and their wide-ranging applications, including causality, Markov random fields, and other structured models. We formulate the learning problems arising from these models within the general framework of model selection, a major problems in statistics that encompasses hypothesis testing and variable selection as important special cases. Adopting the minimax perspective, we develop understandings of statistical efficiency and computational consideration in model selection, with a particular focus on graphical models and/with their basic subroutines, which themselves are instances of model selection problems. Together, these contributions provide an answer to the research question posed at the outset of this dissertation and offer broader insights into the nature of model selection for the statistics community.

1.1 Related Work

Conditional Independence Testing CI testing forms a crucial building block in machine learning and reasoning tasks. Given a triplet of random variables (X, Y, Z), the problem asks whether X is conditionally independent of Y given Z. Numerous methods [e.g. Dawid, 1979, Fukumizu et al., 2007, Ramsey, 2014, Li and Fan, 2020] have been developed to address this fundamental problem. One prominent class of methods relies on measuring the distance between the conditional distribution $P(X, Y \mid Z)$ and the product of the marginals $P(X \mid Z)P(Y \mid Z)$. For Gaussian variables, the problem reduces to testing whether the partial correlation is zero [Fisher, 1915]. For discrete variables, hypothesis tests based on chi-squared statistics are widely employed Ireland and Kullback [1968], Darroch et al. [1980]. Recent advancements have also explored kernel-based methods Gretton et al. [2007], Zhang et al. [2011] and information-theoretic measures such as conditional mutual information [Runge, 2018, Berrett and Samworth, 2019] that can capture complex nonlinear dependencies in a nonparametric fashion.

From the theoretical standpoint, CI testing has been investigated from the perspective of minimax optimality. For discrete distributions, the problem is well-understood and the minimax optimal rates have been established [Chan et al., 2014, Diakonikolas and Kane, 2016, Canonne et al., 2018]. In the nonparametric setting, Neykov et al. [2021] studied the fundamental limits of CI testing, deriving minimax optimal rates under smoothness assumptions. In addition, practical CI tests have been proposed [Kim et al., 2022, 2024]. In particular, Jamshidi et al. [2024] devised a Von Mises estimator for mutual information as a valid CI test, and showed it achieves the parametric rate once the density is sufficiently smooth. These theoretical results provide valuable insights into the inherent difficulty of CI testing under different distributional setups.

Graphical Model Structure Learning Learning the structure of graphical models from observational data is a fundamental problem in machine learning. For undirected graphical models, the problem reduces to support recovery of the precision matrix [Meinshausen and Bühlmann, 2006, Friedman et al., 2008, Cai et al., 2011, Liu et al., 2009]. For DAG learning, approaches can be broadly categorized into three main paradigms: score-based methods (e.g. greedy equivalence search) [Chickering, 2002, Nandy et al., 2018], constraint-based methods (e.g. PC algorithm) [Spirtes and Glymour, 1991, Friedman et al., 2013], and methods that leverage specific distributional assumptions that allow for effective learning such as non-gaussianity [Shimizu et al., 2006, Shimizu, 2014, Wang and Drton, 2020], non-linearity [Hoyer et al., 2008, Peters et al., 2014, Zhang and Hyvärinen, 2009] or equal error variances [Peters and Bühlmann, 2014, Ghoshal and Honorio, 2017b, 2018, Chen et al., 2019, Gao et al., 2020]. Especially, constraint-based methods rely on valid CI tests and the faithfulness assumption to operate [Kalisch and Bühlman, 2007, Marx et al., 2021], thus development of reliable and efficient CI tests directly impacts the performance of these structure learning methods. For tree-structured models, the Chow-Liu algorithm [Chow and Liu, 1968, Chow and Wagner, 1973] provides an efficient method to find the optimal undirected tree structure by constructing a maximum weight spanning tree based on pairwise mutual information. Learning poly-trees/forests is generally more complex than learning undirected trees models, while remaining tractable compared to learning general DAGs [Rebane, 1987, Srebro, 2003, Tan et al., 2010, 2011]. Recent developments include [Gao and Aragam, 2021, Azadkia et al., 2021, Jakobsen et al., 2022]. Furthermore, learning poly-trees/forests is relevant in various applications, including reasoning and density estimation [Kim and Pearl, 1983, Liu et al., 2011].

Sample Complexity of Structure Learning Understanding the sample complexity of learning graphical model structures is crucial for assessing the feasibility and reliability of structure learning algorithms. For undirected graphical models, optimal sample

complexities have been derived for various classes of undirected graphs, such as sparse graphs with bounded degree [Wang et al., 2010a, Santhanam and Wainwright, 2012, Bresler, 2015, Vuffray et al., 2016, Abbe, 2018, Misra et al., 2020]. In contrast, determining the sample complexity for DAG learning is considerably more challenging due to the inherent asymmetry, the larger space of possible graph structures, and the non-identifiability. These help explain the relative scarcity of sharp results in the literature and underscore the importance of pursuing this line of research. We mention prior work that considers consistency and upper bounds under faithfulness [Kalisch and Bühlman, 2007, Nandy et al., 2018, Rothenhäusler et al., 2018], relaxation and improvement on classical methods [Chickering, 2020, Marx et al., 2021, Lam et al., 2022]. Existing work on finite sample bounds often rely on specific assumptions about the underlying data distributions. Ghoshal and Honorio [2017a] provided lower bounds for structure learning of general DAG models. Our work in Gao et al. [2022] on optimality for equal variance DAG has been further extended to linear dynamic models [Veedu et al., 2024] and people try to improve the dependence on other problem parameters [Daskalakis et al., 2025]. For the nonparametric setting, Jamshidi et al. [2024] applied the devised Von Mises CI test to the PC algorithm while no matching lower bound is provided.

Variable Selection and Neighbourhood Selection As a fundamental and long-lasting problem, the existing work on variable selection is rich in nature. We intend to review the most relevant work to our focus of minimax optimality, particularly under the lens of general design matrices, without diving into every aspect of variable selection. The concept of variable selection traces back to the foundational work of ANOVA [Fisher, 1936b,a, 1970]. Numerous textbook methods have been developed since then, including the stepwise regression [Efroymson, 1960, Miller, 1984, Draper, 1998], best subset selection [Hocking and Leslie, 1967], various types of information criteria [Mallows, 2000, Akaike, 1974, Schwarz, 1978, Konishi and Kitagawa, 1996], and cross-validation [Stone, 1974]. The large sample

properties of these methods, e.g. asymptotic efficiency and (in)consistency, have been obtained in different asymptotic regimes [Nishii, 1988, Shao, 1993, 1997b]. The modern era of high-dimensional data has seen substantial interest in ℓ_1-based methods. Important representatives include Lasso [Tibshirani, 1996] and Orthogonal Matching Pursuit [Tropp and Gilbert, 2007]. These methods achieve variable selection consistency under conditions like irrepresentability or mutual incoherence [Zhao and Yu, 2006, Wainwright, 2009b, Zhang, 2011, Cai and Wang, 2011]. Extensions to them include thresholding the Lasso-type estimates [Meinshausen and Yu, 2009], replacing the ℓ_1 penalty with other nonconvex choices [Fan and Li, 2001, Loh and Wainwright, 2017, Zhang, 2010], and multi-stage methods combining estimation power and thresholding thereby [Ndaoud and Tsybakov, 2020, Wang et al., 2020, Wasserman and Roeder, 2009, Ji and Jin, 2012]. While these approaches are efficient, they typically require specific assumptions on the covariance, e.g. bounded norms or (restricted) eigenvalues of the design matrices. While in this work, we aim for the direction of arbitrary, strong (graphical) dependence among the covariates, where these assumptions may fail in general. Finally, [Hastie et al., 2020, Guo et al., 2022b] study the effect of the signal-to-noise ratio on regression problems.

We focus on analyzing the sample complexity of exact recovery under general and structured designs. Optimality for the standard design, which considers $\Sigma = I_d$ and is related to compressed sensing [Candes and Tao, 2007, Akçakaya and Tarokh, 2009], has been derived as $\Theta(\log d/\beta_{\min}^2 \vee \log \binom{d-s}{s})$ [Rad, 2011, Fletcher et al., 2009, Aeron et al., 2010], where $\beta_{\min}$ is the assumed lower bound of nonzero entries of β (cf. Section 4.1). The optimality can be achieved by efficient methods [Wainwright, 2009b, Ndaoud and Tsybakov, 2020]. By contrast, for general or structured design, the efficient techniques no longer apply due to the dependence between variables, and the sample complexity analysis is more complicated. Existing bounds are present in Wainwright [2009a], Wang et al. [2010b], Shen et al. [2012, 2013] along with analysis of best subset selection (BSS),

but matching bounds with exact dependence on the problem parameters is not established yet. Moving beyond the known sparsity level, while consistency of information criterion-based methods is well-documented [Nishii, 1988], their finite-sample behavior, especially under general design, requires further study. Verzelen [2012] provides impossibility results for variable selection in ultra-high dimensions, but only shows results for fixed design, and the lower bound does not depend on Σ. Finally, even when Σ is known, we emphasize that general design is nontrivial, since a simple preconditioning step $X^\top \beta = (\Sigma^{-1/2} X)^\top (\Sigma^{1/2} \beta)$ can destroy sparsity in β; e.g. Kelner et al. [2022].

In graphical models, variable selection is mainly used for structure learning, i.e. estimating the underlying graph G. For undirected graphs, neighbourhood selection reduces to recovery of the precision matrix, which is well-studied [Meinshausen and Bühlmann, 2006, Wang et al., 2010a, Misra et al., 2020]. For directed acyclic graphs (DAGs), neighbourhood selection is widely used for both linear [e.g. Shojaie and Michailidis, 2010, Loh and Bühlmann, 2014, Bühlmann et al., 2014] and nonlinear [e.g. Margaritis and Thrun, 1999, Aliferis et al., 2010, Peters et al., 2014, Bühlmann et al., 2014, Azadkia et al., 2021] models. This is closely related to Markov boundary learning, for which many algorithms based on greedy search have been proposed [Tsamardinos et al., 2003b,a, Pena et al., 2007, Aliferis et al., 2010, Gao and Ji, 2016]. More recently, a growing line of work concerns ordering based DAG learning methods [Peters and Bühlmann, 2013, Ghoshal and Honorio, 2017b, Chen et al., 2019, Gao et al., 2020, Rajendran et al., 2021], which first estimates the topological ordering of the underlying DAG, then performs variable selection for each node to identify the parents. This prior work mostly focuses on consistency and upper bounds. By contrast, we explicitly focus on optimality with respect to Σ while allowing for diverging eigenvalues. In doing so, we allow for a much richer class of SEM. We mention here also that the effect of path cancellation has been noted previously [Wasserman and Roeder, 2009, Bühlmann et al., 2010, Genovese et al., 2012].

Finally, exact subset selection is known to be computationally hard [Natarajan, 1995, Foster et al., 2015]. Over the years, much progress has been made toward solving the BSS programming more efficiently. Notable developments include mixed integer programming [Bertsimas et al., 2016], coordinate descent [Hazimeh and Mazumder, 2020], and binary convex reformulation [Bertsimas and Van Parys, 2020]. In particular, Zhu et al. [2020] employs a sequencing-and-splicing technique to solve the programming in polynomial time, albeit imposing a sparse restricted condition (SRC). On the other hand along with these computational advances, gaps in sample complexity performance between efficient methods and theoretical optimality have been established for many statistical problems [Kunisky et al., 2019, Bandeira et al., 2022, Moitra and Wein, 2023], e.g. sparse PCA [Berthet and Rigollet, 2013], low-rank matrix problems [Oymak et al., 2015, Ma and Wu, 2015], and Gaussian mixture models [Diakonikolas et al., 2017]. In the context of linear model, Zhang et al. [2014] has demonstrated a gap between the minimax prediction risk and the performance achievable by any polynomial-time algorithms by a factor of restricted eigenvalue [Raskutti et al., 2010]. While in this work, we focus on the variable selection aspect of linear model.

1.2 Summary of Contributions

In the remainder of this thesis, we first present self-contained preliminaries for the main results. Specifically, Section 1.3.1 introduces graphical models and structure learning, while Section 1.3.2 formalizes the model selection framework and notions of sample complexity and statistical optimality.

In Chapter 2, we begin with parametric causal graphical models and study two identifiable classes for structure learning. Specifically, we focus on the Gaussian linear DAG model (1.1). In Section 2.2, we analyze the equal-variance DAG model and establish the first optimality result for structure learning in the literature. This is achieved by modi-

fying an existing algorithm and deriving a matching information-theoretic lower bound. This corresponds to the work in Gao et al. [2022]. In Section 2.3, we consider the classical identifiability condition of faithfulness and focus on poly-tree graphs, a rich yet tractable subclass of DAGs. We propose a new efficient algorithm, `PC-tree`, based on the classical PC algorithm, and establish its optimality with explicit dependence on the faithfulness parameter. This corresponds to the work in Wang et al. [2024]. In both settings, the optimal sample complexity exhibits a logarithmic dependence on the dimensionality / number of nodes.

In Chapter 3, we move beyond parametric assumptions and continue to study the role of the faithfulness condition in determining optimal sample complexity. Rather than relying on specific distributional properties as in Section 2.3, we develop a general reduction framework linking conditional independence (CI) testing and structure learning. This framework shows that optimality in CI testing directly translates to optimality in poly-forest learning. The approach is broadly applicable, including fully nonparametric settings, and is illustrated through several examples. This corresponds to the work in Gao et al. [2025b].

In Chapter 4, we turn to neighborhood selection, a key subroutine for structure learning in general DAGs, in order to better understand optimality in this broader setting. We again consider the linear SEM (1.1) and begin by formulating neighborhood selection as a generic variable selection problem in the linear model (4.1). As a complement to existing literature, we establish (with exact dependence to the relevant problem parameters) that Best Subset Selection (BSS) achieves optimal sample complexity under general design covariance, regardless of whether the sparsity level is known. At the same time, we derive a polynomial-time sample complexity lower bound showing that such optimality cannot, under the common conjecture in complexity theory, be achieved efficiently, thereby highlighting the central role of BSS in this setting. This corresponds to the work in Gao and

Aragam [2025]. We then return to variable selection under structured design induced by linear SEMs, and show that BSS can be improved in this setting where the underlying causal structure can be helpful. We propose a new estimator, KL-BSS, which generalizes BSS and achieves superior performance both theoretically and empirically. This corresponds to the work in Gao et al. [2025a].

Finally, Chapter 5 concludes the thesis and discusses several potential directions for future research on this topic.

1.3 Preliminaries

1.3.1 Causal Graphical Model

Given a directed acyclic graph (DAG) $G = (V, E)$, $V = \{1, 2, \ldots, d\} = [d]$ and E is the edge set. For node $k \in V$, $\text{pa}(k) = \{j : (j,k) \in E\}$ denotes the parents of node k, and $\text{ch}(k) = \{j : (k,j) \in E\}$ denotes the children of node k. A directed path is a sequence of distinct nodes $(h_1, \ldots, h_\ell)$ such that $(h_j, h_{j+1}) \in E$. Descendants $\text{de}(k)$ denotes the nodes that can be reached by k via directed path(s) and $\text{nd}(k) = V \setminus \{k\} \setminus \text{de}(k)$ denotes the nondescendants. $\text{an}(k)$ is the set of nodes that have directed path(s) to node k The skeleton of G, $\text{sk}(G)$, is the undirected graph formed by removing directions of all the edges in G. A triplet (j, ℓ, k) is called unshielded if both j, k are adjacent to ℓ but not adjacent to each other, graphically $j - \ell - k$; and is called a v-structure if additionally j, k are parents of ℓ, i.e. $j \to \ell \leftarrow k$. A path is a sequence of distinct nodes $(h_1, \ldots, h_\ell)$ such that (h_j, h_{j+1}) is in edge set of $\text{sk}(G)$. A tree is an undirected graph where any two nodes are connected by exactly one path. A poly-tree is a DAG whose skeleton is a tree. A directed tree is a poly-tree where $|\text{pa}(k)| \leq 1$ for all k. A forest is an undirected graph where any two nodes are connected by at most one path. A poly-forest is a DAG whose skeleton is a forest. Denote the set of all directed trees over d nodes as $\mathcal{T}'$. Denote the set

of all poly-trees over d nodes as $\mathcal{T}$. Denote the set of all poly-forests over d nodes as $\widetilde{\mathcal{T}}$.

A distribution p satisfies the Markov property with respect to a DAG G with d nodes if the following factorization holds

$$p(X) = p(X_1, \ldots, X_d) = \prod_{k=1}^{d} p(X_k \,|\, \mathrm{pa}(k))$$

We consider the problem of *structure learning*, also known as *causal discovery*, where we assume the existence of a DAG G such that p is Markov to G, and we aim to recover G given n i.i.d. samples from p. In general, it is well-known that the DAG is not identifiable from observational data alone. A common assumption is faithfulness, under which G is identified up to its Markov equivalence class (MEC), which is the set of DAGs that encode the same set of conditional independencies as G and is represented by completed partially directed acyclic graph (CPDAG), denoted by $\overline{G}$. We assume the reader is familiar with standard graphical concepts such as d-separation; We refer the readers to Koller and Friedman [2009] for more preliminaries on graphical models.

Definition 1 (Faithfulness)**.** We say a distribution P is faithful to a DAG G if for any $j, k \in V$ and $S \subseteq V \setminus \{j, k\}$,

$$X_j \perp\!\!\!\perp X_k \,|\, X_S \Rightarrow j \text{ and } k \text{ are d-separated by } S\,.$$

Equivalently, for any two nodes j and k not d-separated by set S, faithfulness requires $X_j \not\perp\!\!\!\perp X_k \,|\, X_S$. A standard approach to learning a CPDAG under faithfulness is the PC algorithm [Spirtes and Glymour, 1991], which relies on conditional independence testing to recover the skeleton and orient the edges. While faithfulness can be a strong assumption [Uhler et al., 2013], it is known that weaker assumptions suffice. For example:

Definition 2 (Restricted faithfulness)**.** We say a distribution P is restricted faithful to a DAG G if

1. For any $(j,k) \in E$, $S \subseteq V \setminus \{j,k\}$, $X_j \not\perp\!\!\!\perp X_k \,|\, X_S$;

2. For any unshielded triple $j - \ell - k$, if this is a v-structure, then $X_j \not\perp\!\!\!\perp X_k \,|\, S$ for any $S \subseteq V \setminus \{j,k\}$ with $\ell \in S$; if not, then $X_j \not\perp\!\!\!\perp X_k \,|\, X_S$ for any $S \subseteq V \setminus \{j,k,\ell\}$.

Under general faithfulness, all conditional independence relationships imply d-separations in a DAG. In other words, all instances of d-connections lead to conditional dependence. On the contrary, restricted faithfulness requires only a subset of d-connections to imply conditional dependence. Conventionally, the first part of Definition 2 is also named *adjacency-faithfulness* , which is used for the first step of PC algorithm, i.e. to identify the skeleton; and the second part is named *orientation-faithfulness*, which is used for the first step to return correct separation set such that the second step of PC algorithm will correctly determine whether the unshielded triples are v-structures.

In this dissertation, we will constantly work with Gaussian linear DAG model, also known as Gaussian linear structural equation model (SEM):

$$X_k = \sum_{j \in \mathrm{pa}(k)} b_{jk} X_j + \epsilon_k \qquad \epsilon_k \sim \mathcal{N}(0, \sigma_k^2) \qquad \forall k \in [d]\,. \tag{1.1}$$

See Drton [2018] for an introduction to linear SEM. Although our results extend to more general settings, we focus on it for simplicity. It follows from (1.1) that X has a multivariate Gaussian distribution $X \sim \mathcal{N}(\mathbf{0}_d, \Sigma)$ that is Markov to G; note the dependence on G in the parent sets above. More compactly, let $B = (\beta_{jk})$ denote the coefficient matrix such that $\beta_{jk} \neq 0$ is equivalent to the existence of the edge $j \to k$. Then letting $\epsilon = (\epsilon_1, \ldots, \epsilon_d)$ we have

$$X = B^\top X + \epsilon\,.$$

Denote $\Lambda = \mathrm{diag}(\sigma_1^2, \sigma^2, \ldots, \sigma_d^2)$, it follows that

$$\Sigma = (I - B)^{-T} \Lambda (I - B)^{-1} .$$

1.3.2 Model Selection and Optimal Sample Complexity

We study *structure learning* and *variable selection* (also known as *support recovery* or *neighbourhood selection* in the context of linear SEM) in this dissertation. Both problems are special cases of model selection, so are the statistical techniques used to establish optimality (cf. Appendix D.2) specialized for. Therefore, we provide a light formulation of model selection and provide the definition of (optimal) sample complexity, which is the main statistical object to study in this dissertation.

Let Θ denote a parameter space, for each $\theta \in \Theta$, let p_θ be the data distribution specified by θ. Given this parametrization, model selection can be framed as choosing a subset of Θ. Specifically, Let $\mathcal{M}$ be an index set, where each element $m \in \mathcal{M}$ corresponds to a statistical model defined by a restricted parameter space $\Theta_m \subseteq \Theta$. Suppose the data come from a distribution p_{θ^*} with $\theta^* \in \Theta_{m^*}$ for some unknown true model $m^* \in \mathcal{M}$. The goal of model selection is to design an estimator $\widehat{m}$ for the true model m^* based on the data. For two models m_1 and m_2, the restricted parameter spaces could be nested, e.g. $\Theta_{m_1} \subseteq \Theta_{m_2}$. In this case, we are interested in the "smallest" model defined by $\Theta_{m^*} := \cap_{m:\theta \in \Theta_m} \Theta_m$. The true model is necessarily identifiable for certain model selection problem, in which case, identifiability conditions are needed on Θ to ensure meaningful selection. Assuming such an identifiability, i.e. m^* could be uniquely determined from p_{θ^*} (also from θ^*), we denote the model selection problem class to be

$$\mathcal{P} = \left\{ (m, \theta) \mid m \in \mathcal{M}, \theta \in \Theta_m \right\} .$$

For structure learning problem, the model is the DAG G (or its MEC $\overline{G}$ as the estimand) that generates the distribution. For the linear SEM (1.1), the parameters include the coefficient matrix and noise variances $\theta = (B, \Lambda)$. The model effectively determines the support of B. We aim to design an estimator, which is a measurable function $\widehat{S}$ of the data X into the space of DAGs (or CPDAGs). For variable selection problem (to be formally defined in Section 4.1), the model is the support of the linear regression vector $S_* := \mathrm{supp}(\beta)$. The parameters include the regression vector, design covariance, and the noise variance, i.e. $\theta = (\beta, \Sigma, \sigma^2)$, while only β is affected by the model. The estimator is then a measurable function $\widehat{S}$ of the data (X, Y) into the power set $2^{[d]}$, i.e. $\widehat{S}(X, Y) \subseteq [d]$.

Next, we introduce the notion of statistical optimality in this dissertation. In particular, we focus on model selection as the generic problem to study. We consider the minimax optimality from the sample efficiency perspective. Specifically, the *Sample complexity* of an estimator $\widehat{m}$, denoted as $n^*(\widehat{m})$, is the minimum sample size (up to constants) such that

$$\sup_{(m,\theta)\in\mathcal{P}} \mathbb{P}_\theta(\widehat{m} \neq m) \leq \delta\,, \qquad \delta > 0\,. \tag{1.2}$$

That is, we require the probability of incorrect selection to be uniformly controlled over the problem class $\mathcal{P}$, in line with the minimax perspective. We study the dependence of the sample complexity on key problem parameters, such as the dimension d, sparsity s, and certain signal strengths of the problem. We view δ as a fixed constant and its dependence is not of primary interest.

The *optimal sample complexity* of a model selection problem is the minimum sample size (up to constants) such that there exists an estimator $\widehat{m}$ with (1.2) holds, i.e.

$$n^* = \inf_{\widehat{m}} n^*(\widehat{m})\,. \tag{1.3}$$

An estimator $\widehat{S}$ is called optimal if it achieves the optimal sample complexity in (1.3).

Characterizing the optimal sample complexity requires establishing both sufficient and necessary conditions on the sample size for uniform selection over $\mathcal{P}$. Typically, one first derives upper bounds for n^* by designing an estimator with provable guarantees, and then establishes matching lower bounds for n^* via information-theoretic techniques (see Appendix D.2). When these upper and lower bounds match (up to problem-independent constants and/or logarithmic factors of the sparsity of graph or regression vector), the resulting sample size is optimal. In particular, we do not suppress logarithmic factors of the dimension d of the problem.

1.3.3 Notations

For any nonnegative integer m, let $[m] := \{1, \ldots, m\}$. Throughout, S and T are subsets of $[d]$, write $S \triangle T = (S \setminus T) \cup (T \setminus S)$ to be the symmetric difference, and let $|S|$ be the cardinality of set S. Denote set of all possible supports of dimension d, and sparsity s to be $\mathcal{T}_{d,s} := \{S \subseteq [d] : |S| = s\}$, and bounded sparsity to be $\mathcal{T}_d^{\bar{s}} := \cup_{s=0}^{\bar{s}} \mathcal{T}_{d,s} = \{S \subseteq [d] : |S| \leq \bar{s}\}$. Let $\mathbb{S}_{++}^d$ be all positive definite matrices, $\mathbb{R}_{>0}$ be all positive real numbers. The 2-norm of a vector x is $\|x\| = (\sum_j x_j^2)^{1/2}$, the operator 2-norm of a matrix A is $\|A\| = \|A\|_{\text{op}} = \sup_{\|x\|=1} \|Ax\|$. The largest and smallest eigenvalues of A are $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$. Write x_S to be the $|S|$-dimensional sub-vector indexed by S. Similarly, for a matrix A, write A_S to be the sub-matrix with columns indexed by set S, and A_{TS} to be the sub-matrix with rows and columns indexed by T and S. Let $\mathbb{R}_+, \mathbb{Z}_+, \mathbb{S}_{++}^d$ be positive numbers, positive integers, and positive definite matrices. For a covariance matrix $\Sigma \in \mathbb{S}_{++}^d$, denote the conditional covariance matrix of the variables S given the variables T by $\Sigma_{S \mid T} := \Sigma_{(S\setminus T)(S\setminus T)} - \Sigma_{(S\setminus T)T} \Sigma_{TT}^{-1} \Sigma_{T(S\setminus T)}$. Here $\Sigma_{S \mid T}$ is of size $|S \setminus T| \times |S \setminus T|$, and when $S \cap T = \emptyset$, $|S \setminus T|$ reduces to $|S|$. For a set $\Theta \subseteq \mathbb{R}^d$, write $\Theta_S = \{\beta_S : \beta \in \Theta\}$ for the subspace of coordinates indexed by S. Let $\mathbf{1}_m, \mathbf{0}_m$ be all one's and all zero's vector of dimension m, and $\mathbb{1}\{E\}$ be the indicator of event E. Denote the support of a vector

by $\mathrm{supp}(x) = \{j : x_j \neq 0\}$. We say $a \lesssim b$ and $a \gtrsim b$ if $a \leq Cb$ and $a \geq cb$ for some positive constants C and c, and $a \asymp b$ if both $a \lesssim b$ and $a \gtrsim b$. $a \vee b$ and $a \wedge b$ are the maximum and minimum between two numbers a and b, and $\lfloor a \rfloor$ is the largest integer small than or equal to a. For remainder of the paper, with a little abuse of notation we use (X, Y) to denote the data matrix ($\mathbb{R}^{n\times d} \otimes \mathbb{R}^n$) and random variables interchangeably. Write $\Pi_S := X_S(X_S^\top X_S)^{-1}X_S^\top$ and $\Pi_S^\perp := I_n - \Pi_S$ for projection matrices onto and out of the subspace spanned by X_S.

CHAPTER 2

OPTIMAL STRUCTURE LEARNING IN PARAMETRIC MODELS

We start by studying statistical optimality for parametric causal models. We consider Gaussian linear SEM (1.1). In general, the DAG is not identifiable despite of this simple parametric assumption. Therefore, we impose additional identifiability conditions and proceed to study the optimality for each identifiable class. We mainly consider two popular identifiability conditions: equal variance DAG [Peters and Bühlmann, 2014] and faithful tree model. We specify them in the next section. All the proofs for this chapter are in Appendix A.

2.1 Problem Setup

Equal Variance DAG We consider the Gaussian linear SEM in (1.1) with equal variances. Our results will extend for more general cases, see Remark 2 for a discussion of generalizations. Specifically, consider (1.1) with additionally

$$\mathrm{var}(\epsilon_k) = \sigma_k^2 \equiv \sigma^2 \,.$$

The variance σ^2 of each noise term is assumed to be the same; this is the key identifiability assumption that is imposed on the model [Peters and Bühlmann, 2014].

To study the finite sample complexity, we consider the following equal DAG model class. The largest in-degree of any node is denoted by q, i.e.

$$q := \sup_{k \in [d]} |\mathrm{pa}(k)|. \tag{2.1}$$

The absolute values of the linear coefficients are lower bounded by $b_{\min}$, i.e.

$$b_{\min} := \min_{j,k} \{|b_{jk}| : b_{jk} \neq 0\}. \tag{2.2}$$

Furthermore, assume the covariance matrix $\Sigma = \mathbb{E}[XX^\top]$ satisfies

$$M^{-1} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq M \tag{2.3}$$

for some constant $M > 1$. Let the class of distributions satisfying the above conditions (2.1), and (2.3) be denoted by $\mathcal{F}_{d,q}$. Finally, we introduce the variance gap:

$$\Delta \equiv \min_{k \in [d]} \min_{\substack{A \subseteq \mathrm{nd}(k) \\ \mathrm{pa}(k) \setminus A \neq \emptyset \\ A \subseteq \mathrm{nd}(\mathrm{pa}(k) \setminus A)}} \mathbb{E}_A \operatorname{var}(X_k \mid A) - \sigma^2$$

where the subscript indicates that the expectation is being taken over the random variables in A. This is the missing conditional variance on ancestors if not all the parents are conditioned on, which serves as the identifiability signal for the main algorithm. The size of Δ will depend on σ^2 and $b_{\min}$.

Faithful Poly-tree The next identifiable linear SEM we consider is by assuming faithfulness Definition 1. Meanwhile, we assume the underlying graph structure is a poly-tree as a starting point towards understanding general DAGs. We introduce an estimator based on the classic PC algorithm [Spirtes and Glymour, 1991] and analyze its sample complexity under faithfulness. Crucially, we provide matching bounds with depending on a faithfulness parameter, which offers insights into the difficulty of structure learning under faithfulness.

The poly-tree structure allows us to relax the usual notion of faithfulness in Section 1.3.1:

Definition 3 (Tree-faithfulness)**.** A distribution p is tree-faithful to a poly-tree $G \in \mathcal{T}$ if

1. For two nodes connected $X_j - X_k$, we have $X_k \not\perp\!\!\!\perp X_j \mid X_\ell$ for all $\ell \in V \cup \{\emptyset\} \setminus \{k, j\}$;
2. For any v-structure $X_k \to X_\ell \leftarrow X_j$, we have $X_k \not\perp\!\!\!\perp X_j \mid X_\ell$.

Tree-faithfulness comprises two components, each corresponding to adjacency-faithfulness and orientation-faithfulness respectively in restricted faithfulness (cf. Definition 2). In comparison to adjacency-faithfulness, tree-faithfulness solely requires conditional dependence for neighbouring nodes with conditioning sets of size at most one. Likewise, compared to orientation faithfulness, tree-faithfulness only needs conditional dependence for v-structures given the the collider. Let $\rho(X_j, X_k \mid X_\ell)$ be the conditional correlation coefficient between X_k and X_j given X_ℓ, which is a valid dependence measure under Gaussianity. As usual, in order to establish uniform, finite-sample results, we need the following concept of c-strong tree-faithfulness:

Definition 4 (c-strong tree-faithfulness)**.** A distribution p is c-strong tree-faithful to a poly-tree $G \in \mathcal{T}$ if

1. For two nodes connected $j - k$, we have $\rho(X_k; X_j \mid X_\ell) \geq c$ for $\ell \in V \cup \{\emptyset\} \setminus \{k, j\}$;
2. For any v-structure $k \to \ell \leftarrow j$, we have $\rho(X_k; X_j \mid X_\ell) \geq c$.

We will study the sample complexity dependence of learning poly-trees on dimension d and signal strength c assuming strong tree-faithfulness.

2.2 Optimality in Equal Variance DAG

2.2.1 Upper Bound and Algorithm

We begin with stating the sufficient conditions on the sample size for DAG recovery under the equal variance condition. Namely, we present an algorithm (Algorithm 1) that

takes samples from a distribution $p \in \mathcal{F}_{d,q}$ as an input and returns the DAG G with high probability. We first state an upper bound for the number of samples required in Algorithm 1 in Theorem 2.2.1.

Theorem 2.2.1. *For any $F \in \mathcal{F}_{d,q}$, let $\widehat{G}$ be the DAG return by Algorithm 1 with $\gamma = \Delta/2$. If*

$$n \gtrsim \frac{M^5}{\Delta}\left(q \log \frac{d}{q} + \log \delta\right),$$

then $\mathbb{P}(\widehat{G} = G) \gtrsim 1 - \delta$.

The proof of this result can be found in Appendix A.1. The obtained sample complexity depends on the variance gap Δ, which serves as signal strength, and covariance matrix norm M, which shows up when estimating conditional variances. Treating these parameters as fixed, the sample complexity scales with $q \log d/q$. The order of this complexity arises mainly from counting all possible conditioning sets. The proof follows the correctness of Algorithm 1, which consists of two main steps: *Learning ordering* and *Learning parents*.

Algorithmically, the first step is the same as Chen et al. [2019], however, our analysis is sharper: We separately analyze the estimation of each conditional variance directly rather than indirectly via the inverse covariance matrix. This leads to the improved sample complexity in Theorem 2.2.1. This step is where we exploit the equal variance assumption: The conditional variance $\mathrm{var}(X_k \mid C)$ of each random variable X_k is a constant σ^2 if and only if $\mathrm{pa}(k) \subseteq C$ for any nondescendant set C. This implies that the variance of any non-source node in the corresponding subgraph would be larger than σ^2. Therefore, when all conditional variances v_{kC} are correctly estimated with error within some small factor of the signal Δ (see Lemma A.1.1), identifying the node with the smallest σ_k yields a source node in the underlying subgraph. Recall that σ_k is the minimum variance estimation that node k can achieve conditioned on at most q nondescendants. Finally, recursively apply-

ing the above step leads to a valid topological sort.

In the second step, given the correct ordering, we use Best Subset Selection (BSS) along with a backward phase to learn the parents for each node. Note that BSS is already applied in the step **1.(c).i.** of Algorithm 1 and the candidate set C_j can be stored for each $\widehat{\tau}_j$, thus there is no additional computational cost. Again, when all conditional variances are well approximated by their sample counterpart v_{kC}, C_j would be a superset of the true parents of current node $\widehat{\tau}_j$, otherwise the minimum would not be achieved. Meanwhile, removal of any true parent $i \in \mathrm{pa}(\widehat{\tau}_j)$ from C_j would induce a significant change in conditional variances, which is quantified by Δ as well. This is used to design a tuning parameter γ in the backward phase for pruning C_j. Finally, we show the tail probability of conditional variance estimation error is well bounded to get the desired sample complexity in Lemma A.1.2.

When the true variance gap Δ is unknown, we can select the tuning parameter γ according to the following theorem:

Theorem 2.2.2. *For any $F \in \mathcal{F}_{d,q}$, let $\widehat{G}$ be the DAG return by Algorithm 1 with tuning parameter*

$$\gamma \asymp \frac{2M^5 q \log d/q}{n}.$$

If

$$n \gtrsim \frac{M^5}{\Delta} q \log \frac{d}{q},$$

then $\mathbb{P}(\widehat{G} = G)) \gtrsim 1 - \exp(-q \log d/q)$.

The proof of this result can be found in Appendix A.2.

Remark 1. A computationally attractive alternative to BSS is the Lasso, or ℓ_1-regularized least squares regression. Unlike BSS, the Lasso requires restrictive incoherence-type conditions. If these conditions (or related conditions such as irrepresentability) are imposed on each parent set, then the Lasso can be used to recover the full DAG under a similar

Algorithm 1 LEARNDAG algorithm

Input: Sample covariance matrix $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top$, backward phase threshold γ
Output: $\widehat{G}$.

1. *Learning Ordering*:
   (a) Initialize empty ordering $\widehat{\tau} = []$
   (b) Denote $v_{kC} := \widehat{\Sigma}_{kk} - \widehat{\Sigma}_{kC}\widehat{\Sigma}_{CC}^{-1}\widehat{\Sigma}_{Ck}$
   (c) For $j = 1, 2, \ldots, d$
      i. Calculate $\sigma_k := \min_{C \subseteq \widehat{\tau}, |C| \leq q} v_{kC}$
      ii. Update $\widehat{\tau} = [\widehat{\tau}, \arg\min_k \sigma_k]$
2. *Learning Parents*:
   (a) Initialize empty graph $\widehat{G} = \emptyset$
   (b) For $j = 1, 2, \ldots, d$
      i. Let $C_j = \arg\min_{C \subseteq \widehat{\tau}_{[1:j-1]}, |C| \leq q} v_{\widehat{\tau}_j C}$
      ii. Set

$$\mathrm{pa}_{\widehat{G}}(\widehat{\tau}_j) = C_j \setminus \left\{ i \in C_j \middle| |v_{\widehat{\tau}_j C_j} - v_{\widehat{\tau}_j C_j \setminus i}| \leq \gamma \right\}$$

3. *Return* $\widehat{G}$

sample complexity scaling [see e.g. Wainwright, 2009b].

2.2.2 Lower Bound

We will now present the necessary conditions on the sample size for DAG recovery under the equal variance assumption. Namely, we present a subclass of $\mathcal{F}_{d,q}$ such that any estimator that successfully recovers the underlying DAG in this subclass with high probability requires a prescribed minimum sample size.

Theorem 2.2.3. *Assume $q \leq d/2$. If*

$$n \leq (1-2\delta)\max\left(\frac{\log d}{b_{\min}^2}, \frac{q\log d/q}{M^2-1}\right)$$

then for any estimator $\widehat{G}$,

$$\sup_{p\in\mathcal{F}_{d,q}} \mathbb{P}(\widehat{G} \neq G) \geq \delta - \frac{\log 2}{\log d}.$$

In Theorem 2.2.3, we state two sample complexity lower bounds: $q\log d/q/(M^2-1)$ and $\log d/b_{\min}^2$. Though the first one dominates when fixing other parameters as constants, the second one reveals the dependency on the signal strength $b_{\min}$. We will present two ensembles for each bound. The first one is the whole set of sparse DAGs $\mathcal{G}_{d,q}$ with d nodes and in-degree bounded by q, and the second is the set of DAGs with only one edge, which is constructed to study the dependency on the coefficient $b_{\min}$.

For the first ensemble, we borrow the ideas from Santhanam and Wainwright [2012] to count the number of DAGs in $\mathcal{G}_{d,q}$. The only difference is we consider DAGs instead of undirected graphs. Also, it is easy to bound the KL divergence between any two distributions in this ensemble due to Gaussianity, which would lead to the bound $(q\log d/q)/(M^2-1)$. For the second ensemble, it is easy to count the size of this ensemble since we consider the DAGs with only one edge. Then all possibilities of any different pair of edges are analyzed to bound the KL divergence. This ensemble gives us the bound $\log d/b_{\min}^2$. The detailed proof can be found in Appendix A.3.

For comparison, Ghoshal and Honorio [2017a] previously established a lower bound for general Gaussian DAGs (i.e. without equal variances) of

$$\Omega\Big(\sigma^2\frac{q\log d/q + q^2/d}{\sigma^2 + 2w_{\max}^2(1+w_{\max}^2)}\Big),$$

where $w_{\max}$ depends on the ℓ_2 norms of the regression coefficients. Holding σ^2 constant,

$w^2_{\max}$ is similar to the maximum marginal variance of the variables, which is comparable with our definition of M as an upper bound on $\|\Sigma\|$. By contrast, under the stronger assumption of equal variances, our lower bound is

$$\Omega\Big(\frac{\log d}{b^2_{\min}} \vee \frac{q \log d/q}{M^2 - 1}\Big),$$

which is a comparable lower bound. This is interesting since by restricting to simpler equal variance models (i.e. a smaller family), the problem should become easier, however, our analysis shows this is not the case. In particular, our lower bounds do not follow from previous work, and require a slightly different analysis as outlined in Appendix A.3.

Remark 2. Both our upper and lower bounds can be generalized as follows: Although we assume Gaussianity for simplicity, everything extends to subgaussian families without modification. This is because the upper bound analysis relies only on subgaussian concentration, and the lower bounds easily extend to subgaussian models (i.e. since subgaussian also contains Gaussian as a subclass). Furthermore, the equal variance assumption can be relaxed to more general settings as long as B can be identified by Algorithm 1. Examples include (a) the "unequal variance" condition from Ghoshal and Honorio [2018] (see Assumption 1 therein) and (b) if noise variances are known up to some ratio as in Loh and Bühlmann [2014]. Moreover, both of these identifiability conditions include the naive equal variance condition as a special case, hence the lower bounds still apply. This implies more general optimality results for a wider class of causal DAGs.

Algorithm 2 `PC-Tree` algorithm

Input: n i.i.d. samples $(X_1^{(i)}, \ldots, X_d^{(i)})$

1. Let $\widehat{E} = \emptyset$.
2. For each pair (j,k), $0 \leq j < k \leq d$:
   (a) For all $\ell \in [d] \cup \{\emptyset\} \setminus \{j,k\}$:
      i. Test $H_0 : X_j \perp\!\!\!\perp X_k \,|\, X_\ell$ vs. $H_1 : X_j \not\perp\!\!\!\perp X_k \,|\, X_\ell$, store the results.
   (b) If all tests reject, then $\widehat{E} \leftarrow \widehat{E} \cup \{j-k\}$.
   (c) Else (if some test accepts), let $S(j,k) = \{\ell \in [d] \cup \{\emptyset\} \setminus \{j,k\} : X_j \perp\!\!\!\perp X_k \,|\, X_\ell\}$.

Output: $\widehat{G} = ([d], \widehat{E})$, separation set S.

2.3 Optimality in Faithful Poly-tree

2.3.1 Upper Bound and Algorithm

We now switch to faithful poly-tree learning. Especially, we establish *exact* dependence on the problem parameter c for the optimal sample complexity. We develop the `PC-Tree` algorithm for learning poly-trees as a modification to the classical PC algorithm, outlined in Algorithm 2, effectively identifying the poly-tree's skeleton. An important by-product is the separation set resulted from the CI testing, which is used to obtain the CPDAG by applying an `ORIENT` step (Algorithm 6) as in the original PC algorithm.

In contrast to the original PC algorithm, `PC-Tree` distinguishes itself in two key aspects. Firstly, when assessing the presence of an edge between any two nodes, instead of exploring all potential conditioning sets, `PC-Tree` simplifies the process by exclusively testing marginal independence and conditional independence given only one other node. Furthermore, a notable departure from the original PC algorithm is that `PC-Tree` combines marginal independence tests and conditional independence tests, as opposed to ignoring the latter once marginal independence is established. `PC-Tree` will rely on sam-

ple (conditional) correlation coefficient for all the (conditional) independence tests when running the algorithm.

Now we are ready to provide the sample complexity of `PC-Tree` in the following theorem, whose proof is postponed to Appendix A.4.

Theorem 2.3.1. *For any $G \in \mathcal{T}$, assuming p is c-strong tree-faithful to G, apply `PC-tree` with sample correlation for CI testing, if the sample size*

$$n \gtrsim \frac{1}{c^2}\left(\log d + \log(1/\delta)\right),$$

then $\mathbb{P}(\widehat{G} = \mathrm{sk}(G)) \geq 1 - \delta$, and $\mathbb{P}(\textsc{Orient}(\widehat{G}, S) = \overline{G}) \geq 1 - \delta$

We may compare this upper bound $(\log d)/c^2$ with some of existing results on structure learning. Compared to learning equal variance general DAGs above with optimal rates being $q \log(d/q)$, tree structure simplifies the problem by removing the factor of in-degree q. As against recovering undirected graph in MRF [Misra et al., 2020], whose optimal sample complexity is $(s \log d)/\kappa^2$, we are able to improve the rate by the maximum degree s. Moreover, if $T \in \mathcal{T}' \subset \mathcal{T}$ is a directed tree, the proposition below shows c to be a constant under mild assumption on the parametrization of (1.1), which assures possible concern of dependence on c.

Proposition 2.3.2. *Consider $G \in \mathcal{T}'$ and p is parameterized using $\{\beta_k, \sigma_k^2\}_{k=1}^d$ as (1.1) according to G. If there exists a constant $M > 1$ such that for any $k \in [d]$,*

$$\begin{aligned} |\beta_{kj}| &\in [M^{-1}, M], \qquad \forall \beta_{kj} \neq 0 \\ \sigma_k^2 &\in [M^{-1}, M], \end{aligned}$$

then P is c-strong tree-faithful to G for some $c \asymp 1$.

2.3.2 Lower Bound

Having provided the sample complexity upper bound for learning faithful poly-tree, we continue to derive a matching lower bound:

Theorem 2.3.3. *Assuming c-strong tree-faithfulness, and* $c^2 \leq 1/5$, $d \geq 4$, *if the sample size is bounded as*

$$n \leq \frac{1-2\delta}{8} \times \frac{\log d}{c^2},$$

then for any estimator $\widehat{G}$ *for* $\overline{G}$,

$$\inf_{\widehat{G}} \sup_{\substack{G \in \mathcal{T} \\ p \text{ is } c\text{-strong} \\ \text{tree-faithful to } G}} \mathbb{P}(\widehat{G} \neq \overline{G}) \geq \delta - \frac{\log 2}{\log d}.$$

The lower bound in Theorem 2.3.3 implies the optimal sample complexity is $\Theta(\log d / c^2)$, where the dependence on $1/c^2$ term characterizes the hardness from "how (Tree-)faithful" the distribution is; and $\log d$ term comes from the cardinality of all poly-trees, which is much smaller compared to number of all general DAGs.

To prove this lower bound, we again employee Fano's inequality (cf. Corollary D.2.3) and consider a subclass of $\mathcal{T}'$ to exploit the property that any node in directed tree has at most one parent. This subclass of directed trees has large enough cardinality by Cayley's formula of undirected trees. With the parametrization of edge weights appropriately calibrated, we show the KL divergence between the distributions consistent with any two instances from the subclass is well controlled, which leads to the final lower bound. The detailed proof can be found in Appendix A.6.

Remark 3. The optimality results in this section also extend to directed tree, polyforest and Markov chain. Since the lower bound is constructed using directed trees, the opti-

mality applies. For polyforest, which is essentially poly-tree but allows for disconnected component, `PC-Tree` algorithm is able to identify the correct skeleton. On the other hand, poly-tree is a subclass of polyforest, thus the lower bound in Theorem 2.3.3 applies. These motivate our further work in Chapter 3. For Markov chain, the algorithm is modified to dismiss marginal independence test, and the lower bound construction considers all Markov chains with the same way of parametrization as in Theorem 2.3.3. All these graphical models share the optimal sample complexity $\Theta(\log d/c^2)$.

The results in this chapter reveal for the first time how the optimal sample size in structure learning depends on the problem parameters: The dimensionality d comes in as log factor, while the in-degree q (characterizing the graph sparsity) comes in directly in front of the $\log d$ (Section 2.2). Moreover, we characterize how the (strong) faithfulness assumption plays a role in the optimal sample complexity with an eye on the tree learning (Section 2.3).

CHAPTER 3

NONPARAMETRIC MODELS IN TREES: REDUCTION FROM CONDITIONAL INDEPENDENCE TESTING

The optimal sample complexity established in previous section looks at the parametric models and exploits specialized techniques in the analysis. Now we move to the nonparametric models. In this case, we want to avoid such a exploitation of parametric properties. We seek for a general recipe to study the optimality in a nonparametric way. We pursue this by looking at CI testing, which is mentioned to be very relevant to structure learning and serves as important building block for constraint-based methods. Meanwhile, CI testing as an important topic in statistics, it has been extensively studied and many optimality results are established. However, its relation with structure learning is often understood from the algorithmic perspective, but not from the perspective of sample efficiency. So the question is, can we leverage the existing optimality results in CI testing, hopefully in a black-box way, to help us study the optimality in structure learning.

3.1 Problem Setup

To establish reduction between CI testing and structure learning in terms of sample complexity, we formally define these two problems in this section. As in Section 2.3, we stick with tree structures, but extending to poly-forest, a more general subclass of DAG. To derive uniform, finite-sample bounds, we also need conditions on the minimum signal strength, as is standard in model selection and testing literature. To make our result applicable in the fully general and nonparametric setting, we first define a generic notion of dependence measure that will be used to quantify the signal strength.

Definition 5 (Dependence measure)**.** Let $(X, Y, Z) \sim p$ be a triplet of random variables subject to some distribution p. A *(conditional) dependence measure* $m(X; Y \mid Z)$ is a func-

tional of p such that (1) $m(X; Y \mid Z) \geq 0$; and (2) $m(X; Y \mid Z) = 0$ if and only if $X \perp\!\!\!\perp Y \mid Z$ under p.

If $Z = \emptyset$, we simplify the notation by writing the *marginal dependence* as $m(X; Y) = m(X; Y \mid \emptyset)$. This general measure of dependence is used to simplify our main theorem statement; in our examples (Section 3.3), we specify the dependence measure as the usual correlation coefficient for Gaussian distributions, total variation with the product of marginals for Bernoulli distributions, and total variation distance to the nearest independent instance for nonparametric continuous distributions [Canonne et al., 2018, Neykov et al., 2021, Wang et al., 2024]. In more general settings (e.g. user-defined models), the dependence measure can be chosen based on modeling preference, as long as there exist consistent or efficient CI tests achieving provable error guarantees, it can be embedded into our framework. Using this dependence measure, we can now define strong tree-faithfulness (Definition 4) in a generic form with adaption to poly-forest below:

Definition 6 (c-strong tree-faithfulness under m)**.** A distribution p is c-strong tree-faithful to a poly-forest $G \in \widetilde{\mathcal{T}}$ with respect to the dependence measure m if

1. For two nodes connected $j - k$, we have $m(X_k; X_j \mid X_\ell) \geq c$ for $\ell \in V \cup \{\emptyset\} \setminus \{k, j\}$;
2. For any v-structure $k \to \ell \leftarrow j$, we have $m(X_k; X_j \mid X_\ell) \geq c$.

Since we aim to establish a statistical connection between CI testing and structure learning, we define each problem and the associated statistical quantities of interest as follows. In particular, given the close relationship between these two problems and the fact that structure learning often relies on CI testing as a subroutine, we first introduce the CI testing problem before defining the poly-forest learning problem.

We formally introduce the CI testing problem in a generic form, which can be instantiated under various distributional settings. Based on the elements of CI testing, we will then proceed to define the poly-forest learning problem such that the relationship

between the two problems is explicit and clear. At its core, CI testing is a fundamental statistical problem concerned with distinguishing distributions over triplet of variables (X, Y, Z).

Definition 7 (Conditional independence testing)**.** A conditional independence testing problem $\mathcal{C}(\mathcal{P}, m)$ is defined by a class of distributions $\mathcal{P}$ and dependence measure m, and aims to distinguish two hypotheses of distributions:

$$\begin{aligned} \mathcal{H}_0: &\quad p(X,Y,Z) \quad \text{s.t. } m(X;Y \mid Z) = 0 \iff X \perp\!\!\!\perp Y \mid Z \\ \mathcal{H}_1: &\quad p(X,Y,Z) \quad \text{s.t. } m(X;Y \mid Z) \geq c \end{aligned}$$

where $p \in \mathcal{P}$, and c is the signal strength to distinguish the two hypotheses. A conditional independence test ψ is a function that takes n i.i.d. samples from p and outputs a binary decision, i.e. $\psi : \{(X^{(i)}, Y^{(i)}, Z^{(i)})\}_{i=1}^{n} \mapsto \{0, 1\}$ where $\psi = 0$ indicates selecting the null $\mathcal{H}_0$. Fixing the sample size n, the minimax optimal testing radius of $\mathcal{C}(\mathcal{P}, m)$ is the infimum of $c = c_n > 0$ in terms of n (up to constants) such that there exists a test whose Type-I and Type-II errors are controlled:

$$\inf_{\psi} \left\{ \sup_{p \in \mathcal{H}_0} \mathbb{E}_p[\psi] + \sup_{p \in \mathcal{H}_1} \mathbb{E}_p[1 - \psi] \right\} \leq \frac{1}{10}.$$

For a certain CI testing problem, one needs to specify the model class $\mathcal{P}$, and the dependence measure m. Consequently, the hypothesis classes $\mathcal{H}_0, \mathcal{H}_1$ are determined. One goal of studying $\mathcal{C}(\mathcal{P}, m)$ is to derive the minimax testing radius c_n such that the conditionally dependent and independent instances can be distinguished accurately. Crucially, the hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, together with the dependence measure m, can be used to characterize the c-strong tree-faithfulness (Definition 6). We proceed to introduce the poly-forest learning problem using this set-up.

Definition 8 (Poly-forest learning problem)**.** A poly-forest learning problem $\mathcal{F}(\mathcal{P}, m, c)$ is defined by the following family of distributions (i.e. graphical model):

$$\mathcal{F} = \Big\{ (p, G) \;\Big|\; p \text{ is Markov and } c\text{-strong tree-faithful to } G \in \widetilde{\mathcal{T}} \text{ with respect to } m, \text{ and } \forall j, k, \ell \in [d], p_{X_j, X_k, X_\ell} \in \mathcal{H}_0 \text{ or } \mathcal{H}_1 \Big\}.$$

Recall that $\mathcal{H}_0$ and $\mathcal{H}_1$ are the conditionally independent and dependent hypotheses in the distribution class $\mathcal{P}$, and $\mathcal{H}_1$ depends on parameter c. The goal is to learn the Markov equivalence class of G given i.i.d. samples from p. We are interested in the optimal sample complexity (defined in (1.3)) of learning $\mathcal{F}(\mathcal{P}, m, c)$ in terms of the number of nodes d and the strong tree-faithfulness parameter c.

The poly-forest model is defined by requiring the marginal distributions of all node triplets to fall into the hypothesis classes $\mathcal{H}_0$ or $\mathcal{H}_1$, which follow from $\mathcal{P}$ and m through $\mathcal{C}(\mathcal{P}, m)$ defined in Definition 7. This requirement is minimal as it essentially repeats the c-strong tree-faithfulness while defining p to share the distributional properties of the distributions $\mathcal{P}$ considered in the CI testing problem, e.g. Gaussianity or smoothness. We focus on the optimal sample complexity of recovering the CPDAG for the poly-forest in $\mathcal{F}(\mathcal{P}, m, c)$. Compared to CI testing problem, where the minimax testing radius is the statistical object to study, the distinction lies in which parameter (of the sample size n and signal strength c) is treated as fixed and which quantity is being studied/optimized. These roles can of course be inverted if desired, however, we chose to present these the way they are usually presented. The number $1/10$ in Definitions 7 or the δ in Definition 8 (cf. (1.3)) can be replaced with any fixed small constant independent with other problem parameters, e.g. 0.05, as its dependence is not of primary interest.

3.2 Reduction from CI Testing to Structure Learning

We present our main result below, establishing a fundamental connection between CI testing $\mathcal{C}(\mathcal{P}, m)$ and structure learning $\mathcal{F}(\mathcal{P}, m, c)$.

Theorem 3.2.1. *Given a conditional independence testing problem $\mathcal{C}(\mathcal{P}, m)$ with an optimal test ψ achieving the minimax testing radius $c_n \asymp n^{-1/\alpha}$, if there exist hard instances $p_0 \in \mathcal{H}_0$ and $p_1 \in \mathcal{H}_1$ that are Markov and c-strong tree-faithful, then the optimal sample complexity of learning $\mathcal{F}(\mathcal{P}, m, c)$ is*

$$n \asymp \frac{\log d}{c^\alpha}, \tag{3.1}$$

which is achieved by `PC-tree` *with ψ.*

Theorem 3.2.1 establishes a fundamental statistical reduction between the two problems: the optimality of a CI testing problem implies the optimality of corresponding structure learning for poly-forests. In particular, the inherent difficulty in CI testing directly translates to the difficulty in poly-forest learning, and once the minimax solution (an optimal CI test ψ) is obtained, it can be directly plug-in for learning the poly-forest.

A full detailed statement of this result with explicit constants, technical conditions, additional discussions, and proof is provided in Theorem 3.2.1 in Appendix B.

The condition in Theorem 3.2.1 looks for two hard instances $p_0 \in \mathcal{H}_0$ and $p_1 \in \mathcal{H}_1$ that are difficult to distinguish in the sense of small KL-divergence, see (C2) for the detailed requirement. We stress that this condition is imposed merely on triplet of variables (cf. Definition 7), and is a standard step to establish lower bounds when studying the minimax optimality for CI testing. Additional requirements on the graphical properties, i.e. Markov and strong tree-faithfulness, of these instances are introduced to ensure their validity in the context of poly-forest learning. As will be shown in Section 3.3.4, this technical requirement sometimes necessitates minor modifications on existing constructions

in the literature of CI testing.

In the optimal sample complexity of structure learning (3.1), the effect of dimensionality comes in as a factor of $\log d$, and the exponent α on dependence of signal draws connection between CI testing and structure learning problems, as it directly translates from the optimal testing radius to the optimal sample complexity. For parametric distributions, we typically have $\alpha = 2$, corresponding to the parametric rate. While for nonparametric distributions, it is often the case that $\alpha > 2$ and depends on the smoothness conditions. We provide examples on both cases in Section 3.3.

For the proof of upper bound, we construct a test based on the given test ψ, and apply the median trick [Motwani, 1995] to obtain an exponential decay testing error probability bound based on the guarantee of type I and II error control (Proposition B.2.2). The construction of the new test is for the theoretical sake. In practice, one can directly apply the given ψ in `PC-tree` For the proof of lower bound, we embed the hard instances p_0 and p_1 into the local three-node structure in graphical models, and apply the Tsybakov's method (Corollary D.2.4).

Compared to our result in Section 2.3, which crucially relies on the specific properties of the Gaussian distribution and therefore does not extend to discrete or nonparametric models. In contrast, this black-box reduction between testing and structure learning is model-agnostic and applicable to general, including fully nonparametric, settings where partial correlation is no longer useful and linearity fails. It allows us to obtain minimax-optimal sample complexity guarantees without relying on specific parametric properties of the Gaussian distribution. Moreover, such reduction result highlights the fundamental connection with CI testing in the view of sample efficiency.

To apply Theorem 3.2.1 on concrete problems to obtain optimality, one needs to specify the distributional assumptions in the CI testing problem, design an optimal CI test, and find proper instances from the (in)dependent hypotheses that are close enough in

KL-divergence. In the next section, we show how to apply Theorem 3.2.1 to Bernoulli, Gaussian, and nonparametric continuous distributions. In practice, given a powerful test ψ tailored to the models satisfying certain distributional assumptions, one can directly integrate it into `PC-tree` algorithm for efficient poly-forest learning.

3.3 Applications

In this sections we illustrate by examples the broad applicability of the main optimality result (Theorem 3.2.1) through a series of representative distributions. For each example, we specify the associated model class, provide a valid CI test, and give hard instances that satisfy the condition of Theorem 3.2.1. For the sake of space, we state and discuss the key conclusions and defer full details of the model class descriptions to Appendix B.1.

3.3.1 Application to Bernoulli Distribution

We start with Bernoulli distribution. As mentioned, it suffices to define the CI testing problem to set the stage. We consider all multivariate Bernoulli distributions with dimension being three, i.e. $\mathbb{P}(X = x, Y = y, Z = z) = p(x, y, z), \forall (x, y, z) \in \{0, 1\}^3$ and $\sum_{x,y,z} p(x, y, z) = 1$. Formal details are deferred to Appendix B.1.1. We measure dependence using

$$m^{Ber}(X, Y \mid Z) = \sum_{z} p(z) \sum_{x,y} \Big| p(x, y \mid z) - p(x \mid z) p(y \mid z) \Big|$$

which is the total variation distance with the product of its marginals and is a valid dependence measure. These elements lead to CI testing problem $\mathcal{C}^{Ber}$ and the poly-forest learning problem $\mathcal{F}^{Ber}$, where *Ber* denotes "Bernoulli". $\mathcal{F}^{Ber}$ effectively contains all multivariate Bernoulli distributions with dimension d and Markov and c-strong tree-faithful

to some poly-forest $G \in \widetilde{\mathcal{T}}$:

$$(X_1, X_2 \ldots, X_d) \sim \mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_d = x_d),$$
$$\forall (x_1, x_2 \ldots, x_d) \in \{0,1\}^d .$$

Now we proceed to construct a test based on thresholding the estimation of the dependence measure by the sample counterpart, inspired by the classical χ^2 independence test:

$$\widehat{m}^{Ber}(X, Y \mid Z) = \sum_z \widehat{p}(z) \sum_{x,y} \left| \widehat{p}(x, y \mid z) - \widehat{p}(x \mid z)\widehat{p}(y \mid z) \right| \tag{3.2}$$

$$\psi^{Ber} = \mathbb{1}\{\widehat{m}^{Ber}(X, Y \mid Z) \geq c/2\} \tag{3.3}$$

where $\widehat{p}(x) = \sum_i \mathbb{1}\{X_i = x\}/n$ for $x \in \{0,1\}$ and other estimates are analogously defined.

For the hard instance constructions, we consider $p_0^{Ber}(X, Y, Z)$ under which X, Y, Z are independent Bern($\frac{1}{2}$) random variables, and the alternative distribution $p_1^{Ber}(X, Y, Z)$ is given as follows:

$$Z \sim \text{Bern}\Big(\frac{1}{2}\Big), \quad X \mid Z \sim \begin{cases} \text{Bern}(\frac{1}{2} + c) & Z = 1 \\ \text{Bern}(\frac{1}{2} - c) & Z = 0 \end{cases},$$
$$Y \mid X \sim \begin{cases} \text{Bern}(\frac{1}{2} + c) & X = 1 \\ \text{Bern}(\frac{1}{2} - c) & X = 0 \end{cases}.$$

It is easy to see that p_0^{Ber} is constructed to be Markov and tree-faithful to an empty graph, while p_1^{Ber} is constructed to follow a three-node chain graph $Z \to X \to Y$, both of which are poly-forests.

Then the following theorem checks the validity of the test and construction, and con-

cludes the optimality of learning Bernoulli poly-forest via Theorem 3.2.1. See proof in Appendix B.3.

Theorem 3.3.1. *ψ^{Ber} is optimal for $\mathcal{C}^{Ber}$ with optimal testing radius being $c_n \asymp n^{-1/2}$, and the optimal sample complexity of learning $\mathcal{F}^{Ber}$ is*

$$n \asymp \frac{\log d}{c^2},$$

which is achieved by `PC-tree` *with ψ^{Ber}.*

3.3.2 Application to Gaussian Distribution

We next consider Gaussian distribution as another parametric example to illustrate the main result. Since this application will recover the established optimality in Section 2.3, we only collect the key elements and defer most details to Appendix B.1.2. We consider all multivariate Gaussian distributions with dimension being three and measure dependence using the partial correlation coefficient (cf. (B.1)). The Gaussian CI testing problem $\mathcal{C}^{Gauss}$ and the Gaussian poly-forest learning problem $\mathcal{F}^{Gauss}$ are defined, where *Gauss* represents "Gaussian".

We employ a valid CI test ψ^{Gauss} similar to ψ^{Ber} by thresholding the sample partial correlation (cf. (B.2)). For the hard instances, let p_0^{Gauss} be $\mathcal{N}(\mathbf{0}_3, I_3)$ and p_1^{Gauss} be generated as

$$\begin{aligned} Z \sim \mathcal{N}(0,1), \quad X = \beta Z + \mathcal{N}(0, 1-\beta^2)\,, \\ Y = \beta X + \mathcal{N}(0, 1-\beta^2)\,, \end{aligned}$$

where $\beta = 2c$. Likewise in the Bernoulli case, p_0^{Gauss} is constructed to be Markov and tree-faithful to an empty graph, and p_1^{Gauss} is constructed for the chain graph $Z \to X \to Y$. It

remains to check the validity of ψ^{Gauss}, p_0^{Gauss} and p_1^{Gauss}. Applying Theorem 3.2.1, we have the optimality of learning Gaussian poly-forest (proof is in Appendix B.4):

Theorem 3.3.2. *ψ^{Gauss} is optimal for $\mathcal{C}^{Gauss}$ with optimal testing radius being $c_n \asymp n^{-1/2}$, and the optimal sample complexity of learning $\mathcal{F}^{Gauss}$ is $n \asymp \log d / c^2$, which is achieved by `PC-tree` with ψ^{Gauss}.*

3.3.3 Extension to Optimal Poly-tree Learning

The optimality results for both Bernoulli and Gaussian distributions can be extended to poly-tree learning, whose problem definition simply replaces the poly-forest in Definition 8 by poly-tree. Since poly-tree is a sub-class of poly-forest, the upper bound result of `PC-tree` applies. Then it suffices to derive a lower bound to match $\log d/c^2$ dependence, which requires the construction to be Markov and faithful to not just poly-forest, but further poly-tree.

Here we show the same extension holds for Bernoulli poly-tree due to the parametric nature. The main difficulty lies in the lower bound construction, for which we cannot directly adopt the construction in Theorem 2.3.3 due to the discrepancy between continuous and discrete distributions. While analogously, we consider all possible directed Markov chains in form of $X_{\pi_1} \to X_{\pi_2} \to \dots X_{\pi_d}$ for some permutation π, which is a subset of poly-trees, and specify the construction as follows: For each Markov chain T, consider P_T given by $X_{\pi_1} \sim \text{Bern}(1/2)$ and

$$X_{\tau_k} \,|\, X_{\pi_{k-1}} \sim \begin{cases} \text{Bern}(\frac{1}{2}+c) & X_{\pi_{k-1}} = 1 \\ \text{Bern}(\frac{1}{2}-c) & X_{\pi_{k-1}} = 0 \end{cases}, \qquad \forall k = 2, 3, \cdots, d\,. \tag{3.4}$$

The idea is to add small perturbation according to the amount of faithfulness parameter such that the KL divergence between distributions induced by different Markov chains

can be bounded. The validity of this construction is due to Lemma B.5.1 and B.5.2 proved in Appendix B.5. We conclude the optimality below:

Theorem 3.3.3. *The optimal sample complexity of Bernoulli or Gaussian poly-tree learning is* $n \asymp \log d/c^2$*, which is achieved by* `PC-tree` *with* ψ^B *or* ψ^G*.*

3.3.4 Application to Nonparametric Continuous Distribution

For both Bernoulli and Gaussian distributions, the optimal sample complexity for learning poly-forest is $\Theta(\log d/c^2)$, i.e. $\alpha = 2$ in Theorem 3.2.1. This result primarily arises from the parametric nature of these distributions. To explore the implications beyond the parametric setting, we now shift our focus to a nonparametric continuous distribution. This allows us to examine how the nonparametric nature influences the value of α, leading to a distinct theoretical outcome.

We adopt the framework in Neykov et al. [2021] and provide the essential elements in Appendix B.1.3, see details therein. We consider all continuous distributions over $[0,1]^3$, which admit continuous densities $p(X, Y, Z)$. We measure the dependence using

$$m^{NP}(X, Y \mid Z) = \inf_{q \in \mathcal{Q}} \|p - q\|_1 ,$$

where $\mathcal{Q}$ is the class of continuous distributions $q(X, Y, Z)$ over $[0,1]^3$ such that $X \perp\!\!\!\perp Y \mid Z$, and $\|p - q\|_1 = \int |p(x,y,z) - q(x,y,z)| dxdydz$. This measures the distance to the closest (conditionally) independent distributions. In addition, smoothness conditions are imposed on the densities (cf. Definitions 12-13), characterized by a smoothness parameter s and used to contrast with the parametric cases. Together, these defines the CI testing class $\mathcal{C}^{NP}$ and the poly-forest learning problem $\mathcal{F}^{NP}$, where NP denotes "nonparametric".

Now we proceed to verify the applicability of Theorem 3.2.1 for this nonparametric

setting. We use the minimax optimal CI test introduced in Section 5.3 of Neykov et al. [2021], denoted as ψ^{NP}, which is based on classic U-statistics to measure the (conditional) dependence between the X and Y. ψ^{NP} is a valid choice and satisfies the type I and II error controls with $\alpha = \frac{5s+2}{2s}$ (see Theorem 5.6 therein).

For the hard instances, we design the constructions, denoted as p_0^{NP} and p_1^{NP}, based on a modification of the ones used for proving lower bounds for nonparametric CI testing in Neykov et al. [2021]. Although the original constructions of p_0 and p_1 are close in KL divergence, the issue lies in the unfaithfulness of the distribution, thus they cannot be directly applied for poly-forest learning setting. Specifically, the original construction of the alternative p_1 only characterizes the conditional dependence but accidentally leads to marginal independence between variables, which cannot be faithful to any poly-forest (over triplet of nodes).

We modify the construction by adding back the marginal dependence with extra complexity in the analysis. Specifically, let p_0^{NP} be independent uniform distributions $Unif^3[0,1]$, which is Markov to the empty graph. For p_1^{NP}, we consider a V-structure $X \to Z \leftarrow Y$, which is a three-node poly-forest. Under this graph, a faithful distribution is supposed to have $X \perp\!\!\!\perp Y$ while $X \not\perp\!\!\!\perp Y \,|\, Z$ and $X \not\perp\!\!\!\perp Z, Y \not\perp\!\!\!\perp Z$. Thus p_1^{NP} is specified as follows: $X, Y \sim Unif[0,1]$ and $p_{Z \,|\, X,Y}$ is given by a perturbation to the uniform depending on X and Y:

$$p(Z \,|\, X, Y) = 1 + \gamma_\Delta(X, Y)\eta_\nu(Z)\,,$$

for some functions γ_Δ and η_ν governing the perturbation, which are specified in the proof. This construction is in the same form as the one in Neykov et al. [2021]. However, the functions are constructed such that the X (or Y) and Z are marginally dependent, thus faithfulness is satisfied. See more details about these constructions in the proof (Appendix B.6). Applying Theorem 3.2.1, we have the optimality of learning nonparametric

continuous poly-forest.

Theorem 3.3.4. *ψ^{NP} is optimal for $\mathcal{C}^{NP}$ with optimal testing radius being $c_n \asymp n^{-2s/(5s+2)}$, and the optimal sample complexity of learning $\mathcal{F}^{NP}$ is*

$$n \asymp \frac{\log d}{c^{\frac{5s+2}{2s}}},$$

which is achieved by `PC-tree` *with ψ^{NP}.*

Theorem 3.3.4 highlights the larger sample complexity of nonparametric setting compared to the parametric one for poly-forest learning. The difference lies in the dependence on the strong tree-faithfulness parameter c, where the resulting exponent $\alpha = \frac{5s+2}{2s} > 2$ comes directly from the intrinsic hardness of nonparametric CI testing. Moreover, the modification for p_0^{NP} and p_1^{NP} illustrates the additional complexity in the analysis to ensure faithfulness.

In summary, we establish a theoretical connection between CI testing and structure learning through the lens of sample complexity. In particular, we show that optimality guarantees for a local problem (CI testing) can be directly translated to a global problem (structure learning) via a reduction. This connection is both powerful and broadly applicable, although our current results are established for poly-forests, a class of tree-structured graphs. In the next chapter, we move towards the general DAGs by developing a deeper understanding of neighborhood selection, another important subroutine of structure learning.

CHAPTER 4

TOWARD GENERAL STRUCTURES: RETHINKING VARIABLE SELECTION IN CAUSAL GRAPHICAL MODELS

In this chapter, we move toward causal structure learning in general DAGs. A natural first step is to understand *neighborhood selection*, a basic subroutine of structure learning that aims to identify the set of neighbors of a target node in the graph. Arguably, accurate neighborhood selection is a necessary condition for successful structure learning: Once the underlying graph is correctly recovered, the neighborhood of each node can be directly inferred from the graph structure. Consequently, any statistical difficulty in neighborhood selection necessarily translates to the structure learning problem. We therefore focus on understanding the statistical properties of neighborhood selection.

We begin by revisiting variable selection in linear regression models, a classical problem that has been extensively studied in statistics. Despite its simplicity and decades of prior work, the optimal sample complexity under general design conditions remains not fully understood. We close this gap and further show a statistical-computational tradeoff in this problem. We then consider the linear SEM in (1.1), under which neighborhood selection can be formulated as a variable selection problem with a structured design covariance. We analyze the sample complexity in this setting and show that this additional *unknown* structure can help to achieve improved performance.

4.1 Problem Setup

We begin by introducing the variable selection problem in linear regression in a general form, and then formulate neighborhood selection within this framework. Specifically, we

consider the standard linear model with Gaussian noise:

$$Y = X^\top \beta + \epsilon\,, \qquad X \sim \mathcal{N}(\mathbf{0}_d, \Sigma)\,, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2)\,, \qquad X \perp\!\!\!\perp \epsilon. \tag{4.1}$$

The coefficient vector β is sparse in the sense that $\|\beta\|_0 = s \le \bar{s}$. We assume exact knowledge of s in Sections 4.2.1 and 4.3, then relax to unknown sparsity setting in Section 4.2.2 where only an upper bound $\bar{s} \ge s$ is provided. Results in Section 4.3 can be extended to unknown sparsity as well [Gao et al., 2025a].

Under the general design, we put the most basic assumptions on β and Σ and study the optimal dependence on the signal strength implied by these assumptions. Specifically, we consider the following parameter spaces:

$$\Theta_{d,s}(\beta_{\min}) := \Big\{\beta \in \mathbb{R}^d : \|\beta\|_0 = s, \min_{j \in \mathrm{supp}(\beta)} |\beta_j| \ge \beta_{\min} > 0\Big\},$$
$$\Omega_{d,s}(\omega) := \Big\{\Sigma \in \mathbb{S}^d_{++} : \min_{S,T \in \mathcal{T}_{d,s}} \lambda_{\min}(\Sigma_{S \setminus T \,|\, T}) \ge \omega > 0\Big\}.$$

The space $\Theta_{d,s}(\beta_{\min})$ consists of all sparse vectors with exact s many nonzero entries, and each of them is bounded away from zero by at least $\beta_{\min}$, which measures the signal for recovery and is commonly assumed in literature [van de Geer et al., 2011]. The covariance matrix space $\Omega_{d,s}(\omega)$ imposes a lower bound on the minimum eigenvalue of the conditional covariance over subsets of size s, which essentially requires that the variance in X_S to not be fully explained by X_T, otherwise S and T would be indistinguishable. For the special case of standard design ($\Sigma = I_d$), $\min_{S,T \in \mathcal{T}_{d,s}} \lambda_{\min}(\Sigma_{S \setminus T \,|\, T}) = 1$ for all $S, T \in \mathcal{T}_{d,s}$. We can extend these parameter spaces to the unknown sparsity setting as follows:

$$\Theta^{\bar{s}}_d(\beta_{\min}) := \Big\{\beta \in \mathbb{R}^d : \|\beta\|_0 \le \bar{s}, \min_{j \in \mathrm{supp}(\beta)} |\beta_j| \ge \beta_{\min} > 0\Big\},$$
$$\Omega^{\bar{s}}_d(\omega) := \Big\{\Sigma \in \mathbb{S}^d_{++} : \min_{S,T \in \mathcal{T}^{\bar{s}}_d, S \not\subseteq T} \lambda_{\min}(\Sigma_{S \setminus T \,|\, T}) \ge \omega > 0\Big\}.$$

We enlarge the β vector space by relaxing the exact sparsity to being upper bounded by $\bar{s}$. For Σ space, we relax the sizes of the candidate subsets while requiring one is not fully contained in the other. In addition, we denote

$$\begin{aligned} \mathcal{M} &:= \left\{(\beta, \Sigma, \sigma^2) : \beta \in \Theta_{d,s}(\beta_{\min}), \Sigma \in \Omega_{d,s}(\omega)\right\} \\ \overline{\mathcal{M}} &:= \left\{(\beta, \Sigma, \sigma^2) : \beta \in \Theta_d^{\bar{s}}(\beta_{\min}), \Sigma \in \Omega_d^{\bar{s}}(\omega)\right\}, \end{aligned} \tag{4.2}$$

to be the model classes of known and unknown sparsity settings. We often suppress the dependence on $(d, s, \bar{s}, \beta_{\min}, \omega, \sigma^2)$ to avoid notation clutter.

We aim to estimate the support of β, denoted as $S_* := \mathrm{supp}(\beta)$. A support estimator $\widehat{S}$ is a measurable function of i.i.d. observations (X, Y) to the power set of $[d]$. As in (1.3), we study the optimal sample complexity of variable selection in terms of the problem parameters $d, s, \sigma^2, \beta_{\min}, \omega$ such that the error probability of exact recovery of S_* is upper bounded by any small constant $\delta > 0$ uniformly over $\mathcal{M}$ (or $\overline{\mathcal{M}}$):

$$\sup_{(\beta,\Sigma,\sigma^2)\in\mathcal{M}} \mathbb{P}(\widehat{S} \neq S_*) \leq \delta\,. \tag{4.3}$$

In addition, as commonly imposed in literature, we assume $s \leq \bar{s} \leq d/2$ to simplify results.

Remark 4. Crucially, unlike the common assumptions in the literature for general design, which impose various types of bounded "eigenvalues", we do *not* treat either problem parameter $\beta_{\min}$, ω or σ^2 as fixed constants; instead, we are interested in studying how selection performance in terms of sample complexity depends on these quantities.

Finally, an important quantity in our result is the restricted eigenvalue (RE) of the design matrix, which appears extensively in previous work on ℓ_1-penalized estimators and related methods. It relaxes the typical dependence of the estimation error on the minimum eigenvalue of the empirical covariance (which vanishes when $n > d$). We

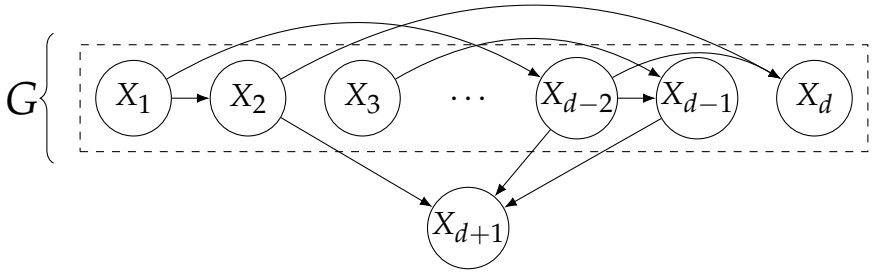


Figure 4.1: A graphical model over $X = (X_1, \ldots, X_d)$ with one more target node X_{d+1} appended to it. The corresponding G will refer to a DAG over X (ignoring X_{d+1}). The neighbourhood of X_{d+1} is $\{X_2, X_{d-2}, X_{d-1}\}$ in this example.

recall the definition here:

Definition 9. The RE constant of a design matrix $X \in \mathbb{R}^{n \times d}$ is

$$\gamma(X) := \min_{S \in \mathcal{T}_{d,s}} \min_{\|\theta_{S^c}\|_1 \leq 3\|\theta_S\|_1} \frac{\|X\theta\|^2/n}{\|\theta\|^2}.$$

Neighbourhood Selection in Linear SEM A basic primitive of structure learning is neighbourhood selection: Learning the Markov boundary S (often called the *neighbourhood*) of each node X_k with respect to a set $A \subseteq V$ such that $X_k \perp\!\!\!\perp X_{A \setminus S} \mid X_S$. This is known to reduce to recovering the support of regression model of X_k on X_A in SEM; see Appendix C.5. Therefore, to model the variable selection problem in SEM, we append one more node X_{d+1} to the graph by directing edges from a subset of $X_1, \ldots, X_d$ to it. We treat X_{d+1} as the target node and aim to learn the neighbourhood with respect to $X_1, \ldots, X_d$. See Figures 4.1 for an illustration. To further align the notation, we will denote the target X_{d+1} by Y, and denote the set of candidate variables $X_1, \ldots, X_d$ by $X = (X_1, \ldots, X_d)$. Thus, the problem reduces to a prototypical regression problem between Y and X (4.1). Furthermore, we will let G be the DAG over the variables X (i.e. ignoring Y), since it is easy to obtain the full DAG by adding the node Y and edges $X_k \to Y$ for $k \in \mathrm{pa}(Y)$. This implies X are the nondescendants of Y in the full DAG, and thus the neighbourhood becomes the parents of Y.

Formally, we are effectively considering the linear regression (4.1) with a particular class of the covariance space that arise from linear SEMs, defined as following:

$$\Omega = \Omega_d(\sigma^2_{\min}) := \Big\{ \operatorname{cov}(X) : X \text{ is generated by (1.1) for some DAG } G \\ \text{and } \sigma_k^2 \geq \sigma^2_{\min}, \forall k \in [d] \Big\}. \tag{4.4}$$

Similar to the beta-min condition on $\Theta = \Theta_{d,s}(\beta_{\min})$ and the eigenvalue condition on $\Omega_{d,s}(\omega)$, the constraint of $\sigma^2_{\min}$ also quantifies the signal for successful neighbourhood selection. When $\sigma^2_{\min} \to 0$, observe that Ω collapses to all of $\mathbb{S}^d_{++}$, paralleling the behavior of $\Omega_{d,s}(\omega)$ as $\omega \to 0$. The key distinction between Ω and $\Omega_{d,s}(\omega)$ lies in the particular structural parametrization of the covariance induced by the linear SEM. The space Ω is also treated as unknown; this only arises in the analysis and is not directly used by our method (cf. Remark 8). We are interested in conditions on $\Sigma \in \Omega$ under which BSS can be improved. Finally, define the model space for neighbourhood selection (NS) in linear SEM by

$$\mathcal{M}^{\mathrm{NS}} := \Big\{ (\beta, \Sigma, \sigma^2) : \beta \in \Theta_{d,s}(\beta_{\min}), \Sigma \in \Omega_d(\sigma^2_{\min}) \Big\}. \tag{4.5}$$

4.2 Optimality and Computational Barriers in Variable Selection

We start with studying variable selection under general design. The optimality is established for both known and unknown sparsity cases (Sections 4.2.1-4.2.2), achieved by BSS. Finally, we show a polynomial-efficient sample complexity lower bound in Section 4.2.3, which emphasizes the optimal position of subset selection, and the critical role played by restricted eigenvalues.

4.2.1 Optimality of BSS for Variable Selection When Sparsity Is Known

In this and the next section, we derive the optimal sample complexity for variable selection in linear models under general design with known and unknown sparsity. To achieve this, we derive a new lower bound to match existing upper bounds. For completeness, we begin by reviewing one such upper bound before detailing our lower bound. First recall the definition of BSS:

$$\widehat{S}^{\mathrm{BSS}} := \operatorname*{arg\,min}_{S \in \mathcal{T}_{d,s}} \|\Pi_S^{\perp} Y\|^2 , \tag{4.6}$$

which estimates the true support by minimizing the residual variance of Y over all possible supports $\mathcal{T}_{d,s}$. Recall that $\Pi_S^{\perp}$ is the projection matrix out of the column space of X_S.

For the setting with the knowledge of $s = \|\beta\|_0$, we start with defining the generic signal to distinguish S_* from any other alternatives T to be:

$$\Delta_1 := \min_{T \in \mathcal{T}_{d,s} \setminus \{S_*\}} \frac{1}{|S_* \setminus T|} \frac{\beta_{S_* \setminus T}^{\top} \Sigma_{S_* \setminus T \,|\, T} \beta_{S_* \setminus T}}{\sigma^2} . \tag{4.7}$$

We can find the intuition of this signal by looking at the numerator, which is the expected residual variance contributed by S_* but not fully captured by the alternative support T, and depends on the difference set $(S_* \setminus T)$. The denominator is the noise variance of ϵ. Since BSS minimizes the residual variance for variable selection, (4.7) measures how much signal-noise-ratio that BSS can exploit. Then we have the following generic lemma on statistical guarantee of BSS:

Lemma 4.2.1. *Assuming $s \leq d/2$, for any $(\beta, \Sigma, \sigma^2) \in \mathcal{M}$, let $S_* = \mathrm{supp}(\beta)$, given n i.i.d.*

samples from $P_{\beta,\Sigma,\sigma^2}$, if the sample size

$$n \gtrsim \max_{\ell \in [s]} \frac{\log \binom{d-s}{\ell} + \log(1/\delta)}{(\ell \Delta_1) \wedge 1}, \tag{4.8}$$

then $\mathbb{P}_{\beta,\Sigma,\sigma^2}(\widehat{S}^{BSS} = S_*) \geq 1 - \delta$.

The proof is postponed to Appendix C.1. We emphasize the key idea behind the proof lies in the scaling factor $|S_* \setminus T|$ in (4.7): the signal to distinguish S_* and T is actually proportional to $|S_* \setminus T|$. This means when T deviates from S_* a lot (by the number of missing true covariates), it is easier to tell them apart. At the same time, the total number of alternatives T to S_* also grows with their difference $|S_* \setminus T|$. Therefore, these two effects cancel each other, leading to the desired sample complexity $\log(d-s)/\Delta_1 \vee \log \binom{d-s}{s}$. By applying Lemma 4.2.1 to $\mathcal{M}$, we obtain an upper bound on the sample complexity for this model class, which was first described in Wainwright [2009a]:

Theorem 4.2.2 (Wainwright, 2009a, Theorem 1)**.** *Assuming $s \leq d/2$, for any $(\beta, \Sigma, \sigma^2) \in \mathcal{M}$, given n i.i.d. samples from $P_{\beta,\Sigma,\sigma^2}$, if the sample size*

$$n \gtrsim \max \left\{ \frac{\log(d-s) + \log(1/\delta)}{\beta_{\min}^2 \omega / \sigma^2}, \log \binom{d-s}{s} + \log(1/\delta) \right\}, \tag{4.9}$$

then $\mathbb{P}_{\beta,\Sigma,\sigma^2}(\widehat{S}^{BSS} = S_*) \geq 1 - \delta$.

Proof. The proof is given by the following chain of inequalities: For any $(\beta, \Sigma, \sigma^2) \in \mathcal{M}$ and any $T \in \mathcal{T}_{d,s} \setminus \{S_*\}$, we have

$$\begin{aligned} \beta_{S_* \setminus T}^\top \Sigma_{S_* \setminus T \,|\, T} \beta_{S_* \setminus T} &\geq \|\beta_{S_* \setminus T}\|^2 \lambda_{\min}(\Sigma_{S_* \setminus T \,|\, T}) \\ &\geq |S_* \setminus T| \beta_{\min}^2 \lambda_{\min}(\Sigma_{S_* \setminus T \,|\, T}) \geq |S_* \setminus T| \beta_{\min}^2 \omega, \end{aligned}$$

which yields $\Delta_1 \geq \beta_{\min}^2 \omega / \sigma^2$ and completes the proof. □

Theorem 4.2.2 establishes the upper bound $\log(d-s)/(\beta_{\min}^2\omega/\sigma^2) + \log\binom{d-s}{s}$ for variable selection and provides characterization of the dependence on (β, Σ) via $\beta_{\min}$ and ω separately. $\Theta_{d,s}(\beta_{\min})$ requires each variable in S_* has large enough effect on Y, and $\Omega_{d,s}(\omega)$ demands for any two distinct supports S and T, the variables therein cannot fully explain each other. Both parameters are indispensable for the uniform consistency of successful variable selection.

Now we switch gears to obtain lower bounds for the risk over $\mathcal{M}$ to match the ones in Theorem 4.2.2. The lower bound provided in Theorem 2 of Wainwright [2009a] is

$$n \gtrsim \max\left\{\frac{\log\binom{d}{s}}{\omega_{bu}\beta_{\min}^2/\sigma^2}, \frac{\log(d-s)}{\omega_{ave}\beta_{\min}^2/\sigma^2}\right\},$$

where

$$\omega_{bu} := \mathbb{E}_S[\min_{z_S\in\mathbb{R}^s, |z_j|\geq 1, \forall j} z_S^\top \Sigma_{SS} z_S]$$

$$\omega_{ave} := \mathbb{E}_S[\min_{t\in S} \min_{z_u: u\in\{t\}\cup S^c, |z_u|\geq 1/\sqrt{2}} \sum_{u,v\in\{t\}\cup S^c} (\Sigma_{uu}z_u^2 + \Sigma_{vv}z_v^2 - 2\Sigma_{uv}z_u z_v)],$$

and the expectation is taken over $S \sim Unif(\mathcal{T}_{d,s})$. However, this lower bound is derived using different set of parameters ω_{bu} and ω_{ave} rather than ω, which do not exactly match the upper bound in Theorem 4.2.2. This is mainly because $\omega_{ave} \geq \omega$ and $\omega_{bu}/s \geq \omega$ in general. For the special case of standard design $\Sigma = I_d$, we have $\omega_{ave} = \omega = 1$ and $\omega_{bu} = s$. Nonetheless, the difference between them can be large: A simple 2×2 covariance $\Sigma = \left[\begin{smallmatrix} 1 & b \\ b & 1+b^2 \end{smallmatrix}\right]$ with $s = 1$ and some moderate $b > 0$ leads to $\omega_{ave} \wedge \omega_{bu} \geq 1$ while $\omega = 1/(1+b^2)$. Therefore, here we want to derive a lower bound that precisely matches the upper bound with the same set of parameters, i.e. $\beta_{\min}$, σ^2, and ω.

The first lower bound construction aims to match the first term in upper bound (4.9), i.e. to characterize the dependence on the problem parameters $\beta_{\min}, \sigma^2$, and especially

ω, which is the variance that cannot be explained by variables outside of S_*. The construction is built upon a equi-correlation matrix where any pair of variables are equally correlated with correlation specified by ω, i.e. a rank one perturbation of identity matrix:

$$\Sigma_\omega := \omega I_d + (1-\omega)\mathbf{1}_d\mathbf{1}_d^\top .$$

When ω is close to zero, Σ_ω is dense in the off-diagonal entries and exhibits strong dependence among all variables X. In particular, we have $\omega_{ave} = \omega$ in this construction. Fixing this choice of Σ, we consider s many ensembles by enumerating all possible candidate supports according to their difference to the truth $|S_* \setminus T|$, and then combine the lower bounds obtained into one final lower bound. The proof is in Appendix C.2.

Theorem 4.2.3. *Assuming $\omega < 1$, given n i.i.d. samples from $P_{\beta,\Sigma,\sigma^2}$ with $(\beta,\Sigma,\sigma^2) \in \mathcal{M}$, if the sample size is bounded as*

$$\begin{aligned} n &\le \frac{1-2\delta}{2} \times \max_{\ell\in[s]} \frac{\log \binom{d-s}{\ell}}{\ell\beta_{\min}^2\omega/\sigma^2} \\ &\asymp \frac{1-2\delta}{2} \times \frac{\log(d-s)}{\beta_{\min}^2\omega/\sigma^2}, \end{aligned} \tag{4.10}$$

then for any estimator $\widehat{S}$ for $S_* = \mathrm{supp}(\beta)$,

$$\inf_{\widehat{S}} \sup_{(\beta,\Sigma,\sigma^2)\in\mathcal{M}} \mathbb{P}_{\beta,\Sigma,\sigma^2}(\widehat{S} \neq S_*) \ge \delta - \frac{\log 2}{\log \binom{d-s}{s}} .$$

To match the second term in the upper bound (4.9) that only depends on dimension parameters, we invoke Theorem 1 of Wang et al. [2010b] in Theorem 4.2.4 below. We will assume $\beta_{\min}^2/\sigma^2$ is upper bounded by some constant, but it should be presumably small. The construction in Wang et al. [2010b] fixes standard design $\Sigma = I_d$ and $\beta_{S_*} = \beta_{\min}\mathbf{1}_s$, then considers the ensemble of all possible supports $\mathcal{T}_{d,s}$. The analysis further relies on

the fact that the differential entropy of a continuous random variable is maximized by a Gaussian distribution with matched variance.

Theorem 4.2.4 (Wang et al., 2010b, Theorem 1)**.** *Given n i.i.d. samples from $P_{\beta,\Sigma,\sigma^2}$ with $\beta \in \Theta_{d,s}(\beta_{\min})$, $\Sigma = I_d$. If the sample size is bounded as*

$$n \leq 2(1-\delta) \times \frac{\log \binom{d}{s} - 1}{\log(1 + s\beta_{\min}^2/\sigma^2)}, \tag{4.11}$$

then for any estimator $\widehat{S}$ for $S_* = \mathrm{supp}(\beta)$,

$$\inf_{\widehat{S}} \sup_{\beta \in \Theta_{d,s}(\beta_{\min})} \mathbb{P}_{\beta, I_d, \sigma^2}(\widehat{S} \neq S_*) \geq \delta\,.$$

The standard design $\Sigma = I_d$ considered in Theorem 4.2.4 is a special case of general design of $\mathcal{M}$, thus the lower bound obtained also applies to $\mathcal{M}$. Combined with Theorem 4.2.3, the two lower bounds (4.10)-(4.11) match the upper bound (4.9), and we verify the folklore and conclude BSS is optimal for variable selection problem under general design with knowledge of the sparsity level.

4.2.2 Optimality of BSS for Variable Selection When Sparsity Is Unknown

Having concluded the optimal sample complexity in the known sparsity case, we now extend this result to the setting where s is unknown but has a known upper bound $\bar{s}$. Formally, we will derive new sample complexity upper bound for $\overline{\mathcal{M}}$ defined in (4.2). The minimax estimator is achieved by modifying BSS with an additive penalty depending on the dimensionality and the model parameters—similar to information criteria such as BIC—to help us target the truth S_*. The finite sample analysis for this estimator and its optimality is new to the best of our knowledge.

Given a tuning parameter τ, define an estimator BSSu (where "u" stands for "un-

known") by:

$$\widehat{S}^{\text{BSSu}} = \underset{S \in \mathcal{T}_d^{\bar{s}}}{\arg\min} \frac{\|\Pi_S^{\perp} Y\|^2}{n - |S|} + |S|\tau \,. \tag{4.12}$$

The first term is the residual variance objective of BSS and the second term is a penalty term.

Theorem 4.2.5. *Assuming $\bar{s} \leq d/2$, for any $(\beta, \Sigma, \sigma^2) \in \overline{\mathcal{M}}$, let $S_* = \text{supp}(\beta)$, given n i.i.d. samples from $P_{\beta,\Sigma,\sigma^2}$, choose $\tau = \frac{1}{4}\omega\beta_{\min}^2$, if the sample size*

$$n \gtrsim \max\left\{ \frac{\log d + \log(1/\delta)}{\beta_{\min}^2 \omega/\sigma^2}, \log\binom{d}{\bar{s}} + \log(1/\delta) \right\}, \tag{4.13}$$

then $\mathbb{P}_{\beta,\Sigma,\sigma^2}(\widehat{S}^{BSSu} = S_*) \geq 1 - \delta$.

Since the support space expands from $\mathcal{T}_{d,s}$ to $\mathcal{T}_d^{\bar{s}}$, the upper bound now has dependence on $\bar{s}$. Implicitly, ω in this setting should be perceived as smaller compared to the known sparsity setting. Because given a fixed ω, the covariance matrix spaces have a nested relationship $\Omega_d^{\bar{s}}(\omega) \subseteq \Omega_{d,s}(\omega)$. Overall, compared to Theorem 4.2.2, variable selection with unknown sparsity is harder than the known case, which is intuitively due to the lack of exact knowledge of s.

Remark 5. The delicate choice of $\tau = \omega\beta_{\min}^2/4$ in Theorem 4.2.5 serves as a balance to correctly identify the alternative T without over-penalizing S_*. We emphasize our analysis is finite-sample as opposed to classic MLE theory relying on asymptotics [e.g. Nishii, 1988], thus provides an alternative understanding of regularization choice. In light of (4.12), we can draw connections to two traditional model selection methods: AIC [Akaike, 1974]

and BIC [Schwarz, 1978], given by

$$\widehat{S}^{\mathrm{AIC}} = \arg\min_{S \in \mathcal{T}_d^{\bar{s}}} \frac{\|\Pi_S^{\perp} Y\|^2}{n} + |S| \frac{2}{n} \quad \text{and} \quad \widehat{S}^{\mathrm{BIC}} = \arg\min_{S \in \mathcal{T}_d^{\bar{s}}} \frac{\|\Pi_S^{\perp} Y\|^2}{n} + |S| \frac{\log n}{n},$$

respectively. With the the sample size being large enough compared to the sparsity level $n \gtrsim \bar{s}$, we have $n - |S| \asymp n$. Therefore, AIC and BIC are special instances of BSSu with different choices of τ: $\tau = 2/n$ for AIC and $\tau = (\log n)/n$ for BIC. As a result, our finite sample result in Theorem 4.2.5 complements the asymptotic analysis of AIC, BIC type of model selection methods.

It is worth highlighting the technicality involved in the proof of Theorem 4.2.5 (in Appendix C.3). When sparsity is known, Lemma 4.2.1 (through Lemma C.1.1) only needs to consider residual variances. By contrast, in the unknown sparsity case, model complexity matters: The size of T may also be different than S_*. Lemma C.3.1 bounds the error probability of distinguishing S_* and alternative support T based on (4.12), and illustrates the crucial role played by the additive penalty $|S|\tau$ in the estimator to differentiate S_* from T such that the error exponent shrinks fast enough.

Since the known sparsity setting is a special case of unknown sparsity, the lower bound constructions apply with minor modifications. We start with showing the equicorrelation Σ_ω with the correlation specified by ω satisfies the requirement for $\Omega_d^{\bar{s}}(\omega)$. An inspection of the proof of Theorem 4.2.3 reveals that, it suffices to have one single ensemble whose instances only differ in one entry of the support. Hence, we consider an ensemble of supports with size $s = 1 \leq \bar{s}$ and derive the bound below to match the first term in (4.13).

Theorem 4.2.6. *Assuming $\omega < 1$, given n i.i.d. samples from $P_{\beta,\Sigma,\sigma^2}$ with $(\beta, \Sigma, \sigma^2) \in \overline{\mathcal{M}}$. If*

the sample size is bounded as

$$n \leq (1-2\delta) \times \frac{\log d}{\beta_{\min}^2 \omega/\sigma^2},$$

then for any estimator $\widehat{S}$ *for* $S_* = \mathrm{supp}(\beta)$,

$$\inf_{\widehat{S}} \sup_{P \in \overline{\mathcal{M}}} \mathbb{P}_{\beta,\Sigma,\sigma^2}(\widehat{S} \neq S_*) \geq \delta - \frac{\log 2}{\log d}$$

The proof of Theorem 4.2.3 is in Appendix C.4. To match the second term in (4.13), since $\Theta_{d,\bar{s}} \subseteq \Theta_d^{\bar{s}}$, the lower bound construction in Theorem 4.2.4 directly applies for unknown sparsity setting by simply changing s to $\bar{s}$. Combined with Theorem 4.2.3, the lower bounds match the upper bound in Theorem 4.2.5. Therefore, we are able to conclude BSSu is indeed optimal and put forth the optimality result of BSS from known sparsity setting to unknown sparsity.

4.2.3 Polynomial-Efficient Sample Complexity Lower Bound

While the general optimality of BSS is appealing, its computational cost prohibits its practical use. In the standard design where $\Sigma = I_d$, apart from the optimality of BSS, there exist other computationally efficient estimators achieving the same sample complexity, e.g. directly using the support of Lasso estimate [Wainwright, 2009b] or simply by marginal screening. The natural question to ask is whether the optimal sample complexity under the general design can be achieved by more efficient estimators. In this section, we give a negative answer by showing that any polynomial-efficient support estimator with known sparsity cannot avoid the restricted eigenvalue condition, establishing a gap in the sample complexity between the optimal estimator (BSS) and any efficient algorithms.

Our result is closely related to the established lower bound for prediction risk [Zhang et al., 2014], thus we borrow the notion of polynomial-efficient estimator therein. We

briefly introduce the most relevant concepts here; interested readers are advised to consult Zhang et al. [2014] and books on complexity theory [Arora and Barak, 2009] for details. We start with quantization for any input value x to the accuracy level given by an integer τ by defining an operator $\lfloor x \rfloor_\tau := 2^{-\tau} \lfloor 2^\tau \rfloor$. Let size$(x; \tau)$ be the length of the binary representation of $\lfloor x \rfloor_\tau$, and size$(X, y; \tau)$ be the total length of the discretized data as matrix (X, y). Then the polynomial efficiency is defined by three quantities: 1) a positive integer b for the number of bits needed to encode an estimator as a program; 2) a polynomial function G for the discretization accuracy of the input data; 3) a polynomial function H for the runtime of the program.

Definition 10 (Polynomial-efficient support estimator)**.** Given polynomial functions $G : (\mathbb{Z}_+)^3 \to \mathbb{R}_+$, $H : \mathbb{Z}_+ \to \mathbb{R}_+$ and an integer $b \in \mathbb{Z}_+$, a support estimator $\widehat{S}(X, y)$ is (b, G, H)-efficient if:

- It can be represented by a computer program encoded in b bits;
- For every triplet (n, d, s), it accepts inputs quantized to accuracy $\lfloor \cdot \rfloor_\tau$ with $\tau \in G(n, d, s)$;
- For every input (X, y), it is guaranteed to terminate in time $H(\text{size}(X, y; \tau))$.

We will also invoke a common conjecture in complexity theory, namely $\mathbf{NP} \not\subset \mathbf{P/poly}$, where $\mathbf{P/poly}$ consists of problems solvable in polynomial time by a Turing machine with side-input of polynomial length.

The proof is based on the idea of reduction from variable selection to prediction risk via sample splitting. Once an efficient and consistent estimator outputs the correct support on one split of the sample, the prediction risk is $\sigma^2 s/n$ for the OLS estimate of β_{S_*} on the other split. Assuming $\mathbf{NP} \not\subset \mathbf{P/poly}$, Theorem 1 in Zhang et al. [2014] implies a lower bound for the prediction risk of a linear model for any polynomial-efficient estimator of

β vector. Based on that, we show by reduction that any polynomial-efficient support estimator will have error probability lower bounded if it fails to satisfy the optimal sample complexity achieved by BSS multiplied by an additional factor of restricted eigenvalue of the design.

Lemma 4.2.7. *If* $\mathbf{NP} \not\subset \mathbf{P/poly}$*, then for any* $\delta \in (0,1)$*, any* $b \in \mathbb{Z}_+$*, any polynomial functions* $G : (\mathbb{Z}_+)^3 \to \mathbb{R}_+$ *and* $F, H : \mathbb{Z}_+ \to \mathbb{R}_+$ *with* $G(n,d,s) > \frac{\delta}{2\log 2}\log s$ *for* $s \geq 1$*, there exists a sparsity level* $s \geq 1$ *such that for any* $d \in [4s, F(s)]$*,* $n \in [C_1 s \log d, F(s)]$*, and* $\gamma \in [2^{-G(n,d,s)}, s^{-\delta/2} \wedge 1/24\sqrt{2})$*, there exists a design matrix* $X \in \mathbb{R}^{n\times d}$ *and* $\beta \in \Theta_{d,s}$ *such that:*

1. *The RE constant* $|\gamma(X) - \gamma| \leq 2^{-G(n,d,s)}$*;*
2. *For any* (b, G, H)*-efficient support estimator* $\widehat{S}$ *with knowledge of* s*, the error probability is lower bounded as*

$$\mathbb{P}(\widehat{S} \neq S_*) \geq 1 \wedge \frac{C_2}{n} \times \frac{s^{1-\delta}\log d}{\max_{T\neq S_*} \frac{\|\Pi_T^{\perp} X_{S_*\setminus T}\beta_{S_*\setminus T}\|^2}{n}/\sigma^2} \times \frac{1}{\gamma^2},$$

where C_1, C_2 *are positive constants.*

The proof is in Appendix C.6. Actually, since $G(n,d,s)$ is a polynomial, the requirement for it to surpass $\delta \log s$ is easy to satisfy, e.g. $G(n,d,s) = s$.

Let's interpret this lower bound result. From Lemma 4.2.7, for any efficient support estimator to be consistent, i.e. $\mathbb{P}(\widehat{S} \neq S_*)$ goes to zero, the sample size is required to be at least lower bounded by (since δ can be arbitrarily small)

$$n \gtrsim \frac{s\log d}{\max_{T\neq S_*} \frac{\|\Pi_T^{\perp} X_{S_*\setminus T}\beta_{S_*\setminus T}\|^2}{n}/\sigma^2} \times \frac{1}{\gamma^2}. \tag{4.14}$$

To compare with BSS, let's further introduce some notations for the signals. The term $\|\Pi_T^{\perp} X_{S_*\setminus T}\beta_{S_*\setminus T}\|^2/n$ is the excess error in the prediction risk of OLS estimate when mis-

specifying the support by T instead of S_*, meanwhile, it also characterizes the pairwise signal to distinguish the true support with alternative T. It can be viewed as a fixed design counterpart of (4.7), based on whose definition, we analogously introduce the scaled version of them by the difference in supports $|S_* \setminus T|$:

$$\Delta_u := \max_{T \in \mathcal{T}_{d,s} \setminus \{S_*\}} \frac{1}{|S_* \setminus T|} \frac{\|\Pi_T^{\perp} X_{S_* \setminus T} \beta_{S_* \setminus T}\|^2}{n\sigma^2}$$
$$\Delta_l := \min_{T \in \mathcal{T}_{d,s} \setminus \{S_*\}} \frac{1}{|S_* \setminus T|} \frac{\|\Pi_T^{\perp} X_{S_* \setminus T} \beta_{S_* \setminus T}\|^2}{n\sigma^2}.$$

Unlike model classes (4.2) for uniform bound, the fixed design signals Δ_u and Δ_l are pointwise quantities and depend on X, β, S_*, σ^2. The only difference between them is whether the maximum or minimum of the pairwise signals is taken over all possible alternatives. These notations are helpful to draw conclusion on the same page. Using Δ_l, we can derive below the fixed design version of the sample complexity upper bound for BSS as (4.8), whose proof is exactly the same as that of Lemma 4.2.1 thus omitted:

$$n \gtrsim \frac{\log d}{\Delta_l} \vee \log \binom{d-s}{s}. \tag{4.15}$$

Using Δ_u, we can simplify (4.14) by upper bounding the denominator by $s\Delta_u$ and compare with BSS on the first term of (4.15) (the second term is required for any estimators by Theorem 4.2.4):

Theorem 4.2.8. *Under the conditions in Lemma 4.2.7, the sample complexity gap between the statistical optimality and any polynomial-efficient support estimator is*

$$\underbrace{\frac{\log d}{\Delta_l}}_{\textit{optimal by BSS}} \quad \textit{vs.} \quad \underbrace{\frac{\log d}{\Delta_u} \times \frac{1}{\gamma^2}}_{\textit{polynomial-efficient lower bound}}. \tag{4.16}$$

Theorem 4.2.8 is implied by Lemma 4.2.7. Recall that $\gamma(X)$ is the RE of the design

matrix (cf. Definition 9). Assuming $\Delta_l \asymp \Delta_u$, we observe the unavoidable gap between optimal sample complexity (achieved by BSS) and any polynomial-efficient estimator by a (squared) factor of restricted eigenvalue. This gap consolidates the optimal position of BSS, especially in the general design setting. In other word, there is no polynomial-efficient substitute for BSS that can attain the same performance.

The lower bound in (4.16) is also suggestive in its connection with existing results for polynomial-efficient estimators based on ℓ_1-regularization. For example, many such methods estimate the support by applying various thresholding techniques to a Lasso-based estimator of β, e.g. Meinshausen and Yu [2009] and Ndaoud and Tsybakov [2020]. These methods rely on the consistency of estimating β, and thus typically have dependence on the restricted eigenvalue [Raskutti et al., 2010]. However, these results also impose additional assumptions such as incoherent designs. Under such stronger assumptions, it is possible that (4.16) reflects the optimal sample complexity for any polynomial-efficient support estimator.

Remark 6. The presumption of $\Delta_l \asymp \Delta_u$ in the comparison above requires each covariate in the true support contributes same order effect to Y, via the interplay of conditional variances and corresponding entries in β_{S_*}. The equi-correlation covariance Σ_ω for the lower bound construction in Theorem 4.2.3 is one example. To exactly verify this is non-trivial since the existence claim for the β vector in Theorem 1 of Zhang et al. [2014] does not give an explicit construction. That being said, it remains reasonable to expect $\Delta_l \asymp \Delta_u$ to hold, given the symmetric nature of the construction of the hard instance for the design matrix X.

Remark 7. One recent development proposes a polynomial-time solution to BSS [Zhu et al., 2020], however, like most efficient support estimators, this approach requires additional conditions on the covariance. Specifically, it imposes Sparse restricted condition (SRC), which is closely related to Restricted isometry property (RIP) [Candes and Tao,

2005] and requires for any $|T| \leq 2s$, and $u \neq 0$, $c_- \leq \|X_T u\|^2/(n\|u\|^2) \leq c_+$ for some constants c_- and c_+. This effectively demands any submatrix of the covariance is close to being orthonormal, and is easily violated under general designs. For instance, consider the equi-correlation covariance Σ_ω used in the lower bound construction of Theorem 4.2.3. For any T with $|T| = 2s$, $u = \mathbf{1}_{2s}$, we have as $s \to \infty$, $u^\top \Sigma_{w,TT} u/\|u\|^2 = (2s-1)(1-\omega)+1 \to \infty$ diverges and thus violates SRC in expectation.

4.3 KL-BSS: A Better Choice for Neighbourhood Selection

The optimality results and insights into the statistical–computational trade-off in Section 4.2 highlight the central role of BSS in the general variable selection problem (cf. (4.2)). However, for the purpose of structure learning, the presence of structure itself may benefit learning the neighborhood. This leads to a natural question: does the optimality of BSS persist in the context of neighborhood selection, or can one achieve improved performance by exploiting the structured model in (4.5)? In this chapter, we provide an answer to this question. We show that it is indeed possible to attain improved sample complexity, thereby surpassing the performance of the BSS golden standard. To this end, we introduce a new estimator, KL-BSS, and demonstrate its advantages both theoretically and empirically.

4.3.1 KL-BSS Estimator

We now introduce our estimator of the support S_*, called KL-BSS. Throughout this section we assume that $\Theta = \Theta_{d,s}(\beta_{\min})$ is given, i.e. the knowledge of sparsity level s and the signal strength $\beta_{\min}$, thus the candidate supports are $\mathcal{T}_{d,s}$, which we recall denotes all possible supports of size s. Extensions to unknown sparsity and beta-min are discussed in Section 4.3.3. Since KL-BSS adopts the tournament interpretation of BSS as searching

over all possible supports and conducting pairwise comparisons, we start by introducing the building block of our estimator: Comparing two candidate supports (Algorithm 3).

Comparing Two Candidates For any two candidate supports $S, T \in \mathcal{T}_{d,s}$, instead of directly comparing residual variances as in BSS, Algorithm 3 chooses the "better" candidate using a newly defined score $\mathcal{L}$ with an additional term that arises from the KL-divergence between each candidate model, explained below. In order to avoid notational clutter, let the shared component be $W := S \cap T$, and the difference between two candidate supports be $S' := S \setminus T$ and $T' := T \setminus S$, so that we can write both S and T as $R \cup W$ with $R \in \{S', T'\}$. Then the score for $R \cup W$ is given by

$$\mathcal{L}(R \cup W; (S, T)) := \underbrace{\frac{\|\Pi^{\perp}_{R \cup W} Y\|^2}{n - s}}_{\text{residual variance from BSS}} + \underbrace{\min_{\gamma \in \Theta_R} (\widehat{\gamma} - \gamma)^{\top} \frac{\widetilde{X}_R^{\top} \widetilde{X}_R}{n - (s - r)} (\widehat{\gamma} - \gamma)}_{\text{violation of constraint } \Theta_R}, \quad (4.17)$$

where $\widetilde{X}_R = \Pi^{\perp}_W X_R$ partials out the effect from the shared component X_W on X_R, and $\widehat{\gamma}$ collects the OLS regression coefficients of $\widetilde{Y} = \Pi^{\perp}_W Y$ on $\widetilde{X}_R$, recall that $\Theta_R = \{\beta_R : \beta \in \Theta\}$ and $\Pi^{\perp}_W = I_n - X_W (X_W^{\top} X_W)^{-1} X_W^{\top}$ is the projection matrix.

The first term in $\mathcal{L}$ is exactly the residual variance used in BSS since $\|\Pi^{\perp}_S Y\|^2 = \|Y - X_S^{\top} \widehat{\beta}(S)\|^2$ in (4.6). The second term quantifies the extent to which $\widehat{\gamma}$ "violates" the constraint Θ_R: That is, the partial regression coefficients $\widehat{\gamma}$ need not be in Θ_R when they are close to zero, and the second term measures how far away $\widehat{\gamma}$ is from Θ_R. This term can be interpreted as the (weighted) L^2-projection of $\widehat{\gamma}$ onto Θ_R. When either S or T is the true support, the second term is zero in expectation, while that of the incorrect support can be positive. So this additional term helps to detect when a candidate set has its partial regression coefficients close to zero. In other words, the second term captures available signal that BSS ignores: The two terms in (4.17) can be interpreted as the minimum KL divergence between any two models specified by the two candidate supports S, T given

Algorithm 3 COMPARE algorithm

Input: Data matrix X; response Y; candidate supports $S, T \in \mathcal{T}_{d,s}$; coefficient space Θ.
Output: Estimated support $\widehat{S}$.

1. Let $S' := S \setminus T, T' := T \setminus S, W := S \cap T, r := |S'| = |T'|$
2. Compute $\widetilde{X}_{S'} = \Pi_W^{\perp} X_{S'}, \widetilde{X}_{T'} = \Pi_W^{\perp} X_{T'}, \widetilde{Y} = \Pi_W^{\perp} Y$
3. For $R \in \{S', T'\}$:
   (a) $\widehat{\gamma} = (\widetilde{X}_R^{\top} \widetilde{X}_R)^{-1} \widetilde{X}_R^{\top} \widetilde{Y}$;
   (b) $\mathcal{L}(R \cup W; (S, T)) = \frac{\|\Pi_{R \cup W}^{\perp} Y\|^2}{n-s} + \min_{\gamma \in \Theta_R} (\widehat{\gamma} - \gamma)^{\top} \frac{\widetilde{X}_R^{\top} \widetilde{X}_R}{n-(s-r)} (\widehat{\gamma} - \gamma)$;
4. Output $\widehat{S} = \arg\min_{D \in \{S,T\}} \mathcal{L}(D; (S, T))$.

Algorithm 4 KL-BSS

Input: Data matrix X; response Y; coefficient space Θ; sparsity s.
Output: Estimated support $\widehat{S}$.

1. Let $M = |\mathcal{T}_{d,s}|$, randomly order the elements in $\mathcal{T}_{d,s}$ to be $S_1, S_2, \ldots, S_M$;
2. Initialize $\widehat{S} = S_1$;
3. For $j = 2, 3, \ldots, M$:
   (a) $\widehat{S} = \text{COMPARE}(X, Y, \widehat{S}, S_j, \Theta)$;
4. Output $\widehat{S}$.

their intersection W, which motivates the design of KL-BSS. We refer to Algorithm 3 as the COMPARE algorithm.

The Proposed Estimator Using COMPARE as our workhorse, we now introduce our proposed estimator, which we call KL-BSS since the estimator can be interpreted via a KL divergence decomposition between the true support and the closest alternative. Conceptually, we can line up all the candidates according to some order, then start with comparing the first two candidates using the prescribed score, and proceed with the winner to compete with the third, etc. After running through each pairwise comparison in this order, a winner is declared. In COMPARE, the pairwise comparison between S and T depends on the shared component W, so the relationship between scores $\mathcal{L}$ defined in (4.17) is not necessarily transitive. Therefore, we adopt this conceptual tournament idea (realized in Algorithm 4) to find the final winner using pairwise comparison along some

order.

It is worth mentioning that the program in the second part of (4.17) is nonconvex when Θ is nonconvex. Of course, since BSS is itself solving a nonconvex combinatorial optimization problem, this is to be expected. Moreover, this is out of necessity: Under standard complexity assumptions, polynomial-time algorithms achieving the optimal rate under general dependence (i.e. our setting) cannot exist in a precise sense (cf. Remark 11). If the space Θ_R is formed by $r = |R|$ many "bounded-away-from-zero" constraints (i.e. $|\beta_j| \geq \beta_{\min}, \forall j \in R, r \in [s]$), then this program can be solved via 2^r quadratic programs with box constraints, of which each one can be solved very fast. Moreover, this program can be cast as a standard mixed integer program (Section 4.3.3). Later in Section 4.3.4, this procedure will be implemented and explored on finite samples.

Remark 8. Neither Algorithm 3 nor Algorithm 4 uses Ω as an input. As a consequence, any structural assumptions on Σ (e.g. SEM assumptions) are not explicitly enforced by the algorithm. In this way, KL-BSS *implicitly* exploits unknown structure without explicitly imposing it. The dependence on Ω only arises in the analysis, where the sample complexity will depend on Σ and/or Ω.

In addition to generalizing BSS, KL-BSS has the important property that on average, it performs at least as well as BSS; this will be a consequence of Theorems 4.3.2-4.3.3 in the next section. Thus, even if the model is not necessarily an SEM, KL-BSS still enjoys all of the storied optimality properties of BSS for general design matrices.

Specifically, the difference between BSS and KL-BSS lies in the second term in (4.17), particularly the way it invokes the constraint Θ in its minimization. In fact, when $\beta_{\min} = 0$—i.e. $\Theta = \mathbb{R}^d$—KL-BSS reduces to BSS. This continues to be true as long as $\beta_{\min} \approx 0$, in which case the beta-min condition will never be violated. As $\beta_{\min}$ increases, the second term measures the extent to which partial regression coefficients in the model violate the constraint Θ, and whenever this term is positive, KL-BSS will improve upon BSS. Thus,

there is a precise sense in which KL-BSS generalizes BSS. This will be made formal in the next sections (Sections 4.3.2-4.3.3) and empirically demonstrated in Section 4.3.4.

4.3.2 Analysis of KL-BSS

In this section, we analyze KL-BSS, with a particular focus on comparing its statistical properties with those of BSS. The key takeaways are: 1) It performs at least as well as BSS on average, and often strictly better when the model is an SEM, and 2) It is optimal over a larger family of design matrices.

Eigenvalue Conditions The variable selection literature expresses the difficulty of recovery in terms of eigenvalue-type conditions that capture the signal for recovering the true support. Following in this tradition, we define a corresponding eigenvalue quantity for KL-BSS and show that it captures the pointwise and minimax sample complexity of KL-BSS. Through this subsection, let Σ be an arbitrary positive-definite matrix.

We first recall the following quantity from Wainwright [2009a], which defines an appropriate eigenvalue for BSS:

$$\lambda_{\mathrm{B}}(\Sigma) := \min_{T \in \mathcal{T}_{d,s} \setminus \{S_*\}} \left[\min_{v \in \mathbb{R}^r} \frac{v^\top \Sigma_{S_* \mid T} v}{\|v\|^2} \right] = \min_{T \in \mathcal{T}_{d,s} \setminus \{S_*\}} \lambda_{\min}(\Sigma_{S_* \mid T}), \tag{4.18}$$

where $r := |S_* \setminus T|$ is the size of $\Sigma_{S_* \mid T}$. This quantity characterizes the information carried by covariates in the true support that is unexplained by alternatives, which is connected to the minimum eigenvalue of Σ and also appears in Shen et al. [2012, 2013]. The idea is that larger $\lambda_{\mathrm{B}}(\Sigma)$ means a stronger signal for variable selection and thus a smaller sample complexity.

Definition 11. Given a design matrix Σ and true support S_*, the KL-BSS eigenvalue is

defined as

$$\lambda_{\mathrm{K}}(\Sigma) := \min_{T \in \mathcal{T}_{d,s} \setminus \{S_*\}} \left[\min_{u \in \Theta_{S_* \triangle T}} \frac{u^\top \Sigma_{S_* \cup T \,|\, S_* \cap T} u}{\min_{|R|=r} \|u_R\|^2} \right] \qquad \text{where } r := |S_* \setminus T| \,. \tag{4.19}$$

Compared to (4.18), the size of $\Sigma_{S_* \cup T \,|\, S_* \cap T}$ is $2r = |S_* \triangle T|$. The difference between (4.18) and (4.19) lies in the conditioning set and size for Σ, the additional restriction in the denominator, and the constraints on u. Roughly speaking, (4.19) is minimizing a larger quantity over a more constrained family, which yields a larger signal, and thus a lower sample complexity.

To make this precise, define a class of design matrices by

$$\Omega_\Delta = \Big\{ \Sigma \in \Omega : \min_{T \in \mathcal{T}_{d,s} \setminus \{S_*\}} \lambda_{\min}(\Sigma_{S_* \,|\, T}) < \min_{T \in \mathcal{T}_{d,s} \setminus \{S_*\}} \lambda_{\min}(\Sigma_{T \,|\, S_*}) \Big\}. \tag{4.20}$$

This class will be used frequently in the sequel: It corresponds to models where KL-BSS strictly outperforms BSS. This is captured by the following lemma, which formalizes this idea that KL-BSS exploits more signal than BSS:

Lemma 4.3.1. *For any $\Sigma \succ 0$, $\lambda_{\mathrm{K}}(\Sigma) \geq \lambda_{\mathrm{B}}(\Sigma)$ and if $\Sigma \in \Omega_\Delta$ then $\lambda_{\mathrm{K}}(\Sigma) > \lambda_{\mathrm{B}}(\Sigma)$.*

See Appendix C.8 for a proof. The construction of Ω_Δ is motivated by the asymmetry in the covariance matrix induced by the existence of structure, and will become more clear through the examples in the next section. For brevity, we simply refer to λ_{K} and λ_{B} as eigenvalues in the sequel, more specifically, the BSS eigenvalue and the KL-BSS eigenvalue.

Strict Improvement Over BSS Roughly speaking, as long as these eigenvalues are nonzero, variable selection is possible, and the larger the eigenvalue, the easier recovery will be. To characterize the performance gap and optimality between KL-BSS and BSS, we consider the SEM class Ω introduced in (4.4-4.5).

We begin with a comparison of the pointwise sample complexity (cf. (1.2)) between KL-BSS and BSS:

Theorem 4.3.2. *Assume $s \le d/2$ with $\beta \in \Theta$. For any $\Sigma \in \Omega$, the (pointwise) sample complexities for variable selection of KL-BSS and BSS are*

$$\underbrace{\frac{\log(d-s)}{\lambda_{\mathrm{K}}(\Sigma)\beta_{\min}^2/\sigma^2} \vee \log\binom{d-s}{s}}_{\text{KL-BSS}} \quad \textit{and} \quad \underbrace{\frac{\log(d-s)}{\lambda_{\mathrm{B}}(\Sigma)\beta_{\min}^2/\sigma^2} \vee \log\binom{d-s}{s}}_{\text{BSS}}.$$

In particular, KL-BSS is more efficient than BSS as long as $\lambda_{\mathrm{K}}(\Sigma) > \lambda_{\mathrm{B}}(\Sigma)$.

By Lemma 4.3.1, we have that *strict* improvement holds for any $\Sigma \in \Omega_\Delta$. The implications of Theorem 4.3.2 are twofold: 1) KL-BSS is always *at least* as sample efficient as BSS, and 2) On Ω_Δ, KL-BSS improves BSS and the improvement is strict. A more technical version of this result also holds for any fixed $(\beta, \Sigma, \sigma^2)$; see Remark 12 in Appendix C.7.

Next, we move on to characterize the minimax optimality of KL-BSS in SEM through the parameter space Ω in (4.4). Define for any constant $c_0 > 0$ the following two classes of SEM design matrices:

$$\Omega_{\mathrm{K}} = \left\{\Sigma \in \Omega : \lambda_{\mathrm{K}}(\Sigma) \ge c_0\sigma_{\min}^2\right\}, \qquad \Omega_{\mathrm{B}} = \left\{\Sigma \in \Omega : \lambda_{\mathrm{B}}(\Sigma) \ge c_0\sigma_{\min}^2\right\}. \tag{4.21}$$

Recall that $\sigma_{\min}^2$ is the minimum noise variance in the SEM (cf. (4.4)). By Lemma 4.3.1, we have $\Omega_{\mathrm{B}} \subseteq \Omega_{\mathrm{K}}$. Then the following theorem gives the desired minimax characterization in SEM.

Theorem 4.3.3. *Assume $s \le d/2$ with $\beta \in \Theta$. Then the minimax optimal sample complexity over both Ω_{B} and Ω_{K} is*

$$\frac{\log(d-s)}{\sigma_{\min}^2\beta_{\min}^2/\sigma^2} \vee \log\binom{d-s}{s}.$$

Moreover, KL-BSS achieves the optimal sample complexity over both Ω_B and Ω_K.

For comparison with BSS, it is known that BSS achieves the optimal sample complexity over Ω_B [Wainwright, 2009a], while KL-BSS extends the optimality to $\Omega_K \supseteq \Omega_B$. These results underscore the critical roles played by the eigenvalues λ_K and λ_B. The proofs are in Appendix C.9-C.10.

Remark 9. The optimality results in Theorem 4.3.3 for Ω_K and Ω_B can be extended beyond SEM, e.g. $\Omega'_K = \{\Sigma \in \mathbb{S}^d_{++} : \lambda_K(\Sigma) \geq \omega\}$ and $\Omega'_B = \{\Sigma \in \mathbb{S}^d_{++} : \lambda_B(\Sigma) \geq \omega\}$ with $\sigma^2_{\min}$ replaced by ω. See the proof of Theorem C.10.1 and Remark 14 in Appendix C.10 for details.

Theorems 4.3.2 and 4.3.3 together imply that not only is KL-BSS minimax optimal over a larger family of designs, but also that there is a class of designs—those in Ω_Δ—over which KL-BSS strictly outperforms BSS. The following example sheds some light on just how large this class is:

Example 1 (Models where KL-BSS outperforms BSS)**.** Consider any SEM as in (4.4) with design matrix Σ. Given the target variable Y with true support S_*, let $C = \Sigma_{S_*}$, $D = \Sigma_{S_*^c \mid S_*}$, so we can write (up to a permutation of the rows and columns)

$$\Sigma = \Sigma(A, C, D) = \begin{pmatrix} C & A^\top \\ A & D + AC^{-1}A^\top \end{pmatrix} \tag{4.22}$$

for some matrix $A \in \mathbb{R}^{(d-s)\times s} \neq 0$. Then we have $\Sigma \in \Omega_\Delta$ as long as

$$\lambda_{\min}(D) \geq \lambda_{\min}(C) \iff \lambda_{\min}(\Sigma_{S_*^c \mid S_*}) \geq \lambda_{\min}(\Sigma_{S_*}). \tag{4.23}$$

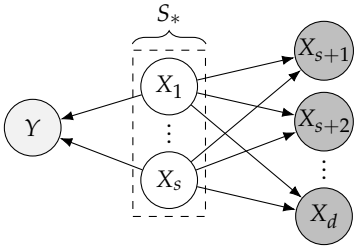
In particular, by Lemma 4.3.1 and Theorem 4.3.2, this gives explicit examples where KL-BSS is strictly more sample efficient than BSS. Since A and C here are essentially arbitrary, the only constraint appears on D through the eigenvalue constraint (4.23). While

this captures a wide range of models, of course this may not always hold. Fortunately, (4.23) is merely a sufficient condition, and the weaker condition in (4.20) substantially relaxes this sufficient condition. See Appendix C.11.

In the next section, we will see how the properties of SEM make these conditions easy to satisfy. This will also help motivate the full relaxation in (4.20), which will prove useful for general SEM.

A Quantitative Example We conclude with a concrete example to demonstrate these ideas and for direct comparison with existing methods.

Example 2 (Comparison with existing methods)**.** Consider the following SEM, with $b > \beta_{\min} > 0$:



$$
\begin{aligned}
X_k &= \begin{cases} \epsilon_k, & k \in [s] \\ \sum_{j \in \mathrm{pa}(k)} b_{jk} X_j + \epsilon_k & k \in \{s+1, \ldots, d\} \end{cases} \\
Y &= \beta_{\min} \times \sum_{k=1}^{s} X_k + \epsilon \\
\epsilon, \epsilon_k &\sim \mathcal{N}(0,1) \qquad b_{jk} = b,\ \forall j,k\,.
\end{aligned}
\tag{4.24}
$$

The uniform choice $b_{jk} \equiv b$ makes the calculation below easier, and helps to expose the dependence on the SEM coefficients b_{jk} more explicitly. The underlying DAG is a bipartite graph with two layers, where the true support is $S_* = [s]$ and the alternative variables $\{s+1, \ldots, d\}$ form the second layer. The covariance and the correlation matrix are

$$
\Sigma = \begin{pmatrix} \Sigma_{S_*S_*} & \Sigma_{S_*S_*^c} \\ \Sigma_{S_*^cS_*} & \Sigma_{S_*^cS_*^c} \end{pmatrix} = \begin{pmatrix} I_s & b\mathbf{1}_s\mathbf{1}_{d-s}^\top \\ b\mathbf{1}_{d-s}\mathbf{1}_s^\top & I_{d-s} + sb^2\mathbf{1}_{d-s}\mathbf{1}_{d-s}^\top \end{pmatrix} \text{ and } \widetilde{\Sigma}_{ij} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}.
$$

Table 4.1: Summary of comparison of KL-BSS with existing methods under model (4.24). The first column is the names of methods under comparison; the second column is the sample complexity upper bound of each method specified under (4.24) or the conclusion about the critical condition of the method; the last column is further explanation about the second column.

	Sample complexity under (4.24)	Comments
KL-BSS	$n \gtrsim \log(d-s)/\beta_{\min}^2 \vee \log\binom{d-s}{s}$	
BSS	$n \gtrsim b^2 s^2 \log(d-s)/\beta_{\min}^2$	
Multi-stage methods	$n \gtrsim b^2 s^2 \log d$	Needed to satisfy RE condition
Lasso	Irrepresentability Fails	$\gamma > \sqrt{s}/2 \to \infty$ violates $\gamma < 1$
OMP	Mutual incoherence Fails	$\mu = \frac{sb^2}{sb^2+1} \to 1$ violates $\mu \leq \frac{1}{2s-1}$

Table 4.1 compares the performance of KL-BSS on this model with existing methods. We include BSS, OMP, Lasso, and Lasso-based multi-stage methods, which depend on additional eigenvalue conditions that BSS and KL-BSS do not require. In this example, these conditions are not met or demand large sample size by the model (4.24). KL-BSS outperforms BSS at least by a factor of s in sample complexity, and other methods either require large sample size (again by a factor of s), or fail to meet the existing conditions for exact recovery. Note also the additional dependence of BSS on the SEM coefficients b^2, which KL-BSS avoids. Recall that we allow s to grow linearly with d, so the gap is indeed severe in high-dimensions. See Table 4.1 for a summary.

4.3.3 Practical Considerations

Here, we discuss several aspects of extensions for usage in practice. Similar to BSS, KL-BSS as described in Algorithm 4 involves an exhaustive search over all candidate supports and assumes various problem parameters such as the sparsity level are known. This is for theoretical convenience and simplicity, and is not necessary in practice. We now discuss how to implement KL-BSS using standard solvers, and extend this implementation to more practical settings where problem parameters are unknown.

Vanilla KL-BSS: A Mixed Integer Reformulation We can reformulate KL-BSS as a standard Mixed Integer Program (MIP), borrowing from Bertsimas et al. [2016]. In this way, we can leverage recent advances in MIP solvers to achieve faster computation as well eliminate the dependence on random ordering. Given $\Theta_{d,s}(\beta_{\min})$, we solve the following program:

$$\min_{\beta,\gamma,z,w} \frac{1}{n-s}\|Y - X\beta\|^2 + (\beta-\gamma)^\top \frac{X^\top X}{n}(\beta-\gamma) \tag{4.25}$$

$$\text{such that} \qquad z \in \{0,1\}^d, \quad w \in \{0,1\}^d, \quad \sum_k z_k = s \tag{4.26}$$

$$\forall k \in [d], \quad (\beta_k, 1-z_k) : \text{SOS-1}\,, \quad (\gamma_k, 1-z_k) : \text{SOS-1}$$

$$\gamma_k + M w_k \geq \beta_{\min} z_k\,, \quad -\gamma_k + M(1-w_k) \geq \beta_{\min} z_k\,,$$

where SOS-1 stands for Ordered Sets of Type 1, which means at most one variable in $(\beta_k, 1-z_k)$ can be nonzero. Each $z_k \in \{0,1\}$ indicates whether the k-th variable is selected, hence the final estimate would be $\text{supp}(z)$. We also introduce integer variables w_k to enforce the nonconvex absolute value constraints $|\gamma_k| \geq \beta_{\min}$, i.e. $\gamma_k \geq \beta_{\min}$ or $\gamma_k \leq -\beta_{\min}$, with a large enough positive constant parameter M, e.g. $M \geq$ upper bound of $|\beta_k| + \beta_{\min}$. In this way, only one of the last two constraints would be effective with M large enough. The equality constraint $\sum_k z_k = s$ can be replaced by $\sum_k z_k \leq \bar{s}$ for the extension in Section 4.3.3.

This formulation is equivalent to Algorithm 4, except that it skips the step of partialling out the support intersection when evaluating scores in Algorithm 3. It follows that the estimator no longer depends on the candidate support ordering, which is convenient in practice for reproducibility. In Section 4.3.4, we evaluate the effect of skipping this step: In practice, this actually slightly *improves* the performance on average. But there is a tradeoff in terms of the worst-case performance; for more details see Appendix C.7.2. We refer to this modified approach that skips the partialing step as Vanilla KL-BSS. Using

standard MIP solvers, Vanilla KL-BSS scales to large problem sizes (up to 1000 variables in our experiments).

Practical Extensions We can extend KL-BSS to the case where both the exact sparsity s and $\beta_{\min}$ are unknown. We start with the unknown sparsity case by assuming an upper bound $\bar{s} \geq s$ is given. In this case, the space of candidate supports expands from $\mathcal{T}_{d,s}$ to $\mathcal{T}_d^{\bar{s}} = \{S \subseteq [d] : |S| \leq \bar{s}\}$. Then we adopt the same strategy of minimizing scores of candidate supports with an additional penalty proportional to their cardinality. Specifically, we modify the MIP objective (4.25) to

$$\min_{\beta,\gamma,z,w} \frac{1}{n-s}\|Y - X\beta\|^2 + (\beta-\gamma)^\top \frac{X^\top X}{n}(\beta-\gamma) + \tau \sum_k z_k$$

and replace the exact sparsity constraint (4.26) by an upper bound $\sum_k z_k \leq \bar{s}$. Here, the parameter τ measures the strength of the penalty. A choice based on $\mathcal{M}$ leads to a sample complexity analogous to that in Theorem 4.3.2, with a modified signal to account for the enlarged support space, see appendix in Gao et al. [2025a] In practice, popular choices are $\tau = \log n$ (BIC) and $\tau = \log d$ (extended BIC). The same extension also applies without the MIP reformulation; e.g. we can modify the output in Algorithm 3 as

$$\widehat{S} := \operatorname*{arg\,min}_{D \in \{S,T\}} \left(\mathcal{L}(D; (S,T)) + \tau |D| \right).$$

Then the algorithm inputs are changed and KL-BSS is called as KL-BSS$(X, Y, \Theta, \bar{s}, \tau)$.

We next consider the case where $\beta_{\min}$ is also unknown. When $\beta_{\min}$ is unknown, we must choose a surrogate value $\widetilde{\beta}_{\min}$ to plug into the input space $\widetilde{\Theta} := \Theta(\widetilde{\beta}_{\min})$. A theoretical choice achieving the sample complexity in Theorem 4.3.3 is provided in Gao et al. [2025a]. In practice, when there is no guidance on the choice of $\widetilde{\beta}_{\min}$, a smaller value of $\widetilde{\beta}_{\min}$ is conservative but safe, because it will not over-penalize the true support. In

Algorithm 5 KL-BSS with unknown sparsity and $\beta_{\min}$

Input: Data matrix X; response Y; candidate choices $\{\beta_{\min}^{\ell}\}_{\ell=1}^{L}$; sparsity upper bound $\bar{s}$; penalty τ

Output: Estimated support $\widehat{S}$.

1. Randomly divide the dataset $\mathcal{D} = (X, Y)$ into K folds: $\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(K)}$, let $\mathcal{D}^{(-k)} = \cup_{j \neq k} \mathcal{D}^{(j)}$;
2. For $\ell = 1, 2, \ldots, L$ and $k = 1, 2, \ldots, K$:
   (a) Let $\widehat{S}_{\ell}^{(k)} = \text{KL-BSS}(\mathcal{D}^{(-k)}, \Theta(\beta_{\min}^{\ell}), \bar{s}, \tau)$;
   (b) Use data $\mathcal{D}^{(-k)}$ to compute $\widehat{\beta}_{\ell}^{(k)}$, the OLS vector of regressing Y onto $\widehat{S}_{\ell}^{(k)}$;
   (c) Let $r_{\ell}^{(k)} := \sum_{i \in \mathcal{D}^{(k)}} (Y_i - X_{i,\widehat{S}^{(k)}}^{\top} \widehat{\beta}_{\ell}^{(k)})^2$;
   (d) Let $r_{\ell} := \sum_{k=1}^{K} r_{\ell}^{(k)}$ be the risk of $\beta_{\min}^{\ell}$;
3. Output $\widehat{S} = \text{KL-BSS}(\mathcal{D}, \Theta(\widehat{\beta_{\min}^{\ell}}), \bar{s}, \tau)$, where $\widehat{\ell} = \arg\min_{\ell \in [L]} r_{\ell}$.

practice, we can also use cross-validation (CV) to choose $\widetilde{\beta}_{\min}$. Given a range of possible choices $\{\beta_{\min}^{\ell}\}_{\ell=1}^{L}$, we consider the standard K-fold CV procedure to estimate regression coefficients on $K-1$ folds of data, then evaluating on the remaining fold using squared residual as criterion.

In Algorithm 5, we provide a consolidated procedure that does not require the knowledge of exact sparsity and $\beta_{\min}$, which outlines the detailed steps of CV. This approach is generic and directly applies to both KL-BSS (Algorithm 4) and Vanilla KL-BSS (MIP formulation). The effectiveness of selecting $\widetilde{\beta}_{\min}$ by CV and the performance of Algorithm 5 are verified by the experiments in Section 4.3.4.

Remark 10 (Standardization). Standardization is a common pre-processing step in regression analyses. A simple approach to handling standardized data is to use a data-driven choice of $\beta_{\min}$, which is also applicable with the CV framework in Algorithm 5: Instead of fixing one $\beta_{\min}$ for all variables, replace $\beta_{\min}$ with $\beta_{\min} \cdot \widehat{sd}_j$ in the definition of Θ, where $\widehat{sd}_j$ is the sample standard deviation of X_j. This adjusts for the effect of standardization, only requires tuning one single parameter $\beta_{\min}$, and yields similar performance to the experiments in Section 4.3.4. Moreover, in the theoretical results, standardization does not

affect λ_{K}, except that the space of minimization $\Theta_{S_* \triangle T}$ in Definition 11 is changed according to the sample standard deviations. Standardization also does not affect the relationship $\lambda_{\mathrm{K}}(\Sigma) \geq \lambda_{\mathrm{B}}(\Sigma)$ in Lemma 4.3.1, and the space of covariance matrices Ω_Δ where strict improvement takes place, i.e. when the requirement of Ω_Δ holds pre-standardization, it still holds post-standardization.

Computational Complexity Since both BSS and KL-BSS can be implemented as an MIP, in practice the required computation for each method is on the same order. In order to provide a more precise worst-case comparison, however, we can use the naïve tournament-style interpretation of each method to compare their respective computational complexity as follows. As such, this is not intended to be a rigourous complexity analysis, but rather a worst-case comparison to highlight the small computational cost of KL-BSS relative to BSS.

Given the sparsity level s, BSS searches over all possible candidates, which has size $\binom{d}{s} \asymp d^s$. For each candidate, BSS needs to evaluate the residual variance. By contrast, KL-BSS conducts $\binom{d}{s} - 1$ pairwise comparisons along the given order, and each comparison requires computations of scores (cf. (4.17)) for both candidates. These can be obtained naïvely by solving 2^s quadratic programs with box constraints, each of which (as well as the residual variance computation for BSS) are standard convex programs. Thus, the computational complexity of KL-BSS is, in the worst-case,

$$\left[\binom{d}{s} - 1\right] \times 2 \times (2^s + 1) \asymp (2d)^s \,.$$

Compared with BSS with complexity d^s, the cost that KL-BSS pays is of smaller order.

Remark 11. A natural question is whether or not neighbourhood selection in SEM can be achieved with efficient (i.e. polynomial-time) estimators, such as a Lasso-based method. It is known that the Lasso needs strong conditions on the design matrix, which we have

already shown are easily violated in SEM (Example 2). In fact, Theorem 4.2.8 makes this strengthened: Under standard conjectures in complexity theory, *any* polynomial-time estimator for variable selection under general design cannot avoid the restricted eigenvalue condition, even if the sparsity is known.

4.3.4 Experiments

In this section, we conduct experiments to empirically evaluate the performance of KL-BSS, in particular compared to BSS and the Lasso. We start with a comprehensive simulation study to show a significant sample complexity gap in randomly generated SEMs. Next, we compare KL-BSS and BSS in the context of structure learning. Finally, we provide a real-data application using gene expression data, assessing both recovery and prediction performance.

Simulation Study: Empirical Sample Complexity Gap We begin with a simulation study where the ground truth is known and we can simulate from different types of SEM. Full experiment results and all implementation details can be found in Appendix E; here we briefly illustrate a representative slice of the results in Figure 4.2 to show the *empirical improvements* in the sample complexity. Results for other metrics, e.g. Hamming distance, TPR, FDR are available in Appendix E.4, particularly a summary across all simulation setups is given in Figure E.1 in Appendix E.5, showing a significant and overall improvement over BSS.

The results cover various combinations of $(d, s, \bar{s})$ and three types of underlying DAGs: Erdös-Rényi (ER) and Scale-Free (SF) graphs with expected number of edges to be $4d$, and Complete graphs where all possible edges are present, and the data is simulated according to (1.1) (specifically (4.1) and (4.4)) where the noise $\{\epsilon_k\}_{k=1}^d$ have mixed, possibly non-Gaussian, distributions, the true supports and SEM coefficients are randomly sam-

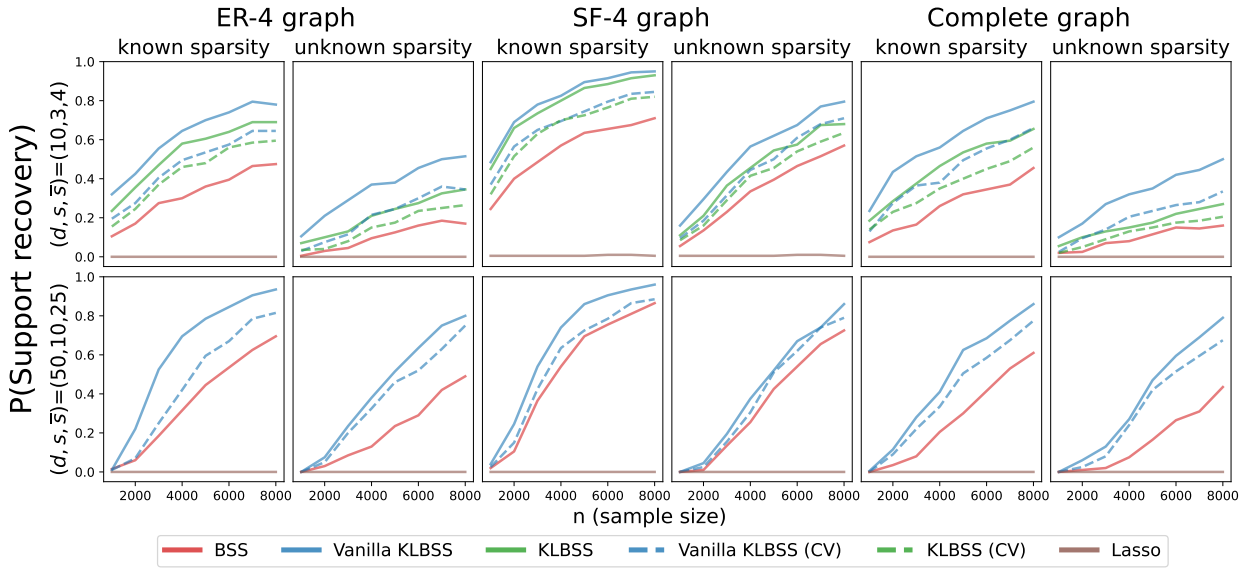


Figure 4.2: Comparison on variable selection performance of BSS, KL-BSS and Lasso under different types of graphs and dimensions $(d, s, \bar{s})$ averaged over 200 replications. The horizontal axis is sample size, the vertical axis is probability of exact recovery. The first/middle/last two columns are for ER graph, SF graph, and complete graph. The left/right columns of each graph type indicate known and unknown sparsity cases. The solid/dashed lines indicate known and unknown $\beta_{\min}$ (by CV). There is a notable performance gap between KL-BSS and BSS. Lasso is never consistent.

pled. For $d = 50$, we implement both KL-BSS and BSS with the MIP reformulation in Section 4.3.3; for unknown sparsity, we input with $\bar{s}$ and apply the BIC penalty discussed in Section 4.3.3; for unknown $\beta_{\min}$, we apply the CV procedure introduced in Algorithm 5. Both ER-4 and SF-4 graphs are sparse random graph ensembles whereas Complete graphs represent a dense graph setting; the latter is particularly interesting since it represents a setting that is closer to the general design setting where BSS is commonly perceived to be optimal. In each instance, the optimality conditions imposed in (4.21) are not always guaranteed to be satisfied, and thus our simulations also represent a more realistic evaluation where our theoretical assumptions are likely to be violated. In particular, path cancellation is ubiquitous and we make no efforts to eliminate path cancellation.

The results in Figure 4.2 confirm that KL-BSS is significantly more sample-efficient compared to BSS and the Lasso and show KL-BSS is robust to deviations from our the-

oretical assumptions. This helps illustrate the benefits and improvement KL-BSS brings in a more *general and practical* class of SEM. Moreover, Vanilla KL-BSS performs slightly better than KL-BSS in the average sense. This does not contradict the minimax optimality of KL-BSS: In Appendix E.6, we demonstrate that there exist hard cases where Vanilla KL-BSS performs worse than KL-BSS. The Lasso fails in all graphs even with large sample size due to the strong dependence structure in Σ.

Structure Learning Since a primary motivation for this work is neighbourhood selection in SEM, we also apply KL-BSS for structure learning, i.e. to recover the DAG G that generates the data X. When the topological ordering of G is known, the problem reduces to variable selection for each variable from the preceding nodes on the ordering with unknown sparsity. We follow the experiment setup as in the previous section with ER/SF graphs and $d = 10$. By setting $\sigma_k \equiv \sigma = 2$ for each k, the DAG G is identifiable [Peters and Bühlmann, 2014] and can be estimated via the EQVAR algorithm [Chen et al., 2019]. We consider two cases: 1) A valid oracle topological ordering is given, and 2) The topological ordering is estimated with the EQVAR algorithm from scratch. The latter probes the robustness of KL-BSS to misspecification of the ordering, and illustrates its applicability for structure learning in practice. A comparison using Structural Hamming Distance (SHD) between the estimated and true graphs is shown in Figure 4.3, where significant improvement can be observed for KL-BSS over BSS, while the difference between KL-BSS and Vanilla KL-BSS is negligible. In the second case where the ordering is estimated from data, overall recovery performance is affected for all the methods, but not by much. Results on Lasso are not shown since it is never consistent and produces SHD around 10-20.

Verification of Ω_Δ In order to explore the prevalence of design matrices where KL-BSS outperforms BSS, in this subsection we explore how often the constraint in Ω_Δ is satisfied in randomly generated SEM. Recall that Ω_Δ defined in (4.20) represents design matrices

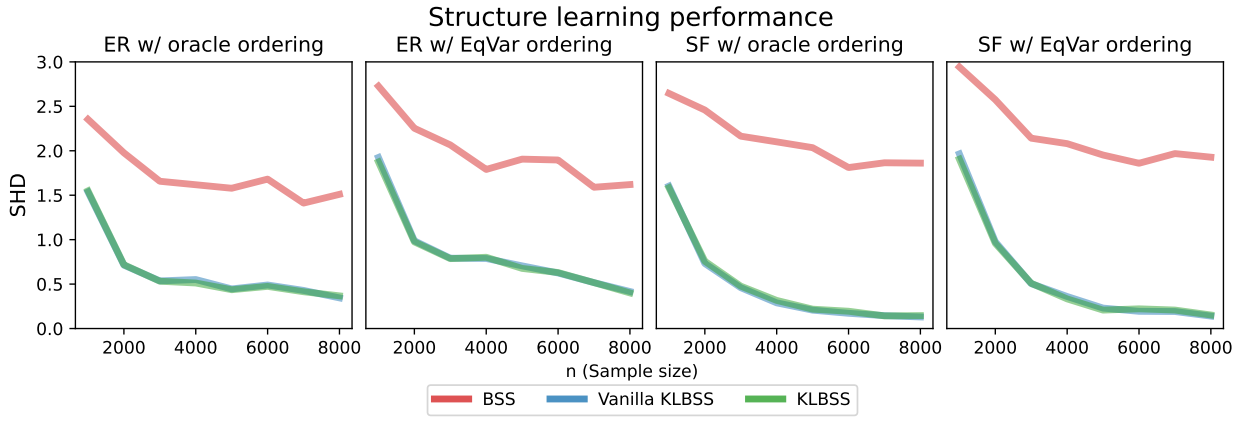


Figure 4.3: Comparison on structure learning performance of BSS and KL-BSS on ER and SF graphs: Structural Hamming Distance (SHD) vs. sample size. The first and third panels are input with oracle valid topological ordering of the graph, the second and forth apply EqVar algorithm for ordering estimation, then conduct neighbourhood selection via KL-BSS or BSS. KL-BSS gives better structure learning performance than BSS. KL-BSS and Vanilla KL-BSS perform similarly thus lines are overlapped.

where KL-BSS has a *provably* better sample complexity according to Theorem 4.3.2. For the random SEMs sampled in previous section, we check if the inequality in (4.20) holds. In this way, we showcase the ubiquitous improvement of KL-BSS in generic SEM class. We vary the sparsity of the random graphs by specifying the expected number of edges in G using both ER/SF graphs ($s_0 \times d$). We consider various sparsity levels s for β as well. We randomly sample 5,000 SEM covariances (and supports S_*) and record the proportion that satisfy Ω_Δ. The result summarized in Figure 4.4 indicates a significant proportion of random SEMs have covariance matrices in Ω_Δ, i.e. showing improvement of KL-BSS. Especially, the proportion grows as the graph becomes denser and the sparsity level of β increases. The effect of the edge density s_0 stands out in ER graphs while it is less significant in SF graphs. Nonetheless, SF graphs still exhibit an average 30% proportion, and ER graphs can reach as high as 60%. This confirms that SEM are indeed likely to fall into the class of design matrices where KL-BSS *strictly* improves the sample efficiency of neighbourhood selection.

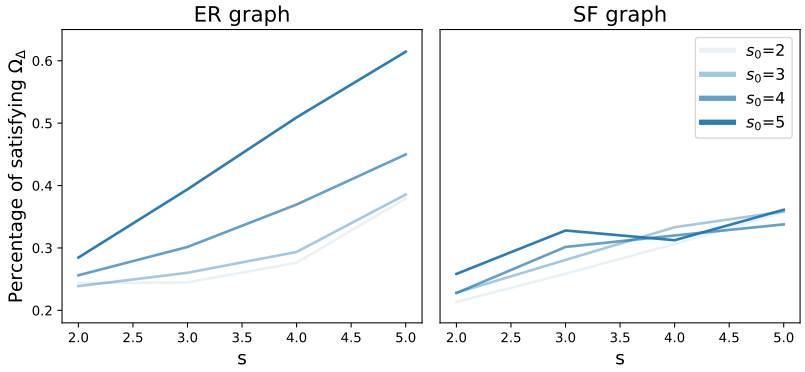


Figure 4.4: The percentage of randomly sampled SEM covariances that satisfy the constraint in Ω_Δ (4.20) for ER/SF graphs with various expected number of edges ($s_0 \times d$) and sparsity s and fixed $d = 12$. Significant proportion of random SEM covariances meet Ω_Δ.

Application to Pan-Cancer Data Finally, to evaluate the performance of KL-BSS on real data, we apply it to a pan-cancer dataset consisting of RNA-Seq gene expression measurements from $n = 801$ patients with 5 different types of cancers [Fiorini, 2016]. Since the linear model is certainly misspecified on such data, and there is no known "ground truth", this dataset allows us to evaluate 1) robustness to misspecification of linear SEM and 2) performance on downstream prediction tasks.

In the first experiment, we use the pan-cancer data for the covariates X and construct the response Y from X by simulation. In this way, we can deal with covariates related by real genetic processes, meanwhile, we also know S_* and are able to evaluate the estimate $\widehat{S}$. Specifically, we group the genes according to their variances into d bins. Then for each replication, we randomly sample one gene from each bin to form the X matrix (of dimension d), with Y simulated as in Figure 4.2. We fix $s = 10$ and show results for increasing d from 50 to 90 in the left panel of Figure 4.5 indicated by solid and dashed lines for KL-BSS and BSS. The barplots depict the gap in the performances of two methods (difference between solid and dashed lines). We can still observe the improved performance against BSS, especially, the gap becomes more pronounced as the dimension gets larger and for smaller sample size.

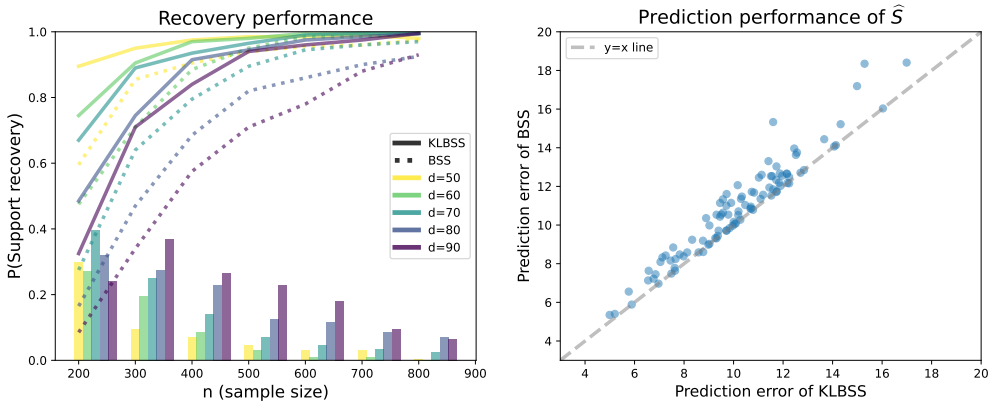


Figure 4.5: (Left) Recovery performance comparison on RNA-Seq gene expression data with $s = 10$. The solid lines are for KL-BSS and dashed lines are for BSS. The barplots indicate the gap between the performances of KL-BSS and BSS, which is more significant for larger d and smaller sample size. (Right) Prediction performance of $\widehat{S}$ given by KL-BSS with CV choice of $\beta_{\min}$ and BSS on gene expression data. Each point is one random sampling of $d = 50$ variables from $20,531$ genes to choose $s = 10$ in training set for prediction in test set. The grey dashed line is the $y = x$ line. The genes selected by KL-BSS result in better prediction error than BSS (above the $y = x$ line).

In the second experiment, we avoid simulations altogether. Since there is no "true" support, we instead evaluate the selected models by the prediction performance on the gene with the largest variance (Y) using the support estimated from the remaining genes as candidate predictors (X). For each replication, we randomly choose $d = 50$ genes as X, and randomly split the dataset in training / test sets. We apply KL-BSS and BSS with $s = 10$ on the training set and compute prediction error on the test set. We use CV for the choice of $\beta_{\min}$ for KL-BSS, perform $N = 100$ replications, and display the result by scatterplot of prediction errors of KL-BSS vs. BSS in Figure 4.5. In a majority (73%) of the evaluations, KL-BSS selected genes with a lower out-of-sample prediction error vs. BSS (indicated by the points above the $y = x$ line). Some (24%) points lie exactly on the $y = x$ line because KL-BSS and BSS both estimate the same support $\widehat{S}$. This demonstrates that KL-BSS selects models that yield better out-of-sample predictions compared to BSS, on realistic data where the underlying model may not be an exact SEM.

CHAPTER 5

CONCLUSION

In this work, we developed a comprehensive understanding of statistical optimality in causal structure learning and its fundamental subroutines. For parametric Gaussian models, we in Chapter 2 establish minimax-optimal sample complexity results for learning linear DAGs under the equal-variance condition, and poly-trees under faithfulness condition. The newly proposed/modified algorithms achieve optimal sample complexity.

In Chapter 3, we then establish a general and quantitative connection between structure learning and CI testing, showing that the optimal sample complexity of structure learning for poly-forest is governed by that of CI testing, which is achieved by an efficient algorithm with the optimal test plugged in. This reduction provides a unifying perspective on statistical optimality and demonstrates the central role of CI testing in constraint-based methods. Extending this framework beyond poly-forests to general DAGs remains an important direction for future work.

In Chapter 4, to further understand optimality in general DAGs, we studied variable selection and neighborhood selection as key subroutines. We show that best subset selection (BSS) achieves minimax-optimal sample complexity for variable selection under general design, and established a lower bound for polynomial-time estimators, revealing a fundamental statistical-computational gap depending on the restricted eigenvalue. At the same time, we demonstrate that this folklore optimality of BSS can be improved in structured settings arising from linear SEM. In particular, we proposed a new estimator, KL-BSS, which leverages latent structures without requiring explicit knowledge of the underlying graph to achieve improved sample complexity under weaker conditions, both theoretically and empirically.

Overall, these results highlight the interplay between model structure, statistical efficiency, and computational constraints in causal structure learning. They show that while

optimality can be characterized in several important settings, exploiting additional structure can lead to strictly improved performance, even when such structure is unknown. Important directions for future work include deriving analysis for score-based methods, extending optimality results to general DAGs, designing computationally efficient algorithms that approach statistical limits, and further understanding how latent structure can be leveraged in high-dimensional statistical problems or how powerful modern machine learning techniques can benefit structure learning.

APPENDIX A

PROOFS IN CHAPTER 2

The ORIENT algorithm below is used in PC type algorithms to infer edge directions after skeleton learning. ORIENT requires input with separation sets obtained from CI testing.

Algorithm 6 ORIENT algorithm

Input: Skeleton $\widehat{G}$, separation sets S

Output: CPDAG $\widehat{\widehat{G}}$.

1. For all pairs of nonadjacent nodes j, k with common neighbour ℓ:
   (a) If $\ell \notin S(j,k)$, then directize $j - \ell - k$ in $\widehat{T}$ by $j \to \ell \leftarrow k$
2. In the resulting PDAG $\widehat{G}$, orient as many as possible undirected edges by applying following rules:
   - **R1** Orient $k - \ell$ into $k \to \ell$ whenever there is an arrow $j \to k$ such that j and ℓ are not adjacent
   - **R2** Orient $j - k$ into $j \to k$ whenever there is a chain $j \to \ell \to k$
   - **R3** Orient $j - k$ into $j \to k$ whenever there are two chains $j - \ell \to k$ and $j - i \to k$ such that ℓ and i are not adjacent
   - **R4** Orient $j - k$ into $j \to k$ whenever there are two chains $j - \ell \to i$ and $\ell - i \to k$ such that ℓ and i are not adjacent
3. Return $\widehat{G}$ as $\widehat{\widehat{G}}$.

A.1 Proof of Theorem 2.2.1 (Upper Bound for EqVar DAG)

Proof. We first show that if all conditional variances are estimated sufficiently well, then Algorithm 1 is able to identify the true DAG.

Lemma A.1.1. *If for all $k \in V$ and $C \subset V \setminus \{k\}$, $|C| \le q$,*

$$|\operatorname{var}(X_k \mid X_C) - v_{kC}| \le \Delta/4$$

then $\widehat{G} = G$.

Next we bound the estimation error tail probability:

Lemma A.1.2. *For all $k \in V$ and $C \subset V \setminus \{k\}$, $|C| \le q$,*

$$\mathbb{P}(|\operatorname{var}(X_k \mid X_C) - v_{kC}| \ge \epsilon) \le A_1 \exp(-A_2 n\epsilon/M^5 + q)$$

for some constants A_1, A_2.

Combine Lemma A.1.1 and A.1.2, we can have success probability:

$$\begin{aligned}
\mathbb{P}(\widehat{G} \ne G) &\le \mathbb{P}\Bigg(\bigcup_{\substack{k\in V \\ C\subset V\setminus\{k\} \\ |C|\le q}} \{|\operatorname{var}(X_k \mid X_C) - v_{kC}| \ge \epsilon\} \Bigg) \\
&\le \sum_{\substack{k\in V \\ C\subset V\setminus\{k\} \\ |C|\le q}} \mathbb{P}\Big(|\operatorname{var}(X_k \mid X_C) - v_{kC}| \ge \epsilon \Big) \\
&\lesssim d \times \left(\binom{d-1}{1} + \ldots + \binom{d-1}{q} \right) \times \exp(q - n\epsilon/M^5) \\
&\lesssim d \times q \times e^{q\log(d/q)} \times \exp(q - n\epsilon/M^5) \\
&= \exp(q\log(d/q) + \log d + \log q + q - n\epsilon/M^5) \\
&\asymp \exp(q\log(d/q) - n\epsilon/M^5)\,.
\end{aligned}$$

The last equality is by $(\frac{d}{q})^q = (\frac{d}{q})^{q-1} \times (\frac{d}{q}) \geq 2^{q-1} \times (\frac{d}{q}) \geq q \times \frac{d}{q} = d$, thus $q \log(d/q) \gtrsim \log d$. In the end, replace ϵ by $\Delta/4$, solve for the sample size n such that

$$\exp(q \log(d/q) - n\Delta/M^5) \lesssim \delta\,,$$

we can have the desired sample complexity. □

We proceed to prove Lemmas A.1.1 and A.1.2.

Proof of Lemma A.1.1. We start by showing that $\widehat{\tau}$ is a valid ordering for G, which is equivalent to saying $\widehat{\tau}_j$ is a source node of the subgraph $G[V \setminus \widehat{\tau}_{[1:j-1]}]$ for all j. We proceed by induction. For $j = 1$, it reduces to compare marginal variances.

$$\begin{cases} \mathrm{var}(X_k) = \sigma^2 & k \text{ is a source node} \\ \mathrm{var}(X_\ell) > \sigma^2 + \Delta & \ell \text{ is not a source node} \end{cases}$$

For any non-source node ℓ and any source node k,

$$\begin{aligned} \sigma_\ell &= v_{\ell\emptyset} \\ &\geq \mathrm{var}(X_\ell) - \Delta/4 \\ &> \Delta \times \frac{3}{4} + \sigma^2 \\ &> \Delta/4 + \sigma^2 \\ &= \mathrm{var}(X_k) + \Delta/4 \\ &\geq v_{k\emptyset} = \sigma_k \end{aligned}$$

Thus k is preferred over ℓ. Given that $\widehat{\tau}_{[1:j-1]}$ correctly identified, by the equal variance

assumption,

$$\min_{C \subseteq \widehat{\tau}_{[1:j-1]}, |C| \leq q} \operatorname{var}(X_k \mid C) = \operatorname{var}(X_k \mid \operatorname{pa}(k)) = \sigma^2$$

if k is a source node of $G[V \setminus \widehat{\tau}_{[1:j-1]}]$

$$\min_{C \subseteq \widehat{\tau}_{[1:j-1]}, |C| \leq q} \operatorname{var}(X_\ell \mid C) > \Delta + \sigma^2$$

if ℓ is not a source node of $G[V \setminus \widehat{\tau}_{[1:j-1]}]$

Therefore, for any ℓ that is not a source node and for any k that is a source node,

$$\begin{aligned}
\sigma_\ell &= \min_{C \subseteq \widehat{\tau}_{[1:j-1]}, |C| \leq q} v_{\ell C} \\
&\geq \min_{C \subseteq \widehat{\tau}_{[1:j-1]}, |C| \leq q} \operatorname{var}(X_\ell \mid C) - \Delta/4 \\
&> \frac{3}{4}\Delta + \sigma^2 \\
&> \Delta/4 + \sigma^2 \\
&= \min_{C \subseteq \widehat{\tau}_{[1:j-1]}, |C| \leq q} \operatorname{var}(X_k \mid X_C) + \Delta/4 \\
&\geq \min_{C \subseteq \widehat{\tau}_{[1:j-1]}, |C| \leq q} v_{kC} = \sigma_k
\end{aligned}$$

Thus the first step of Algorithm 1 will always include k instead of ℓ into $\widehat{\tau}$. This implies that $\widehat{\tau}$ is a valid topological ordering.

Now we look at the second step of Algorithm 1, this step is to remove false parents from candidate set returned by Best Subset Selection. For any j, let $\widehat{\tau}_j = j$ for ease of notation. Given that $\widehat{\tau}$ is a valid ordering, $\operatorname{pa}(j) \subseteq \widehat{\tau}_{[1:j-1]}$. We first conclude $\operatorname{pa}(j) \subseteq C_j$,

otherwise there exists $C'_j \subseteq \widehat{\tau}_{[1:j-1]}$ with $\mathrm{pa}(j) \subseteq C'_j$ such that

$$\begin{aligned} v_{jC_j} &\geq \mathrm{var}(X_j \mid C_j) - \Delta/4 \\ &> \sigma^2 + \Delta - \Delta/4 \\ &> \sigma^2 + \Delta/4 \\ &= \mathrm{var}(X_j \mid C'_j) + \Delta/4 \\ &\geq v_{jC'_j} \end{aligned}$$

Then C_j should not lead to minimum. Then for any $k \in \mathrm{pa}(j)$ and any $\ell \in C_j \setminus \mathrm{pa}(j)$,

$$\begin{aligned} v_{jC_j} - v_{jC_j \setminus k} &\geq \mathrm{var}(X_j \mid C_j) - \mathrm{var}(X_j \mid C_j \setminus k) - \Delta/2 \\ &> \Delta/2 = \gamma \\ v_{jC_j} - v_{jC_j \setminus \ell} &\leq \mathrm{var}(X_j \mid C_j) - \mathrm{var}(X_j \mid C_j \setminus \ell) + \Delta/2 \\ &= \Delta/2 = \gamma \end{aligned}$$

Thus ℓ will be removed while k will stay. Then $\mathrm{pa}_{\widehat{G}}(j) = \mathrm{pa}_G(j)$ for all j. □

Proof of Lemma A.1.2. Denote the covariance between k and set of nodes C at

$$\Sigma_t = \begin{pmatrix} \Sigma_{kk} & \Sigma_{kC} \\ \Sigma_{Ck} & \Sigma_{CC} \end{pmatrix}$$

Note that $\|\Sigma_{kC}\|, \|\Sigma_{CC}\|, \|\Sigma_{CC}^{-1}\| \leq \|\Sigma_t\| \leq \|\Sigma\| \leq M$.

Then the estimation error for conditional variance

$$
\begin{aligned}
&\left|\operatorname{var}(X_k \mid X_C) - \widehat{\operatorname{var}}(X_k \mid X_C)\right| \\
&= \left|(\Sigma_{kk} - \Sigma_{kC}\Sigma_{CC}^{-1}\Sigma_{Ck}) - (\widehat{\Sigma}_{kk} - \widehat{\Sigma}_{kC}\widehat{\Sigma}_{CC}^{-1}\widehat{\Sigma}_{Ck})\right| \\
&\leq \left|\widehat{\Sigma}_{kk} - \Sigma_{kk}\right| + \left|\widehat{\Sigma}_{kC}(\widehat{\Sigma}_{CC}^{-1} - \Sigma_{CC}^{-1})\widehat{\Sigma}_{Ck}\right| \\
&\qquad + \left|(\widehat{\Sigma}_{kC} - \Sigma_{kC})\Sigma_{CC}^{-1}\widehat{\Sigma}_{Ck}\right| + \left|\Sigma_{kC}\Sigma_{CC}^{-1}(\widehat{\Sigma}_{Ck} - \Sigma_{Ck})\right| \\
&\leq \left|\widehat{\Sigma}_{kk} - \Sigma_{kk}\right| + \|\widehat{\Sigma}_{CC}^{-1} - \Sigma_{CC}^{-1}\|\|\widehat{\Sigma}_{kC}\|^2 \\
&\qquad + M\|\widehat{\Sigma}_{kC} - \Sigma_{kC}\|\|\widehat{\Sigma}_{Ck}\| + M^2\|\widehat{\Sigma}_{Ck} - \Sigma_{Ck}\| \\
&\leq \left|\widehat{\Sigma}_{kk} - \Sigma_{kk}\right| \\
&\qquad + 2\|\widehat{\Sigma}_{CC}^{-1} - \Sigma_{CC}^{-1}\|\|\widehat{\Sigma}_{kC} - \Sigma_{kC}\|^2 + 2M^2\|\widehat{\Sigma}_{CC}^{-1} - \Sigma_{CC}^{-1}\| \\
&\qquad + M\|\widehat{\Sigma}_{kC} - \Sigma_{kC}\|^2 + M\|\widehat{\Sigma}_{kC} - \Sigma_{kC}\| \\
&+ M^2\|\widehat{\Sigma}_{Ck} - \Sigma_{Ck}\| \\
&\leq \left|\widehat{\Sigma}_{kk} - \Sigma_{kk}\right| \\
&\qquad + 2\|\widehat{\Sigma}_{CC}^{-1} - \Sigma_{CC}^{-1}\|\|\widehat{\Sigma}_{t} - \Sigma_{t}\|^2 + 2M^2\|\widehat{\Sigma}_{CC}^{-1} - \Sigma_{CC}^{-1}\| \\
&\qquad + M\|\widehat{\Sigma}_{t} - \Sigma_{t}\|^2 + M\|\widehat{\Sigma}_{t} - \Sigma_{t}\| + M^2\|\widehat{\Sigma}_{t} - \Sigma_{t}\|
\end{aligned}
$$

The first inequality is by the triangular inequality, and the second simply bounds $\|\Sigma_{kC}\|$ by M. The third inequality introduces the estimation error of $\widehat{\Sigma}_{kC}$ and the final inequality replaces this with the estimation error of the full covariance matrix $\widehat{\Sigma}_t$. To set the RHS to be smaller than $\epsilon > 0$, we consider three estimation errors. The first two can be controlled via standard sub-exponential concentration, whereas the third can be controlled via Theorem 6.5 from Wainwright [2019]:

$$
\begin{aligned}
\mathbb{P}(|\widehat{\Sigma}_{kk} - \Sigma_{kk}| \geq \zeta) &\leq \exp(-A_3 n\zeta/M) \\
\mathbb{P}(\|\widehat{\Sigma}_{t} - \Sigma_{t}\| \geq \zeta) &\leq \exp(-A_4 n\zeta/M + q)
\end{aligned}
$$

for some constants A_3, A_4. The largest error is from

$$\begin{aligned}
\|\widehat{\Sigma}_{CC}^{-1} - \Sigma_{CC}^{-1}\| &\leq \|\Sigma_{CC}^{-1}\|\|\widehat{\Sigma}_{CC}^{-1}\|\|\Sigma_{CC} - \widehat{\Sigma}_{CC}\| \\
&\leq M\|\widehat{\Sigma}_{CC}^{-1}\|\|\Sigma_{CC} - \widehat{\Sigma}_{CC}\| \\
&= M(\|\widehat{\Sigma}_{CC}^{-1}\| - \|\Sigma_{CC}^{-1}\|)\|\Sigma_{CC} - \widehat{\Sigma}_{CC}\| \\
&\quad + M\|\Sigma_{CC}^{-1}\|\|\Sigma_{CC} - \widehat{\Sigma}_{CC}\| \\
&\leq M\|\widehat{\Sigma}_{CC}^{-1} - \Sigma_{CC}^{-1}\|\|\Sigma_{CC} - \widehat{\Sigma}_{CC}\| \\
&+ M^2\|\Sigma_{CC} - \widehat{\Sigma}_{CC}\|
\end{aligned}$$

After some arrangement, we have

$$\|\widehat{\Sigma}_{CC}^{-1} - \Sigma_{CC}^{-1}\| \leq \frac{M^2\|\Sigma_{CC} - \widehat{\Sigma}_{CC}\|}{1 - M\|\Sigma_{CC} - \widehat{\Sigma}_{CC}\|} \leq \zeta$$

as long as

$$\|\Sigma_{CC} - \widehat{\Sigma}_{CC}\| \leq \frac{\zeta}{M^2 + M\zeta} \leq \frac{\zeta}{M^2}.$$

This is just another Gaussian covariance matrix estimation error, i.e.

$$\mathbb{P}(\|\widehat{\Sigma}_{CC} - \Sigma_{CC}\| > \zeta/M^2) \leq \exp(-A_5 n\zeta/M^3 + q),$$

for some constant A_5. Now we require all the errors to be bounded by $\zeta = \epsilon/M^2$ such that the conditional variance estimation error is within ϵ. Thus

$$\begin{aligned}
&\mathbb{P}(|\operatorname{var}(X_k|X_C) - \widehat{\operatorname{var}}(X_k|X_C)| > \epsilon) \\
&\leq \exp(-A_3 n\epsilon/M^3) + \exp(-A_4 n\epsilon/M^3 + q) \\
&\quad + \exp(-A_5 n\epsilon/M^5 + q) \\
&\leq A_1 \exp(-A_2 n\epsilon/M^5 + q).
\end{aligned}$$

□

A.2 Proof of Theorem 2.2.2 (Upper Bound for EqVar DAG without tuning parameter)

Proof. In the proof of Lemma A.1.1, denote the estimation error to be upper bounded by ϵ, i.e. for all $k \in V$ and $C \subset V \setminus \{k\}$, $|C| \leq q$,

$$|\operatorname{var}(X_k \mid X_C) - v_{kC}| \leq \epsilon\,,$$

then it suffices to have

$$2\epsilon < \gamma < \Delta/2 - 2\epsilon$$

for the correctness of second phase to proceed. Therefore, let $\gamma = 3\epsilon$ and require $\epsilon < \Delta/10$. Finally, set

$$\epsilon \asymp \frac{2M^5 \times q \log(d/q)}{n}$$

Then we have failure probability bounded:

$$\begin{aligned}\mathbb{P}(\widehat{G} \neq G) &\lesssim \exp(q\log(d/q) - n\epsilon/M^5) \\ &= \exp(-q\log(d/q))\,.\end{aligned}$$

And to satisfy the requirement $\epsilon < \Delta/10$, we need sample size

$$n \gtrsim \frac{M^5 q \log(d/q)}{\Delta}\,.$$ □

A.3 Proof of Theorem 2.2.3 (Lower Bound for EqVar DAG)

Proof. We are going to apply Corollary D.2.3 to derive lower bounds. It suffices to find a subclass of original problem which (i) has large cardinality N; (ii) pairwise KL divergence between any two distributions is small.

We start by counting for the number of DAGs with d nodes and in-degree bounded by q.

Lemma A.3.1. *For $q \leq d/2$, the number of DAGs with d nodes and in-degree bounded by q scales as $\Theta(dq \log d/q)$.*

Formally to show the lower bounds, we consider two ensembles:

Ensemble A In this ensemble, we consider all possible DAGs with bounded in-degree. Note that by Lemma A.3.1, we know $N \asymp dq \log d/q$, it remains to provide an upper bound for the KL divergence between any two distributions within the class. For any two $F_j, F_k \in \mathcal{F}_{d,q}$, denote their covariance matrices to be Σ_j, Σ_k. Due to Gaussianity, It is easy to see that

$$\begin{aligned}
\mathbf{KL}(F_i||F_j) &= \frac{1}{2}\Big(\mathbb{E}_{F_j}[X^\top \Sigma_k^{-1} X] - \mathbb{E}_{F_j}[X^\top \Sigma_j^{-1} X]\Big) \\
&= \frac{1}{2}\Big(\mathbb{E}_{F_j}[X^\top \Sigma_k^{-1} X] - d\Big) \\
&= (\mathrm{Tr}(\Sigma_k^{-1}\Sigma_j) - d)/2 \\
&\leq (\sqrt{\mathrm{Tr}(\Sigma_j^{-2})\,\mathrm{Tr}(\Sigma_k^2)} - d)/2 \\
&\leq (M^2 - 1)d\,.
\end{aligned}$$

Therefore, we can establish the first lower bound that

$$n \gtrsim \frac{q \log d/q}{M^2 - 1}$$

Ensemble B For this ensemble, we consider the DAGs with exactly one edge $u \to v$ and coefficient $b_{\min}$, denoted as G^{uv}. There are 2 directions and $d(d-1)/2$ many edges, so the cardinality of this ensemble would be $N = d(d-1) \asymp d^2$. Then denote the distribution

defined according to G^{uv} as F^{uv}, the log likelihood

$$-\log F^{uv} \propto \frac{1}{2\sigma^2}\left[\sum_{i\notin\{u,v\}} X_i^2 + X_u^2 + (X_v - b_{\min}X_u)^2\right] + d\log\sigma$$

and the difference between any two cases is

$$\log F^{uv} - \log F^{jk} = \frac{1}{2\sigma^2}\left[X_v^2 + (X_k - b_{\min}X_j)^2 - X_k^2 - (X_v - b_{\min}X_u)^2\right]$$

Then take expectation over F^{uv} we get the KL divergence:

$$\mathbf{KL}(F^{uv}||F^{jk}) = \frac{1}{2\sigma^2}\left[b_{\min}^2\sigma^2 + \mathbb{E}_{F^{uv}}(b_{\min}^2X_j^2 - 2X_kX_j)\right]$$

For fixed edge (u,v), any other edges (j,k) has relationship and corresponding KL divergence below:

- $j \neq u, k \neq v$, $\mathbf{KL}(F^{uv}||F^{jk}) = b_{\min}^2$
- $j = u, k \neq v$, $\mathbf{KL}(F^{uv}||F^{jk}) = b_{\min}^2$
- $j \neq u, k = v$, $\mathbf{KL}(F^{uv}||F^{jk}) = b_{\min}^2$
- $j = v, k = u$, $\mathbf{KL}(F^{uv}||F^{jk}) = b_{\min}^2 + b_{\min}^4/2 - b_{\min}$
- $j = v, k \neq u$, $\mathbf{KL}(F^{uv}||F^{jk}) = b_{\min}^2 + b_{\min}^4/2$
- $j \neq v, k = u$, $\mathbf{KL}(F^{uv}||F^{jk}) = b_{\min}^2$

Among them, the largest KL between F^{uv}, F^{jk} is $b_{\min}^2 + b_{\min}^4/2$. Therefore, we can conclude a lower bound

$$n \gtrsim \frac{1}{b_{\min}^4 + 2b_{\min}^2} \log d \asymp \frac{\log d}{b_{\min}^2}. \qquad \square$$

We proceed to prove Lemma A.3.1.

Proof of Lemma A.3.1. The proof construction is similar to Santhanam and Wainwright [2012]. We can upper bound by number of directed graphs (DG), and lower bound by one particular subclass of DAGs.

For upper bound, note that a DG has at most d^2 many, and a sparse DAG has at most dq any edges. Since $q \leq d/2$, then we have $dq \leq d^2/2$. Since for $\ell < k \leq d^2/2$, we have $\binom{d^2}{\ell} < \binom{d^2}{k}$, and there are $\binom{d^2}{k}$ DGs with exactly k directed edges, then the numbder of DGs is upper bounded by

$$\log |DGs| = \log \sum_{k=1}^{dq} \binom{d^2}{k} \leq \log dq \binom{d^2}{dq} \asymp dq \log \frac{d}{q}$$

For lower bound, we look at one subclass of DAGs. Suppose $d/(q+1)$ is an integer, otherwise discard remaining nodes. First partition d nodes into $q+1$ groups with equal size $d/(q+1)$. Then for the first group, build directed edges from nodes in group $2, 3, \ldots, q+1$ to group 1, which requires q permutations on $d/(q+1)$ nodes within one particular group. Then the nodes in group 1 has exactly degree q. Similarly, for group 2, build directed edges from group $3, 4, \ldots, q+1$, which requires $q-1$ permutations on $d/(q+1)$ nodes. Therefore, for the subclass of DAGs generated in this way of partition, we have

$$(\frac{d}{q+1}!)^{(q)+(q-1)+\ldots+1} = (\frac{d}{q+1}!)^{q(q+1)/2}$$

many DAGs, any of which is valid DAG and has degree bounded by q. Then the cardi-

nality

$$\log |DAGs| \geq \log(\frac{d}{q+1}!)^{q(q+1)/2} \asymp dq \log \frac{d}{q}$$

Thus the total number of DAGs scales as $\Theta(dq \log \frac{d}{q})$. □

A.3.1 Sample Conditional Correlation Coefficient as CI Tester

`PC-Tree` relies on sample (conditional) correlation coefficient as (conditional) independence tester. Specifically, denote the sample covariance matrix to be $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X^{(i)} X^{(i)\top}$, for any two nodes $j, k \in V$ and any subset $S \subseteq V \setminus \{j, k\}$, which could be $\emptyset$, the sample correlation coefficient is defined by

$$\widehat{\rho}_{jk \mid S} := \frac{\widehat{\Sigma}_{jk} - \widehat{\Sigma}_{jS}\widehat{\Sigma}_{SS}^{-1}\widehat{\Sigma}_{Sk}}{\sqrt{(\widehat{\Sigma}_{jj} - \widehat{\Sigma}_{jS}\widehat{\Sigma}_{SS}^{-1}\widehat{\Sigma}_{Sj})(\widehat{\Sigma}_{kk} - \widehat{\Sigma}_{kS}\widehat{\Sigma}_{SS}^{-1}\widehat{\Sigma}_{Sk})}}$$

Then the conditional independence tester for hypothesis $H_0 : X_j \perp\!\!\!\perp X_k \mid X_S$ is given by a cutoff on the sample correlation coefficient:

$$\text{Output} = \begin{cases} \text{accept } H_0 & \text{if } |\widehat{\rho}_{jk \mid S}| \geq c/2 \\ \text{reject } H_0 & \text{if } |\widehat{\rho}_{jk \mid S}| < c/2 \end{cases}. \tag{A.1}$$

Here the choice of $c/2$ is for theoretical purpose. Since correlation coefficient is normalized between $[-1, 1]$, in practice, the tester can be implemented by choosing a cutoff that is small enough, e.g. 0.05. The analysis of `PC-Tree` crucially relies on the following lemma on the estimation error of the sample (conditional) correlation coefficients:

Lemma A.3.2. *Let $X \in \mathbb{R}^d \sim \mathcal{N}(0, \Sigma)$, for any $j, k \in V$ and any subset $S \subseteq V \setminus \{j, k\}$ with $|S| \leq q$, $\delta \in (0, 1)$, if $n \gtrsim q + 1/\delta^2$, then*

$$\mathbb{P}(|\widehat{\rho}_{jk \mid S} - \rho_{jk \mid S}| \geq \delta) \leq \exp(-C_0(n - q)\delta^2),$$

for some universal constant $C_0 > 0$.

It is clear to see that as long as the (conditional) correlation coefficients are estimated accurately enough, the CI tests are correct due to c-strong Tree-faithfulness. Lemma A.3.2 is more general than needed to analyze `PC-Tree` algorithm. Since Lemma A.3.2 reveals the dependence on the size of conditioning set S, while `PC-Tree` only requires $|S| \leq 1$.

Proof of Lemma A.3.2. The proof is a combination of the following lemmas. We start with analyzing sample marginal correlation of bivariate normal distribution, then extend to conditional correlation.

Lemma A.3.3. *Let $W = (X, Y) \sim \mathcal{N}(0, \Sigma)$ where $\Sigma = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix} \in \mathbb{R}^{2\times 2}$, and $\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$. Let the sample covariance matrix and correlation be*

$$\frac{1}{n}\sum_{\ell=1}^{n} w^{(\ell)} w^{(\ell)\top} = \begin{pmatrix} \widehat{\sigma}_X^2 & \widehat{\sigma}_{XY} \\ \widehat{\sigma}_{XY} & \widehat{\sigma}_Y^2 \end{pmatrix}, \quad \text{and} \quad \widehat{\rho} = \frac{\widehat{\sigma}_{XY}}{\widehat{\sigma}_X \widehat{\sigma}_Y}.$$

For $\delta \in (0, 1)$, if $n \gtrsim 1/\delta^2$, then

$$\mathbb{P}(|\widehat{\rho} - \rho| \geq \delta) \leq \exp(-C_0 n \delta^2),$$

for some constant $C_0 > 0$.

Now look at sample conditional correlation, suppose we want to estimate $\rho_{jk\,|\,S}$ with $|S| = q' \leq q$. Recall the sample covariance matrix is $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X^{(i)} X^{(i)\top}$. Denote $I = \{j, k\}$, then the estimator is given by 2×2 matrix

$$\widehat{\Sigma}_{II\,|\,S} := \widehat{\Sigma}_{II} - \widehat{\Sigma}_{II,S} \widehat{\Sigma}_{SS}^{-1} \widehat{\Sigma}_{S,II}.$$

We borrow a classic result regarding the distribution of $\widehat{\Sigma}_{II\,|\,S}$:

Lemma A.3.4 (Anderson [1958], Theorem 4.3.4)**.** *The sample covariance matrix $\widehat{\Sigma}_{II\,|\,S}$ is distributed as $\frac{1}{n}\sum_{\ell=1}^{n-q'} u^{(\ell)}u^{(\ell)\top}$, where $\{u^{(\ell)}\}_{\ell=1}^{n-q'}$ are independently distributed according to $\mathcal{N}(0, \Sigma_{II\,|\,S})$.*

Then applying the bivariate result from Lemma A.3.3 with covariance matrix $\Sigma_{II\,|\,S}$ and sample size $n - q' \leq n - q$ completes the proof. □

It remains to prove the lemma used in proof above.

Proof of Lemma A.3.3. Let $Z_X = X/\sigma_X$, $Z_Y = Y/\sigma_Y$, then $Z_X, Z_Y \sim \mathcal{N}(0,1)$ and $\rho_{Z_X,Z_Y} = \rho = \mathrm{cov}(Z_X, Z_Y) \in [-1,1]$. Denote the corresponding samples to be $z_X = (z_X^{(1)}, \ldots, z_X^{(n)})$ and $z_Y = (z_Y^{(1)}, \ldots, z_Y^{(n)})$, therefore

$$\widehat{\rho} = \frac{\widehat{\sigma}_{XY}}{\widehat{\sigma}_X\widehat{\sigma}_Y} = \frac{\widehat{\sigma}_{XY}/(\sigma_X\sigma_Y)}{(\widehat{\sigma}_X/\sigma_X)\times(\widehat{\sigma}_Y/\sigma_Y)} = \frac{\langle z_X, z_Y\rangle}{\|z_X\|\|z_Y\|}.$$

Then the deviation

$$\begin{aligned}
|\widehat{\rho} - \rho| &= \left|\frac{\langle z_X, z_Y\rangle}{\|z_X\|\|z_Y\|} - \mathrm{cov}(Z_X, Z_Y)\right| \\
&\leq \left|\frac{\langle z_X, z_Y\rangle/n}{\|z_X\|\|z_Y\|/n} - \frac{\mathrm{cov}(Z_X, Z_Y)}{\|z_X\|\|z_Y\|/n} + \frac{\mathrm{cov}(Z_X, Z_Y)}{\|z_X\|\|z_Y\|/n} - \mathrm{cov}(Z_X, Z_Y)\right| \\
&\leq \left|\frac{1}{\|z_X\|\|z_Y\|/n} - 1\right|\left|\langle z_X, z_Y\rangle/n - \mathrm{cov}(Z_X, Z_Y)\right| + \left|\langle z_X, z_Y\rangle/n - \mathrm{cov}(Z_X, Z_Y)\right| \\
&\quad + \left|\mathrm{cov}(Z_X, Z_Y)\right|\left|\frac{1}{\|z_X\|\|z_Y\|/n} - 1\right|.
\end{aligned}$$

We apply the following lemma to bound the errors:

Lemma A.3.5. *If $(X, Y) \sim \mathcal{N}\Big(0, \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}\Big)$ for $|r| \leq 1$, then the sample variance $\widehat{\sigma}_X^2 = \frac{1}{n}\sum_{i=1}^n {X^{(i)}}^2$, $\widehat{\sigma}_Y^2 = \frac{1}{n}\sum_{i=1}^n {Y^{(i)}}^2$ and sample covariance $\widehat{\sigma}_{XY} = \frac{1}{n}\sum_{i=1}^n X^{(i)}Y^{(i)}$ have the fol-*

lowing bounds: for $\zeta < 1$, *if* $n \geq \frac{2048 \log 7}{\zeta^2}$, *then*

$$\begin{aligned}
\mathbb{P}(|\widehat{\sigma}_X^2 - 1| \geq \zeta) &\leq \exp(-n\zeta^2/16) \\
\mathbb{P}(|\widehat{\sigma}_Y^2 - 1| \geq \zeta) &\leq \exp(-n\zeta^2/16) \\
\mathbb{P}(|\widehat{\sigma}_{XY} - r| \geq \zeta) &\leq \exp(-n\zeta^2/2048)\,.
\end{aligned}$$

Using Lemma A.3.5, with probability at least $1 - 3\exp(-n\zeta^2/2048)$, we have

$$\begin{aligned}
|\|z_X\|^2/n - 1| &\leq \zeta \\
|\|z_Y\|^2/n - 1| &\leq \zeta \\
|\langle z_X, z_Y\rangle/n - \mathrm{cov}(Z_X, Z_Y)| &\leq \zeta\,.
\end{aligned}$$

Then

$$\left|\frac{1}{\|z_X\|\|z_Y\|/n} - 1\right| = \frac{|\|z_X\|\|z_Y\|/n - 1|}{\|z_X\|\|z_Y\|/n} \leq \frac{\zeta}{1-\zeta}\,.$$

Choose $\zeta = \frac{\delta}{3+\delta}$, then $\left|\frac{1}{\|z_X\|\|z_Y\|/n} - 1\right| \leq \delta/3$, $\left|\langle z_X, z_Y\rangle/n - \mathrm{cov}(Z_X, Z_Y)\right| \leq \delta/(3+\delta) \leq \delta/3$. Lastly,

$$|\widehat{\rho} - \rho| \leq \frac{\delta}{3} \times \frac{\delta}{3} + \frac{\delta}{3} + \frac{\delta}{3} \leq \delta\,,$$

with probability at least

$$\begin{aligned}
1 - 3\exp(-n\zeta^2/2048) &= 1 - \exp\Big(-n \times \frac{\delta^2}{(3+\delta)^2}/2048 + \log 3\Big) \\
&\geq 1 - \exp\Big(-n \times \frac{\delta^2}{16 \times 2048} + \log 3\Big) \\
&\geq 1 - \exp(-C_0 n\delta^2)\,,
\end{aligned}$$

for some constant $C_0 > 0$ as long as $n \gtrsim 1/\delta^2$. □

Proof of Lemma A.3.5. We only show variance bound for X. Since $\widehat{\sigma}_X^2 \sim \chi_n^2/n$, using the concentration of χ^2 distribution, we have

$$\mathbb{P}(|\widehat{\sigma}_X^2 - 1| \geq \zeta) = \mathbb{P}(|\chi_n^2 - n|/n \geq \zeta) \leq \exp(-n\zeta^2/16)\,.$$

Now we show bound for covariance. Since bivariate Gaussian (X, Y) can be reparameterized by

$$X = U + W$$
$$Y = V + W$$

where U,V,W are mutually independent with $\mathrm{var}(U) = \mathrm{var}(V) = 1 - r$, $\mathrm{var}(W) = r$. Therefore,

$$\begin{aligned}
&\widehat{\sigma}_{XY} \\
&= \frac{1}{n}\sum_{i=1}^n U^{(i)}V^{(i)} + \frac{1}{n}\sum_{i=1}^n U^{(i)}W^{(i)} + \frac{1}{n}\sum_{i=1}^n V^{(i)}W^{(i)} + \frac{1}{n}\sum_{i=1}^n W^{(i)2} \\
&= \frac{1-r}{2n}\Big[\sum_{i=1}^n \Big(\frac{U'^{(i)} + V'^{(i)}}{\sqrt{2}}\Big)^2 - \sum_{i=1}^n \Big(\frac{U'^{(i)} - V'^{(i)}}{\sqrt{2}}\Big)^2\Big] \\
&\quad + \frac{\sqrt{r(1-r)}}{2n}\Big[\sum_{i=1}^n \Big(\frac{U'^{(i)} + W'^{(i)}}{\sqrt{2}}\Big)^2 - \sum_{i=1}^n \Big(\frac{U'^{(i)} - W'^{(i)}}{\sqrt{2}}\Big)^2\Big] \\
&\quad + \frac{\sqrt{r(1-r)}}{2n}\Big[\sum_{i=1}^n \Big(\frac{V'^{(i)} + W'^{(i)}}{\sqrt{2}}\Big)^2 - \sum_{i=1}^n \Big(\frac{V'^{(i)} - W'^{(i)}}{\sqrt{2}}\Big)^2\Big] + \frac{r}{n}\sum_{i=1}^n W'^{(i)2} \\
&\overset{\mathcal{D}}{\sim} \frac{1-r}{2n}(\chi_{n11}^2 - \chi_{n12}^2) + \frac{\sqrt{r(1-r)}}{2n}(\chi_{n21}^2 - \chi_{n22}^2) + \frac{\sqrt{r(1-r)}}{2n}(\chi_{n31}^2 - \chi_{n32}^2) + \frac{r}{n}\chi_{n4}^2
\end{aligned}$$

where U', V', W' are standard normal random variables, thus $\sum_{i=1}^n (U'^{(i)} \pm V'^{(i)})^2/2$,

$\sum_{i=1}^{n}(U'^{(i)} \pm W'^{(i)})^2/2$, $\sum_{i=1}^{n}(V'^{(i)} \pm W'^{(i)})^2/2$ are χ^2_n random variables. Since $r \leq 1$,

$$\begin{aligned}
\mathbb{P}(|\widehat{\sigma}_{XY} - r| \geq \zeta) &\leq \mathbb{P}\Big(\frac{1-r}{2} \times \frac{1}{n}|\chi^2_{n11} - \chi^2_{n12}| \geq \zeta/4\Big) \\
&\quad + \mathbb{P}\Big(\frac{\sqrt{r(1-r)}}{2} \times \frac{1}{n}|\chi^2_{n21} - \chi^2_{n22}| \geq \zeta/4\Big) \\
&\quad + \mathbb{P}\Big(\frac{\sqrt{r(1-r)}}{2} \times \frac{1}{n}|\chi^2_{n31} - \chi^2_{n32}| \geq \zeta/4\Big) \\
&\quad + \mathbb{P}\Big(r \times |\chi^2_{n41}/n - 1| \geq \zeta/4\Big) \\
&\leq \mathbb{P}\Big(|\chi^2_{n11}/n - 1| \geq \zeta/8\Big) + \mathbb{P}\Big(|\chi^2_{n12}/n - 1| \geq \zeta/8\Big) \\
&\quad + \mathbb{P}\Big(|\chi^2_{n21}/n - 1| \geq \zeta/8\Big) + \mathbb{P}\Big(|\chi^2_{n22}/n - 1| \geq \zeta/8\Big) \\
&\quad + \mathbb{P}\Big(|\chi^2_{n31}/n - 1| \geq \zeta/8\Big) + \mathbb{P}\Big(|\chi^2_{n32}/n - 1| \geq \zeta/8\Big) \\
&\quad + \mathbb{P}\Big(|\chi^2_{n41}/n - 1| \geq \zeta/4\Big) \\
&\leq 7\exp(-n\zeta^2/32^2) \leq \exp(-n\zeta^2/2048)\,.
\end{aligned}$$

The last inequality holds when $n \geq 2048 \log 7/\zeta^2$. □

A.4 Proof of Theorem 2.3.1 (Upper Bound for Faithful Poly-tree)

Proof. We firstly show the correctness of Algorithm 2. We make following notation of sets of nodes:

- $W = \{(j,k) : 1 \leq j < k \leq d\}$ is the set of all pairs of nodes in $[d]$;
- E is the true edge set;
- $A = \{(j,k) : j \text{ and } k \text{ are d-separated by } \emptyset\}$;
- $B = \{(j,k) : \exists \ell \in [d] \setminus \{j,k\}, j \text{ and } k \text{ are d-separated by } \ell\}$
- $C = \{(j,k) : \exists \ell \in [d] \setminus \{j,k\}, j \to \ell \leftarrow k \text{ is a } v\text{-structure}\}$

- $D = \{(j,k) : \exists \ell \in [d] \setminus \{j,k\}, j - \ell - k$ is a unshielded triple but not a v-structure$\}$

We claim that

1. E and $A \cup B$ are disjoint;
2. $W = E \cup A \cup B$;
3. $C \subseteq A$;
4. $D \subseteq B$.

It is easy to see the first claim, since for any pair of nodes connected by an edge, they cannot be d-separated by any set, and vice versa.

For the second claim, it suffices to show that for any pair of nodes not adjacent, it is in either A or B. First of all, for any two nodes j and k not adjacent, there will be one and only one path, denoted as ϕ, with length at least two between them. By property of poly-tree:

- If there is a collider on ϕ, then the path is blocked by $\emptyset$, so $(j,k) \in A$;
- If there is no collider on ϕ, then any node on ϕ will block the path, thus there exists $\ell \in [d] \setminus \{j,k\}$ such that i and j are d-separated by ℓ, so $(j,k) \in B$.

For the third claim, since $j \to \ell \leftarrow k$ is the only path between (j,k), which is blocked by $\emptyset$, thus $C \subseteq A$. For the forth claim, since $j - \ell - k$ is the only path between (j,k), either one of $j \to \ell \to k$ and $j \leftarrow \ell \leftarrow k$ and $j \leftarrow \ell \to k$ will be blocked by ℓ, thus $D \subseteq B$.

We now claim if the CI tests in Step 2 of Algorithm 2 are correct for

- all pairs $(j,k) \in E$ with $\ell \in [d] \cup \{\emptyset\} \setminus \{j,k\}$;
- all pairs $(j,k) \in A$ with $\ell = \emptyset$;
- all pairs $(j,k) \in C$ with ℓ being the collider;

- all pairs $(j,k) \in B$ with ℓ being the corresponding separation node(s) ,

then

1. the returned $\widehat{T}$ has the correct edge set E thus is the correct skeleton;
2. for any $(j,k) \in C$, $\ell \notin S(j,k)$;
3. for any $(j,k) \in D$, $\ell \in S(j,k)$.

For the first claim, if the CI tests conducted in Step 2 are correct for E, then pairs in E will pass all the CI tests and be included into $\widehat{E}$ (which is ensured by adjacency-faithfulness in Tree-faithfulness). But pairs in A will not pass marginal independence tests, and pairs in B will not pass some CI tests with corresponding ℓ (which is ensured by Markov property). Therefore, the returned $\widehat{G}$ is the correct skeleton. The second claim is ensured by orientation-faithfulness in Tree-faithfulness, and the third claim is ensured by Markov property and $D \subseteq B$.

Once the returned $\widehat{G}$ is the correct skeleton, Algorithm 6 will use the returned separation sets to determine v-structure for each possible unshielded triple. Note that

$$\{\text{All unshielded triples}\} = C \cup D\,.$$

For any $(j,k) \in C$, $\ell \notin S(j,k)$, thus it will be oriented as a v-structure; For any $(j,k) \in D$, $\ell \in S(j,k)$; thus it will remain as non-v-structure. Then ORIENT step is correct, which leads to correct CPDAG.

Finally we show the sample complexity of Algorithm 2 with CI tester (A.1). Note that

correct CI tests implies correct estimation. Therefore,

$$
\begin{aligned}
&\mathbb{P}(\widehat{G} \neq \mathrm{sk}(G)) \\
&\leq \mathbb{P}\Bigg(\cup_{\substack{(j,k)\in E \\ \ell\in[d]\cup\{\emptyset\}\setminus\{j,k\}}} \text{ or } \cup_{\substack{(j,k)\in A \\ \ell=\emptyset}} \text{ or } \cup_{\substack{(j,k)\in C \\ \ell=\text{collider}}} \text{ or } \cup_{\substack{(j,k)\in B \\ \ell \text{ d-separates } (j,k)}} |\widehat{\rho}_{ij\,|\,\ell} - \rho_{ij\,|\,\ell}| \geq c/2 \Bigg) \\
&\leq \binom{d}{2} \times (1 + (d-2)) \times \sup_{\substack{(j,k)\in E \\ \ell\in[d]\cup\{\emptyset\}\setminus\{j,k\}}} \text{ or } \sup_{\substack{(j,k)\in A \\ \ell=\emptyset}} \text{ or } \sup_{\substack{(j,k)\in C \\ \ell=\text{collider}}} \text{ or } \sup_{\substack{(j,k)\in B \\ \ell \text{ d-separates } (j,k)}} \mathbb{P}(|\widehat{\rho}_{ij\,|\,\ell} - \rho_{ij\,|\,\ell}| \geq c/2) \\
&\leq \exp(3\log d) \times \sup_{\substack{(j,k)\in E \\ \ell\in[d]\cup\{\emptyset\}\setminus\{j,k\}}} \text{ or } \sup_{\substack{(j,k)\in A \\ \ell=\emptyset}} \text{ or } \sup_{\substack{(j,k)\in C \\ \ell=\text{collider}}} \text{ or } \sup_{\substack{(j,k)\in B \\ \ell \text{ d-separates } (j,k)}} \mathbb{P}(|\widehat{\rho}_{ij\,|\,\ell} - \rho_{ij\,|\,\ell}| \geq c/2) \\
&\leq \exp\Bigg(-C_0(n-1)c^2 + 3\log d \Bigg).
\end{aligned}
$$

The first inequality is because it suffices to have $|\widehat{\rho}_{ij\,|\,\ell} - \rho_{ij\,|\,\ell}| \leq c/2$ for correct CI test. By c-strong Tree-faithfulness, $|\rho_{ij\,|\,S}| \geq c$ for $\rho_{ij\,|\,S} \neq 0$. Therefore,

$$
\begin{cases}
\widehat{\rho}_{ij\,|\,S} > c/2 & \text{if } \rho_{ij\,|\,S} \neq 0 \\
\widehat{\rho}_{ij\,|\,S} \leq c/2 & \text{if } \rho_{ij\,|\,S} = 0
\end{cases}
$$

Thus the cutoff $= c/2$ implies correct CI tests. The last inequality is by Lemma A.3.2 where $q = 1$ and the sample size requirement is satisfied by the stated sample complexity. Set RHS to be smaller than δ, we need sample complexity

$$
n \gtrsim \frac{1}{c^2}\left(\log d + \log \frac{1}{\delta} \right),
$$

which completes the proof. □

A.5 Proof of Proposition 2.3.2

Proof. Since a directed tree G does not have any v-structures, we only need to verify adjacency faithfulness in Definition 4. For any two nodes connected as $j \to k$, we want to

check whether $\rho(X_j, X_k \mid X_\ell)$ is lower bounded by some constant for $\ell \in V \cup \{\emptyset\} \setminus \{j, k\}$. There are four cases of ℓ to consider, see Figure A.1:

- $\ell = \emptyset$: To simplify the notation, we write

$$X_k = \beta_k \times X_j + \eta_k$$

with $\beta_k \in \mathbb{R}$ and $|\beta_k| \in [M^{-1}, M]$, $\mathrm{var}(\eta_k) = \sigma_k^2$. We also write $V_j^2 := \mathrm{var}(X_j) \geq \sigma_j^2$. Hence,

$$\rho(X_j, X_k) = \frac{\beta_k V_j^2}{\sqrt{V_j^2}\sqrt{\beta_k^2 V_j^2 + \sigma_k^2}} = \frac{1}{\sqrt{1 + \sigma_k^2/\beta_k^2 V_j^2}} \geq \frac{1}{\sqrt{1 + \sigma_k^2/\beta_k^2 \sigma_j^2}} \gtrsim 1\,.$$

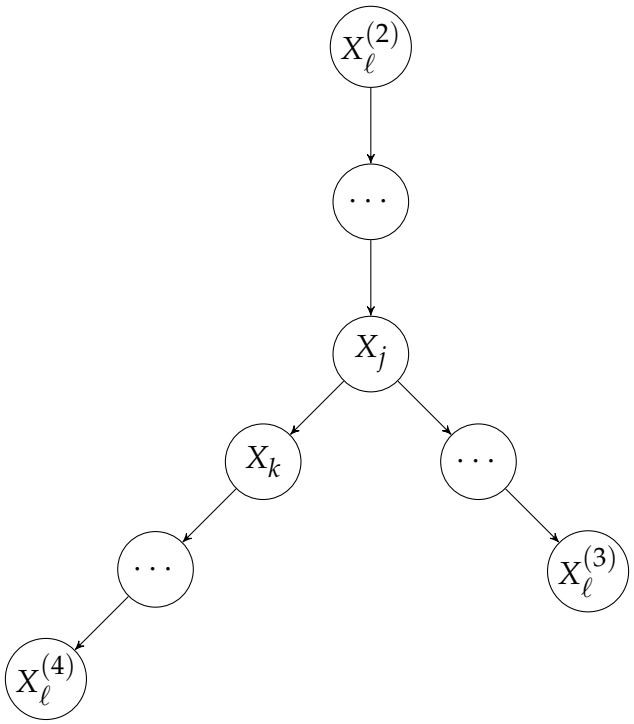


Figure A.1: Four cases of ℓ to verify for c-strong Tree-faithfulness, indicated by the superscript of X_ℓ. The first case is when $\ell = \emptyset$. The second, third and fourth are when ℓ is the ancestor of j, descendant of j and descendant of k.

- $\ell \in \mathrm{an}(j)$: Write $V^2_{j\,|\,\ell} = \mathrm{var}(X_j \,|\, X_\ell) \geq \sigma_j^2$, hence

$$\rho(X_j, X_k \,|\, X_\ell) = \frac{\beta_k V^2_{j\,|\,\ell}}{\sqrt{V^2_{j\,|\,\ell}}\sqrt{\beta_k^2 V^2_{j\,|\,\ell} + \sigma_k^2}} \geq \frac{1}{\sqrt{1 + \sigma_k^2/\beta_k^2\sigma_j^2}} \gtrsim 1\,.$$

- $\ell \in \mathrm{de}(j)$: Suppose the directed path from j to ℓ is $j \to h_1 \to h_2 \to \ldots \to h_q \to \ell$, q can be 0, then we can write

$$X_\ell = b_1 X_j + u_1\,,$$

with

$$b_1 = \beta_\ell \prod_{i=1}^{q} \beta_{h_i}, \qquad u_1 = \eta_\ell + \beta_\ell \sum_{i=1}^{q} \eta_{h_i} \prod_{t=i+1}^{q} \beta_{h_t}\,,$$

and

$$\nu_1^2 := \mathrm{var}(u_1) = \sigma_\ell^2 + \beta_\ell^2 \sum_{i=1}^{q} \sigma_{h_i}^2 \prod_{t=i+1}^{q} \beta_{h_t}^2 \geq \beta_\ell^2 \sigma_{h_1}^2 \prod_{t=2}^{q} \beta_{h_t}^2\,.$$

So we have $b_1^2/\nu_1^2 \leq \beta_{h_1}^2/\sigma_{h_1}^2 \asymp 1$. The covariance among X_j, X_k, X_ℓ is

$$\mathrm{cov}(X_j, X_k, X_\ell) = \begin{pmatrix} V_j^2 & \beta_k V_j^2 & b_1 V_j^2 \\ * & \beta_k^2 V_j^2 + \sigma_k^2 & b_1 \beta_k V_j^2 \\ * & * & b_1^2 V_j^2 + \nu_1^2 \end{pmatrix}.$$

Then the conditional covariance is

$$\mathrm{cov}(X_j, X_k \,|\, X_\ell) \propto \begin{pmatrix} \nu_1^2 & \beta_k \nu_1^2 \\ * & \beta_k^2 \nu_1^2 + \sigma_k^2 b_1^2 + \sigma_k^2 \nu_1^2 / V_j^2 \end{pmatrix}.$$

Therefore,

$$\rho(X_j, X_k \mid X_\ell) = \frac{1}{\sqrt{1 + \frac{\sigma_k^2}{\beta_k^2} \times \frac{b_1^2}{\nu_1^2} + \frac{\sigma_k^2}{V_j^2 \beta_k^2}}} \gtrsim 1 \,.$$

- $\ell \in \mathrm{de}(k)$: Similarly, we can write

$$X_\ell = b_2 X_k + u_2, \qquad \mathrm{var}(u_2) = \nu_2^2 \,,$$

with $b_2^2/\nu_2^2 \lesssim 1$. The covariance among X_j, X_k, X_ℓ is

$$\mathrm{cov}(X_j, X_k, X_\ell) = \begin{pmatrix} V_j^2 & \beta_k V_j^2 & b_2 \beta_k V_j^2 \\ * & V_k^2 & b_2 V_k^2 \\ * & * & b_2^2 V_k^2 + \nu_2^2 \end{pmatrix} .$$

Then the conditional covariance is

$$\mathrm{cov}(X_j, X_k \mid X_\ell) \propto \begin{pmatrix} b_2^2 \sigma_k^2 V_j^2 + \nu_2^2 V_j^2 & \beta_k V_j^2 \nu_2^2 \\ * & \nu_2^2 V_k^2 \end{pmatrix} .$$

Therefore,

$$\rho(X_j, X_k \mid X_\ell) = \frac{1}{\sqrt{(1 + \frac{\sigma_k^2}{\beta_k^2 V_j^2})(1 + \frac{b_2^2}{\nu_2^2} \sigma_k^2)}} \gtrsim 1 \,.$$

In all four cases, $\rho(X_j, X_k \mid X_\ell) \gtrsim 1$, thus c-strong Tree-faithfulness is satisfied with some $c \asymp 1$. □

A.6 Proof of Theorem 2.3.3 (Lower Bound for Faithful Poly-tree)

Proof. We construct a hard ensemble to show the lower bound. The construction is as follows. Consider a subset $\mathcal{T}'' \subset \mathcal{T}' \subset \mathcal{T}$, where $\mathcal{T}''$ is all the directed trees rooted at the first node $k = 1$. $\mathcal{T}''$ has the same cardinality as all undirected trees with d nodes, and the elements in it have different skeletons and no v-structures. Since our target is MEC, which is determined by its skeleton and v-structures, we have at least as many MECs as undirected trees, which leads to cardinality $|\mathcal{T}''| = d^{d-2}$ using Cayley's formula. Thus the size of the ensemble is lower bounded as

$$\log |\mathcal{T}''| = (d-2)\log d \geq \frac{1}{2} d \log d$$

The inequality holds when d is large enough, e.g. $d \geq 4$. Any directed tree has an important property: each node has at most one parent. Then we parameterize $\mathcal{T}''$ as follows

$$X_k = \beta X_{\mathrm{pa}(k)} + \eta_k\,, \quad \forall k \in [d] \tag{A.2}$$

where $\eta_k \sim \mathcal{N}(0,1)$ for all $k \in [d]$. Now we determine $\beta > 0$ to make sure the parametrization satisfies c-strong Tree-faithfulness.

In the subsequent lemma, we assert that the condition $\beta^2 = 2c^2 \asymp c^2$ is adequate for the validity of c-strong Tree-faithfulness, provided that c is sufficiently small:

Lemma A.6.1. *If $\beta = \sqrt{2}c$ and $c^2 \leq 1/5$, then for any $T \in \mathcal{T}''$, the distribution defined in* (A.2)*:*

1. *is c-strong Tree-faithful to T;*
2. *for all $k \in [d]$, $\mathrm{var}(X_k) \leq 1 + \frac{\beta^2}{1-\beta^2}$.*

It remains to bound the KL divergence between any two instances in this ensemble. Before that, we claim that for any instance, we have $\mathrm{cov}(X_k, X_j) > 0$ for all distinct $j, k \in$

$[d]$. This is because for any pair of distinct nodes (j,k), there can be 3 possible paths between them:

- There is a directed path $j \to \phi_1 \to \cdots \to \phi_h \to k$ with length $h+1$, then $\mathrm{cov}(X_j, X_k) = \mathbb{E}[X_j X_k] = \beta^{h+1}\mathbb{E}[X_j^2] > 0$;
- There is a directed path $k \to \phi_1 \to \cdots \to \phi_h \to j$ with length $h+1$, then $\mathrm{cov}(X_j, X_k) = \mathbb{E}[X_j X_k] = \beta^{h+1}\mathbb{E}[X_k^2] > 0$;
- j, k share a common ancestor ℓ and there is a path $j \leftarrow \phi_1 \leftarrow \cdots \leftarrow \phi_h \leftarrow \ell \to \varphi_1 \to \cdots \to \varphi_g \to k$, then $\mathrm{cov}(X_j, X_k) = \mathbb{E}[X_j X_k] = \beta^{h+g+2}\mathbb{E}[X_\ell^2] > 0$.

To compute the KL divergence between distributions P_0 and P_1 induced by any two $T_0, T_1 \in \mathcal{T}''$, let's first look at the covariance matrices of them Σ_0, Σ_1. Under our parametrization, they share the same determinant. To see this, let covariance matrix of η be $\Sigma_\eta = I_d$, for $\ell \in \{0,1\}$, $\det(\Sigma_\ell) = \det(\Sigma_\eta) = \det(I_d) = 1$. Then the KL divergence is:

$$\begin{aligned} D_{\mathrm{KL}}(P_0 \| P_1) &= \mathbb{E}_{P_0} \log \frac{P_0}{P_1} \\ &= \mathbb{E}_{P_0} \log \frac{\exp\left(-\frac{1}{2}\sum_{k=1}^d (X_k - \beta \mathrm{pa}_{T_0}(k))^2 \right) / \sqrt{\det(\Sigma_0)}}{\exp\left(-\frac{1}{2}\sum_{k=1}^d (X_k - \beta \mathrm{pa}_{T_1}(k))^2 \right) / \sqrt{\det(\Sigma_1)}} \\ &= \frac{1}{2}\left[\mathbb{E}_{P_0} \sum_{k=1}^d (X_k - \beta \mathrm{pa}_{T_1}(k))^2 - d \right]. \end{aligned}$$

For all $k \in [d]$, let $\mathrm{pa}_{T_1}(k) = j$, then

$$\begin{aligned}
\mathbb{E}_{P_0}(X_k - \beta \mathrm{pa}_{T_1}(k))^2 &= \mathbb{E}_{P_0}[X_k^2] + \beta^2 \mathbb{E}_{P_0}[X_j^2] - 2\beta \mathbb{E}_{P_0}[X_k X_j] \\
&\le \mathbb{E}_{P_0}[X_k^2] + \beta^2 \mathbb{E}_{P_0}[X_j^2] \\
&\le (1+\beta^2)\left(1 + \frac{\beta^2}{1-\beta^2}\right) \\
&= 1 + \frac{2\beta^2}{1-\beta^2}\,.
\end{aligned}$$

The first inequality is because all covariances are positive; the second one is due to the upper bound for all variances. Thus, we have

$$D_{\mathrm{KL}}(P_0 \| P_1) \le \frac{1}{2}\left(d + \frac{2d\beta^2}{1-\beta^2} - d\right) = d\beta^2 \times \frac{1}{1-\beta^2} \le 2d\beta^2 = 4dc^2$$

The last inequality holds when β^2 is small enough, e.g. $\beta^2 \le 1/2$. The proof follows from applying Fano's inequality with KL divergence upper bound $4dc^2$ and cardinality of ensemble lower bound $\frac{1}{2} d \log d$. □

We end by proving the lemma used in the lower bound proof.

Proof of Lemma A.6.1. Since for any $T \in \mathcal{T}''$, there is no v-structure because each node has at most one parent, thus it suffices to show the first part of Definition 4.

We first show all marginal variances are bounded, i.e. $1 \le \mathrm{var}(X_k) \le 1 + \beta^2/(1-\beta^2)$ for all $k \in [d]$. Starting from the root node r, whose variance is $\mathrm{var}(X_r) = \mathrm{var}(\eta_r) = 1$, we can compute the variances of its children, they are all $\mathrm{var}(X_\ell) = \mathrm{var}(\eta_\ell) + \beta^2 \mathrm{var}(X_r) = 1 + \beta^2$ for all $\ell \in \mathrm{ch}(r)$. Proceed the calculation, $\mathrm{var}(X_j) = \mathrm{var}(\eta_j) + \beta^2 \mathrm{var}(X_\ell) = 1 + \beta^2 + \beta^4$ for all $j \in \mathrm{ch}(\ell)$ and $\ell \in \mathrm{ch}(r)$. Therefore, because the longest path has length

at most $d-1$,

$$1 \le \mathrm{var}(X_k) \le 1 + \beta^2 + \beta^4 + \cdots + \beta^{2d}$$
$$= 1 + \frac{\beta^2}{1-\beta^2} \times (1 - \beta^{2(d-1)}) \le 1 + \frac{\beta^2}{1-\beta^2}, \ \ \forall k \in [d]$$

Now we can show the marginal correlation is lower bounded for any adjacent nodes (j,k). Without loss of generality, let $j = \mathrm{pa}(k)$, then $X_k = \beta X_j + \eta_k$, and the correlation

$$\rho(X_j, X_k) = \frac{\mathbb{E}[X_j X_k]}{\sqrt{\mathrm{var}(X_k)\,\mathrm{var}(X_j)}} = \beta \sqrt{\frac{\mathbb{E}[X_j^2]}{1 + \beta^2 \mathbb{E}[X_j^2]}}$$

Thus $\rho(X_j, X_k) \ge c \Leftrightarrow \beta^2 \mathbb{E}[X_j^2] \ge \frac{c^2}{1-c^2}$. Since $\mathbb{E}(X_j^2) \ge 1$, then $\beta^2 \mathbb{E}(X_j^2) \ge \beta^2 = 2c^2 \ge \frac{c^2}{1-c^2}$ when $c^2 \le 1/2$. Now consider any pair of adjacent nodes (j,k), assuming $j = \mathrm{pa}(k)$, and any other node $\ell \in [d] \setminus \{j,k\}$, there are 4 cases on the relation between ℓ and (j,k):

1. ℓ is ancestor of j, i.e. a directed path ϕ: $\ell \to \phi_1 \to \cdots \to \phi_h \to j$;
2. j and ℓ share the same ancestor w, i.e. a directed path ϕ: $w \to \phi_1 \to \cdots \to \phi_h \to j$ and a directed path φ: $w \to \varphi_1 \to \cdots \to \varphi_g \to \ell$;
3. ℓ is a descendant of k, i.e. a directed path ϕ: $j \to k \to \phi_1 \to \cdots \to \phi_h \to \ell$;
4. ℓ is a descendant of j but not k, i.e. a directed path ϕ: $j \to \phi_1 \to \cdots \to \phi_h \to \ell$ not going through k;

where $h \ge 0$ in either case. We deal with them separately:

- For the first and second case, because $X_\ell \perp\!\!\!\perp \eta_k$, the conditional correlation is

$$\begin{aligned}\rho(X_k, X_j \mid X_\ell) &= \frac{\mathbb{E}[X_k X_j \mid X_\ell]}{\sqrt{\mathbb{E}(X_k^2 \mid X_\ell)\mathbb{E}(X_j^2 \mid X_\ell)}} \\ &= \frac{\beta \mathbb{E}[X_j^2 \mid X_\ell]}{\sqrt{\mathbb{E}(X_j^2 \mid X_\ell)(1+\beta^2\mathbb{E}(X_j^2 \mid X_\ell))}} \\ &= \sqrt{\frac{\beta^2 \mathbb{E}[X_j^2 \mid X_\ell]}{1+\beta^2\mathbb{E}(X_j^2 \mid X_\ell)}}\end{aligned}$$

  Thus $\rho(X_k, X_j \mid X_\ell) \geq c \Leftrightarrow \beta^2\mathbb{E}(X_j^2 \mid X_\ell) \geq \frac{c^2}{1-c^2}$. Since $X_{\phi_h} \perp\!\!\!\perp \eta_j \mid X_\ell$, we have $\mathbb{E}(X_j^2 \mid X_\ell) = 1 + \beta^2\mathbb{E}(X_{\phi_h}^2 \mid X_\ell) \geq 1$, then $\beta^2\mathbb{E}(X_j^2 \mid X_\ell) \geq \beta^2 = 2c^2 \geq \frac{c^2}{1-c^2}$ when $c^2 \leq 1/2$.

- For the third case, denote $v = \mathbb{E}[X_j^2]$, let's compute the covariance matrix of (X_k, X_k, X_ℓ):

$$\begin{pmatrix} v & \beta v & \beta^{h+2}v \\ \beta v & \beta^2 v + 1 & \beta^{h+1}(\beta^2 v + 1) \\ \beta^{h+2}v & \beta^{h+1}(\beta^2 v + 1) & \beta^{2(h+2)}v + \beta^{2(h+1)} + \cdots + \beta^2 + 1 \end{pmatrix}$$

  Denote $V(v, h) = \beta^{2(h+2)}v + \beta^{2(h+1)} + \cdots + \beta^2 + 1$. The covariance matrix of

(X_j, X_k) given X_ℓ is

$$
\begin{aligned}
&\begin{pmatrix} v & \beta v \\ \beta v & \beta^2 v + 1 \end{pmatrix} - \frac{1}{V(v,h)} \begin{pmatrix} \beta^{2(h+2)} v^2 & \beta^{2h+3} v(\beta^2 v + 1) \\ \beta^{2h+3} v(\beta^2 v + 1) & \beta^{2(h+1)} (\beta^2 v + 1)^2 \end{pmatrix} \\
=&\frac{1}{V(v,h)} \Bigg[\begin{pmatrix} \beta^{2(h+2)} v^2 + \beta^{2(h+1)} v + \cdots + \beta^2 v + v & \beta^{2(h+2)+1} v^2 + \beta^{2(h+1)+1} v + \cdots + \beta^3 v + \beta v \\ \beta^{2(h+2)+1} v^2 + \beta^{2(h+1)+1} v + \cdots + \beta^3 v + \beta v & (\beta^{2(h+2)+2} v^2 + \beta^{2(h+1)+2} v + \cdots + \beta^4 v + \beta^2 v \\ & \quad + \beta^{2(h+2)} v + \beta^{2(h+1)} + \cdots + \beta^2 + 1) \end{pmatrix} \\
&- \begin{pmatrix} \beta^{2(h+2)} v^2 & \beta^{2h+3} v(\beta^2 v + 1) \\ \beta^{2h+3} v(\beta^2 v + 1) & \beta^{2(h+1)} (\beta^2 v + 1)^2 \end{pmatrix} \Bigg] \\
=&\frac{1}{V(v,h)} \begin{pmatrix} (\beta^{2(h+1)} + \beta^{2h} + \cdots + \beta^2 + 1) v & (\beta^{2h} + \beta^{2(h-1)} \cdots + \beta^2 + 1) \beta v \\ (\beta^{2h} + \beta^{2(h-1)} \cdots + \beta^2 + 1) \beta v & (\beta^{2h} + \beta^{2(h-1)} + \cdots + \beta^2 + 1)(\beta^2 v + 1) \end{pmatrix}
\end{aligned}
$$

Thus the conditional correlation is

$$
\begin{aligned}
\rho(X_j, X_k \,|\, X_\ell) &= \frac{\beta v \times \frac{1-\beta^{2(h+1)}}{1-\beta^2}}{\sqrt{v \times \frac{1-\beta^{2(h+2)}}{1-\beta^2} \times (1 + \beta^2 v) \times \frac{1-\beta^{2(h+1)}}{1-\beta^2}}} \\
&= \sqrt{\frac{\beta^2 v}{1 + \beta^2 v} \times \frac{1 - \beta^{2(h+1)}}{1 - \beta^{2(h+2)}}}
\end{aligned}
$$

Denote $f(h) = \frac{1-\beta^{2(h+1)}}{1-\beta^{2(h+2)}} = 1 - \frac{(1-\beta^2)\beta^{2(h+1)}}{1-\beta^{2(h+2)}}$, which is increasing in h with minimum value being $f(0) = \frac{1}{1+\beta^2}$. Therefore, $\rho(X_j, X_k \,|\, X_\ell) \geq c \Leftrightarrow \beta^2 v \geq \frac{c^2}{f(h)-c^2}$. Since $v = \mathbb{E}[X_j^2] \geq 1$ for all $j \in [d]$, then $\beta^2 v \geq \beta^2 = 2c^2 \geq \frac{c^2}{1/(1+2c^2)-c^2} \geq \frac{c^2}{f(h)-c^2}$ when $c^2 \leq 1/5$, which yields the bound.

- For the forth case, analogously, denote $v = \mathbb{E}[X_j^2]$, let's compute the covariance

matrix of (X_k, X_k, X_ℓ):

$$\begin{pmatrix} v & \beta v & \beta^{h+1}v \\ \beta v & \beta^2 v + 1 & \beta^{h+2}v \\ \beta^{h+1}v & \beta^{h+2}v & \beta^{2(h+1)} + \beta^{2h} + \cdots + \beta^2 + 1 \end{pmatrix}$$

Denote $W(v,h) = \beta^{2(h+1)} + \beta^{2h}v + + \cdots + \beta^2 + 1$. The covariance matrix of (X_j, X_k) given X_ℓ is

$$\begin{aligned}
&\begin{pmatrix} v & \beta v \\ \beta v & \beta^2 v + 1 \end{pmatrix} - \frac{1}{W(v,h)} \begin{pmatrix} \beta^{2(h+1)}v^2 & \beta^{2h+3}v^2 \\ \beta^{2h+3}v^2 & \beta^{2(h+2)}v^2 \end{pmatrix} \\
=&\frac{1}{W(v,h)} \Bigg[\begin{pmatrix} \beta^{2(h+1)}v^2 + \beta^{2h}v + \cdots + \beta^2 v + v & \beta^{2(h+1)+1}v^2 + \beta^{2h+1}v + \cdots + \beta^3 v + \beta v \\ \beta^{2(h+1)+1}v^2 + \beta^{2h+1}v + \cdots + \beta^3 v + \beta v & (\beta^{2(h+1)+2}v^2 + \beta^{2h+2}v + \cdots + \beta^4 v + \beta^2 v \\ & \quad + \beta^{2(h+1)}v + \beta^{2h} + \cdots + \beta^2 + 1) \end{pmatrix} \\
&- \begin{pmatrix} \beta^{2(h+1)}v^2 & \beta^{2h+3}v^2 \\ \beta^{2h+3}v^2 & \beta^{2(h+2)}v^2 \end{pmatrix} \Bigg] \\
=&\frac{1}{W(v,h)} \begin{pmatrix} (\beta^{2h} + \beta^{2(h-1)} + \cdots + \beta^2 + 1)v & (\beta^{2h} + \beta^{2(h-1)} \cdots + \beta^2 + 1)\beta v \\ (\beta^{2h} + \beta^{2(h-1)} \cdots + \beta^2 + 1)\beta v & (\beta^{2h} + \beta^{2(h-1)} + \cdots + \beta^2 + 1)(\beta^2 v + 1) + \beta^{2h+2}v \end{pmatrix}
\end{aligned}$$

Thus the conditional correlation is

$$\begin{aligned}
\rho(X_j, X_k \mid X_\ell) &= \frac{\beta v \times \frac{1-\beta^{2(h+1)}}{1-\beta^2}}{\sqrt{v \times \frac{1-\beta^{2(h+1)}}{1-\beta^2} \times \left[(1+\beta^2 v) \times \frac{1-\beta^{2(h+1)}}{1-\beta^2} + \beta^{2h} \times \beta^2 v\right]}} \\
&= \sqrt{\frac{\beta^2 v}{1 + \left(1 + \frac{\beta^{2h}}{g(h)}\right)\beta^2 v}}
\end{aligned}$$

where $g(h) = \frac{1-\beta^{2(h+1)}}{1-\beta^2} \geq 1$ for $h \geq 0$. Since $\beta^2 = 2c^2 \leq 1$, then $1 + \beta^{2h}/g(h) \leq 2$. Since $\rho(X_j, X_k \mid X_\ell) \geq c \Leftrightarrow \beta^2 v \geq \frac{c^2}{1 - \left(1 + \frac{\beta^{2h}}{g(h)}\right)c^2}$, and $v = \mathbb{E}[X_j^2] \geq 1$ for all $j \in [d]$,

then $\beta^2 v \geq \beta^2 = 2c^2 \geq \frac{c^2}{1-2c^2} \geq \frac{c^2}{1-\left(1+\frac{\beta^2 h}{g(h)}\right)c^2}$ when $c^2 \leq 1/5$, which completes the proof.

□

APPENDIX B

PROOFS IN CHAPTER 3

B.1 Details of Applications

In this appendix, we detail the formal definitions of the model class considered in each application in Section 3.3. Specifically, we specify $\mathcal{P}$ and m, thereby $\mathcal{H}_1$ and $\mathcal{H}_0$ will follow, for the exampled distributions.

B.1.1 Bernoulli Distribution (Section 3.3.1)

We start by specifying the model class $\mathcal{P}$. Consider all multivariate Bernoulli distributions with dimension being three. They are parametrized by joint probability mass function $p(x, y, z)$:

$$\mathcal{P}^{Ber} = \Big\{ p(x,y,z) : \\ (X,Y,Z) \sim \mathbb{P}(X = x, Y = y, Z = z) = p(x,y,z), \\ (x,y,z) \in \{0,1\}^3, \sum_{x,y,z} p(x,y,z) = 1 \Big\}.$$

We measure dependence using the total variation distance to the product of its marginals:

$$m^{Ber}(X, Y \mid Z) = \sum_z p(z) \sum_{x,y} |p(x,y \mid z) - p(x \mid z)p(y \mid z)| .$$

Then $\mathcal{H}_0^{Ber}$ and $\mathcal{H}_1^{Ber}$ contain all these Bernoulli distributions that are conditionally independent and dependent respectively:

$$\mathcal{H}_0^{Ber} = \{p \in \mathcal{P}^{Ber} : X \perp\!\!\!\perp Y \mid Z\}, \quad \mathcal{H}_1^{Ber} = \{p \in \mathcal{P}^{Ber} : m^{Ber}(X, Y \mid Z) \geq c\}$$

Having defined the elements of CI testing problem $\mathcal{C}^{Ber} := \mathcal{C}(\mathcal{P}^{Ber}, m^{Ber})$, the Bernoulli poly-forest learning problem $\mathcal{F}^{Ber} := \mathcal{F}(\mathcal{P}^{Ber}, m^{Ber}, c)$ contains all multivariate Bernoulli distributions with dimension d and Markov and c-strong tree-faithful to some poly-forest $G \in \widetilde{\mathcal{T}}$:

$$(X_1, X_2 \ldots, X_d) \sim \mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_d = x_d), \quad (x_1, x_2 \ldots, x_d) \in \{0,1\}^d .$$

B.1.2 Gaussian Distribution (Section 3.3.2)

Consider all multivariate Gaussian distributions with dimension being three:

$$\mathcal{P}^{Gauss} = \left\{ p(X, Y, Z) : (X, Y, Z) \sim \mathcal{N}(\mathbf{0}_3, \Sigma), \Sigma \in \mathbb{S}^3_{++} \right\}.$$

We measure dependence using the partial correlation:

$$m^{Gauss}(X, Y \mid Z) = \frac{|\operatorname{cov}(X, Y \mid Z)|}{\sqrt{\operatorname{var}(X \mid Z) \operatorname{var}(Y \mid Z)}} . \tag{B.1}$$

Then $\mathcal{H}_0^{Gauss}$ and $\mathcal{H}_1^{Gauss}$ contain all these Gaussian distributions that are conditionally independent and dependent respectively.

$$\mathcal{H}_0^{Gauss} = \{ p \in \mathcal{P}^{Gauss} : X \perp\!\!\!\perp Y \mid Z \}, \quad \mathcal{H}_1^{Gauss} = \{ p \in \mathcal{P}^{Gauss} : m^{Gauss}(X, Y \mid Z) \geq c \}$$

Consequently, the Gaussian CI testing problem $\mathcal{C}^{Gauss} := \mathcal{C}(\mathcal{P}^{Gauss}, m^{Gauss})$ is defined. Meanwhile, it gives the Gaussian poly-forest learning problem $\mathcal{F}^{Gauss} := \mathcal{F}(\mathcal{P}^{Gauss}, m^{Gauss}, c)$, which include all multivariate Gaussian distributions with dimension d and Markov and c-strong tree-faithful to some poly-forest $G \in \widetilde{\mathcal{T}}$:

$$(X_1, X_2 \ldots, X_d) \sim \mathcal{N}(\mathbf{0}_d, \Sigma), \Sigma \in \mathbb{S}^d_{++} .$$

Similar to ψ^{Ber}, we construct a test by thresholding the sample partial correlation:

$$\widehat{m}^{Gauss}(X, Y \mid Z) = \frac{|\widehat{\Sigma}_{XY} - \widehat{\Sigma}_{XZ}\widehat{\Sigma}_{ZZ}^{-1}\widehat{\Sigma}_{ZY}|}{\sqrt{(\widehat{\Sigma}_{XX} - \widehat{\Sigma}_{XZ}\widehat{\Sigma}_{ZZ}^{-1}\widehat{\Sigma}_{ZX})(\widehat{\Sigma}_{YY} - \widehat{\Sigma}_{YZ}\widehat{\Sigma}_{ZZ}^{-1}\widehat{\Sigma}_{ZY})}} \tag{B.2}$$
$$\psi^{Gauss} = \mathbb{1}\{\widehat{m}^{Gauss}(X, y \mid Z) \geq c/2\},$$

where $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X^{(i)} X^{(i)\top}$ is the sample covariance matrix.

B.1.3 Nonparametric Continuous Distribution (Section 3.3.4)

We follow the CI testing framework in Neykov et al. [2021]. Consider all continuous distributions over $[0,1]^3$ that admit continuous densities $p(X, Y, Z)$. In addition, we impose following two smoothness condition on the densities.

Definition 12 (Lipschitzness)**.** For some constant $L_1 > 0$,

- if $X \perp\!\!\!\perp Y \mid Z$, then $\|p_{X \mid Z=z} - p_{X \mid Z=z'}\|_1 \leq L_1|z - z'|$ and $\|p_{Y \mid Z=z} - p_{Y \mid Z=z'}\|_1 \leq L_1|z - z'|$.
- if $X \not\perp\!\!\!\perp Y \mid Z$, then $\|p_{X,Y \mid Z=z} - p_{X,Y \mid Z=z'}\|_1 \leq L_1|z - z'|$.

Definition 13 (s-Hölder smoothness)**.** For some $s > 0$, let $\lfloor s \rfloor$ be the maximum integer smaller than s. For some constant $L_2 > 0$ and all $x, x', y, y', z \in [0,1]$,

$$\sup_{t \leq \lfloor s \rfloor} \left| \frac{\partial^t}{\partial x^t} \frac{\partial^{\lfloor s \rfloor - t}}{\partial y^{\lfloor s \rfloor - t}} p(x, y \mid z) - \frac{\partial^t}{\partial x^t} \frac{\partial^{\lfloor s \rfloor - t}}{\partial y^{\lfloor s \rfloor - t}} p(x', y' \mid z) \right| \leq L_2 \sqrt{(x - x')^2 + (y - y')^2}^{s - \lfloor s \rfloor}$$

and

$$\sup_{t \leq \lfloor s \rfloor} \left| \frac{\partial^t}{\partial x^t} \frac{\partial^{\lfloor s \rfloor - t}}{\partial y^{\lfloor s \rfloor - t}} p(x, y \mid z) \right| \leq L_2 .$$

The Lipschitzness is used to characterize the smoothness with respect to the conditioning variable Z, while the Hölder smoothness is required for the conditional densities with the conditioning variable fixed, and particularly for the (conditionally) dependent variable pairs. Denote this class of distributions as $\mathcal{P}^{NP}$. As specified in the main paper, we measure the dependence using

$$m^{NP}(X, Y \mid Z) = \inf_{q \in \mathcal{Q}} \|p - q\|_1$$

which is the distance to the closest (conditionally) independent distributions, and should be large enough such that the dependent distributions can be distinguished from the independent ones [Shah and Peters, 2020]. Therefore, the CI testing class is defined $\mathcal{C}^{NP} := \mathcal{C}(\mathcal{P}^{NP}, m^{NP})$, with the null and alternative hypotheses being

$$\begin{aligned}
\mathcal{H}_0^{NP} &= \{p \in \mathcal{P}^{NP} : X \perp\!\!\!\perp Y \mid Z \text{ and satisfies Lipschitzness}\} \\
\mathcal{H}_1^{NP} &= \{p \in \mathcal{P}^{NP} : m^{NP}(X, Y \mid Z) \geq c \\
&\qquad\qquad \text{and satisfies Lipschitzness and } s\text{-Hölder smoothness}\}\,.
\end{aligned}$$

Correspondingly, the nonparametric poly-forest learning problem is defined, which includes all continuous distributions supported on $[0,1]^d$ and Markov and c-strong tree-faithful to some poly-forest $G \in \widetilde{\mathcal{T}}$, denoted as $\mathcal{F}^{NP} := \mathcal{F}(\mathcal{P}^{NP}, m^{NP}, c)$.

B.2 Proof of Theorem 3.2.1

We provide the full version of Theorem 3.2.1 below, followed by the discussion and proof.

Theorem B.2.1. *Given a conditional independence testing problem $\mathcal{C}(\mathcal{P}, m)$, if for some constants A and A' independent with sample size n,*

(C1) There exists a test ψ such that when $c \geq A \times n^{-1/\alpha}$, it holds that

$$\sup_{p \in \mathcal{H}_0} \mathbb{E}_p[\psi] + \sup_{p \in \mathcal{H}_1} \mathbb{E}_p[1 - \psi] \leq \frac{1}{2};$$

(C2) There exist $p_0 \in \mathcal{H}_0$ and $p_1 \in \mathcal{H}_1$ such that $\mathbf{KL}(p_1 \| p_0) \leq A' \times c^\alpha$;

then the minimax testing radius of $\mathcal{C}(\mathcal{P}, m)$ is $c_n \asymp n^{-1/\alpha}$, and ψ is optimal for $\mathcal{C}(\mathcal{P}, m)$. In addition, if

(C3) p_0 and p_1 are Markov and c-strong tree-faithful to some poly-forests of three nodes;

then the optimal sample complexity of learning $\mathcal{F}(\mathcal{P}, m, c)$ is

$$n \asymp \frac{\log d}{c^\alpha},$$

which is achieved by `PC-tree` *with ψ.*

The first two conditions assumed in Theorem B.2.1 are standard steps to study minimax optimality, corresponding to upper and lower bounds of testing radius. It is typical in literature to verify these two conditions when deriving the optimal testing radius:

- (C1) requires finding a powerful test to distinguish independent and dependent instances with sufficiently large signal.
- (C2) looks for two instances from $\mathcal{H}_0$ and $\mathcal{H}_1$ that are hard to distinguished. Typically, p_0 is a "flat" distribution, e.g. uniform distribution or $\text{Bern}(0.5)$, while p_1 is constructed to be a perturbation to the flat distribution with sufficiently large signal c but small deviation to being independent.

- With above two conditions, to conclude optimal poly-forest learning, (C3) requires the instances to be generated by poly-forest models, ensuring their validity in the context of structure learning.

Proof. Under the given conditions, we prove the optimalities for CI testing and poly-forest learning.

Optimal CI testing

The existence of test ψ implies a upper bound of testing radius $c_n \lesssim n^{-1/\alpha}$. The lower bound is given by Le Cam's two point method: by the existence $p_0 \in \mathcal{H}_0, p_1 \in \mathcal{H}_1$, we have

$$\inf_{\psi} \left\{ \sup_{p \in \mathcal{H}_0} \mathbb{E}_p[\psi] + \sup_{p \in \mathcal{H}_1} \mathbb{E}_p[1-\psi] \right\} \geq \frac{1}{2}\Big(1 - n\mathbf{TV}(p_1 \| p_0)\Big),$$

and $\mathbf{TV}(p_1 \| p_0) \leq \mathbf{KL}(p_1 \| p_0) \leq A' \times c^{\alpha}$. This requires the testing radius $nc_n^{\alpha} \gtrsim 1$, which yields $c_n \gtrsim n^{-1/\alpha}$.

Optimal poly-forest learning

Regarding the upper bound, we start with applying median trick [Motwani, 1995] to obtain an exponential decay of error probability. The idea is to divide the full sample into K folds and apply the statistical test on each sub-sample. Take the majority vote of all these tests as the final output. Then the final output makes mistake only when half of the tests make mistake, whose probability can easily computed and bounded in an exponential way and serves our goal when K is appropriately chosen.

Proposition B.2.2. *Suppose there exists a statistical test ψ such that if the sample size $n \gtrsim N$, then*

$$\mathbb{P}(\psi \textit{ is incorrect}) \leq 1/2,$$

then there exists a CI test ψ' such that under the same condition,

$$\mathbb{P}(\psi' \textit{ is incorrect}) \leq \exp\Big(-C_0 n/N\Big),$$

for some constant C_0 independent with n.

Applying Proposition B.2.2, we can derive an error probability bound for ψ with modification. For simplicity, we will stick with ψ with a little abuse of notation. Assuming $n \gtrsim c^{-\alpha}$, we have for some constant C_0,

$$\mathbb{P}(\psi \text{ is incorrect}) \leq \exp\Big(-C_0 n c^{\alpha}\Big).$$

With this error bound in hand, the proof proceeds by following the proof of Theorem 4.3 in Wang et al. [2024] with minor modifications: 1) replacing poly-tree with poly-forest; 2) replacing correlation CI testing with ψ. This completes the proof of upper bound, and shows `PC-tree` with ψ as CI tester achieves the upper bound.

Regarding the lower bound, without loss of generality, we assume d divides 3, otherwise proceed by setting the remaining one or two nodes to be isolated and follow $p_0(X)$. Group all the nodes into $d' = d/3$ clusters:

$$\{1,2,3\}, \{4,5,6\}, \cdots, \{3k+1, 3k+2, 3k+3\}, \cdots, \{3d'+1, 3d'+2, 3d'+3\}.$$

We are going to use the p_1, p_0 to parametrize the construction. Since p_1 and p_0 give different conditional independence statements, they are Markov and c-strong tree-faithful to two different poly-forests, which are denoted as T_0, T_1.

Firstly, construct graph G_0 by stacking d' many T_0's together. Then for each $k \geq 1$, construct graph G_k by stacking $d'-1$ many T_0's and one T_1 together, and the subgraph T_1 is imposed on nodes $\{3k+1, 3k+2, 3k+3\}$, while T_0 is imposed on all remaining triplets.

Under this construction, $\{G_k\}_{k\geq 0}$ are poly-forests with (at least) d' many disconnected subgraphs, and are distinct with each other.

Now we consider distributions for each $k \geq 0$. Let $P_0 = p_0^{\otimes d'}$ and

$$P_k(X) = p_0^{\otimes d'-1} \times P_k(X_{3k+1}, X_{3k+2}, X_{3k+3}) \quad \text{with} \quad P_k(X_{3k+1}, X_{3k+2}, X_{3k+3}) = p_1 \,.$$

Since p_0 and p_1 are Markov and c-strong tree-faithful to T_0 and T_1 respectively, P_k is Markov and c-strong tree-faithful to G_k for all $k \geq 0$. Therefore, $\{P_k\}_{k\geq 0} \in \mathcal{F}(\mathcal{P}, m, c)$.

We also going to apply Tsybakov's method (Corollary D.2.4). Firstly, we have the size of the construction to be lower bounded

$$\log M = \log(d'+1) \geq \frac{1}{2}\log d \,.$$

Then we upper bound the KL divergence between P_k and P_0 for all $k \geq 1$:

$$\mathbf{KL}(P_k \| P_0) = \mathbb{E}_{P_k} \log \frac{P_k}{P_0} = \mathbb{E}_{P_k} \log \frac{p_1}{p_0} = \mathbf{KL}(p_1 \| p_0) \leq A' \times c^{\alpha} \,.$$

Invoking Corollary D.2.4 completes the proof. □

Proof of Proposition B.2.2. Divide the full sample into K folds, which we will specify later, and apply ψ_0 for each of the sub-sample to get outputs $\psi^{(1)}, \ldots, \psi^{(K)}$. Let

$$\psi = \operatorname*{arg\,max}_{t\in\{0,1\}} \sum_{k=1}^{K} \mathbb{1}\{\psi^{(k)} = t\}$$

be the majority vote of all the tests as final output. Therefore, the incorrectness of ψ implies at least half of the sub-tests make mistakes. Suppose that $n/K \gtrsim N$, by the guarantee

of ψ_0, we have

$$\begin{aligned}\mathbb{P}(\psi \text{ is incorrect}) &\leq \mathbb{P}\Big(\text{half of } \{\psi^{(k)}\}_{k=1}^{K} \text{ is incorrect}\Big) \\ &\leq \frac{1}{2^{K/2}} = \exp\Big(-\frac{\log 2}{2} \times K\Big).\end{aligned}$$

Now set $K = C'n/N$ for constant C' large enough such that $n/K \gtrsim N$ is satisfied, let $C_0 = C'(\log 2)/2$, we obtain

$$\mathbb{P}(\psi \text{ is incorrect}) \leq \exp\Big(-C_0 n/N\Big)$$

as desired. □

B.3 Proof of Theorem 3.3.1 (Bernoulli Distribution)

We start by showing the upper bound (C1) of ψ^{Ber}, then proceed to verify the validity of p_0^{Ber} and p_1^{Ber} and show they satisfy (C2) and (C3).

B.3.1 Proof of Upper Bound for Bernoulli Distribution

Proof. We directly show ψ^{Ber} satisfies an exponential bound. Notice that by construction, the concentration of $\widehat{m}$ implies the correctness of testing:

$$|\widehat{m}^{Ber} - m^{Ber}| \leq c/2 \implies \psi^{Ber} \text{ is correct.}$$

We aim to show $|\widehat{m}^{Ber} - m^{Ber}| \leq c/2$ holds with high probability. To achieve this, we employee the result below:

Lemma B.3.1 (Devroye [1983], Lemma 3)**.** *Let $(X_1, X_2, \ldots, X_k)$ be a multinomial $(n, p_1, p_2, \ldots, p_k)$ random vector. Let $\widehat{p} = (X_1, X_2, \ldots, X_k)/n$. For all $\epsilon \in (0, 1)$*

and all k satisfying $k/n \le \epsilon^2/20$, *we have*

$$\mathbb{P}(\|\widehat{p} - p\|_1 > \epsilon) \le 3\exp(-n\epsilon^2/25)\,.$$

Applied to our setup, we consider the concentration for the joint distribution p_{XYZ}, which can be viewed as a multinomial distribution with dimension $k = 2^3 = 8$. Therefore, for ϵ such that $n \ge 160/\epsilon^2$, with probability at least $1 - 3\exp(-n\epsilon^2/25)$, we have

$$\|p_{XYZ} - \widehat{p}_{XYZ}\|_1 := \sum_{xyz} |p(x,y,z) - \widehat{p}(x,y,z)| \le \epsilon\,,$$

which also implies the concentration of p_{WZ} ($W \in \{X, Y\}$) and p_Z. To see this, take $W = X$ for example:

$$\begin{aligned}\|\widehat{p}_{XZ} - p_{XZ}\|_1 &= \sum_{xz} |\widehat{p}(x,z) - p(x,z)| = \sum_{xz} |\sum_{y} \Big(\widehat{p}(x,y,z) - p(x,y,z)\Big)| \\ &\le \sum_{xz}\sum_{y} |\widehat{p}(x,w,z) - p(x,w,z)| = \|p_{XYZ} - \widehat{p}_{XYZ}\| \le \epsilon\,.\end{aligned}$$

Similarly,

$$\begin{aligned}\|\widehat{p}_{Z} - p_{Z}\|_1 &= \sum_{z} |\widehat{p}(z) - p(z)| = \sum_{z} |\sum_{xy} \Big(\widehat{p}(x,y,z) - p(x,y,z)\Big)| \\ &\le \sum_{z}\sum_{xy} |\widehat{p}(x,w,z) - p(x,w,z)| = \|p_{XYZ} - \widehat{p}_{XYZ}\| \le \epsilon\,.\end{aligned}$$

Since the estimator $\widehat{m}^{Ber}$ involves the estimation of $p_{XY\,|\,Z}$ and $p_{W\,|\,Z}$ for $W = X$ or Y, we

proceed to bound the error of them. For any $(x, y, z) \in \{0,1\}^3$,

$$\begin{aligned}
|\widehat{p}(x, y \,|\, z) - p(x, y \,|\, z)| &= \Big|\frac{\widehat{p}(x,y,z)}{\widehat{p}(z)} - \frac{p(x,y,z)}{p(z)}\Big| \\
&= \Big|\frac{\widehat{p}(x,y,z)}{\widehat{p}(z)} - \frac{\widehat{p}(x,y,z)}{p(z)} + \frac{\widehat{p}(x,y,z)}{p(z)} - \frac{p(x,y,z)}{p(z)}\Big| \\
&\le \widehat{p}(x,y,z)\Big|\frac{1}{\widehat{p}(z)} - \frac{1}{p(z)}\Big| + \frac{1}{p(z)}|\widehat{p}(x,y,z) - p(x,y,z)| \\
&\le \widehat{p}(x,y,z)\frac{|\widehat{p}(z) - p(z)|}{p(z)\widehat{p}(z)} + \frac{1}{p(z)}|\widehat{p}(x,y,z) - p(x,y,z)| \\
&= \frac{1}{p(z)}\Big(\widehat{p}(x,y \,|\, z)|\widehat{p}(z) - p(z)| + |\widehat{p}(x,y,z) - p(x,y,z)|\Big) \\
&\le \frac{1}{p(z)}\Big(|\widehat{p}(z) - p(z)| + |\widehat{p}(x,y,z) - p(x,y,z)|\Big).
\end{aligned}$$

Analogously, for $W = X$ or Y and any $(w, z) \in \{0,1\}^2$

$$\begin{aligned}
|\widehat{p}(w \,|\, z) - p(w \,|\, z)| &= \Big|\frac{\widehat{p}(w,z)}{\widehat{p}(z)} - \frac{p(w,z)}{p(z)}\Big| \\
&= \Big|\frac{\widehat{p}(w,z)}{\widehat{p}(z)} - \frac{\widehat{p}(w,z)}{p(z)} + \frac{\widehat{p}(w,z)}{p(z)} - \frac{p(w,z)}{p(z)}\Big| \\
&\le \widehat{p}(w,z)\Big|\frac{1}{\widehat{p}(z)} - \frac{1}{p(z)}\Big| + \frac{1}{p(z)}|\widehat{p}(w,z) - p(w,z)| \\
&\le \widehat{p}(w,z)\frac{|\widehat{p}(z) - p(z)|}{p(z)\widehat{p}(z)} + \frac{1}{p(z)}|\widehat{p}(w,z) - p(w,z)| \\
&= \frac{1}{p(z)}\Big(\widehat{p}(w \,|\, z)|\widehat{p}(z) - p(z)| + |\widehat{p}(w,z) - p(w,z)|\Big) \\
&\le \frac{1}{p(z)}\Big(|\widehat{p}(z) - p(z)| + |\widehat{p}(w,z) - p(w,z)|\Big).
\end{aligned}$$

Denote $\widehat{p}(x,y,z) = p(x,y,z) + \delta_{xyz}$, $\widehat{p}(w,z) = p(w,z) + \delta_{wz}$, $\widehat{p}(z) = p(z) + \delta_z$, thus $\delta_{xyz}, \delta_{wz}$ are bounded correspondingly. Then we are ready to show the concentration of

$\widehat{m}^{Ber}$ below.

$$
\begin{aligned}
&|\widehat{m}^{Ber} - m^{Ber}| \\
&= \sum_z \left\{ \widehat{p}(z) \sum_{xy} |\widehat{p}(x,y \mid z) - \widehat{p}(x \mid z)\widehat{p}(y \mid z)| - p(z) \sum_{xy} |p(x,y \mid z) - p(x \mid z)p(y \mid z)| \right\},
\end{aligned}
$$

where

$$
\begin{aligned}
&\widehat{p}(z) \sum_{xy} |\widehat{p}(x,y \mid z) - \widehat{p}(x \mid z)\widehat{p}(y \mid z)| \\
=&\Big(p(z) + \delta_z\Big) \sum_{xy} \Big| \Big(p(x,y \mid z) + \delta_{xyz}\Big) - \Big(p(x \mid z) + \delta_{xz}\Big)\Big(p(y \mid z) + \delta_{yz}\Big) \Big| \\
\leq& p(z) \sum_{xy} |p(x,y \mid z) - p(x \mid z)p(y \mid z)| \\
&+ p(z) \sum_{xy} \Big\{ |\delta_{xyz}| + |\delta_{xz}\delta_{yz} + \delta_{xz}p(y \mid z) + \delta_{yz}p(x \mid z)| \Big\} \\
&+ \delta_z \sum_{xy} |\widehat{p}(x,y \mid z) - \widehat{p}(x \mid z)\widehat{p}(y \mid z)| \\
\leq& p(z) \sum_{xy} |p(x,y \mid z) - p(x \mid z)p(y \mid z)| + p(z) \sum_{xy} \Big\{ |\delta_{xyz}| + |\delta_{xz}\widehat{p}(y \mid z) + \delta_{yz}p(x \mid z)| \Big\} + 8\delta_z \\
\leq& p(z) \sum_{xy} |p(x,y \mid z) - p(x \mid z)p(y \mid z)| + p(z) \sum_{xy} \Big\{ |\delta_{xyz}| + |\delta_{xz}| + |\delta_{yz}| \Big\} + 8\delta_z \,.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&|\widehat{m}^{Ber} - m^{Ber}| \\
&\leq \sum_z \left(p(z) \sum_{xy} \left\{ |\delta_{xyz}| + |\delta_{xz}| + |\delta_{yz}| \right\} + 8\delta_z \right) \\
&\leq 4\|p_Z - \widehat{p}_Z\|_1 + \|p_{XYZ} - \widehat{p}_{XYZ}\|_1 \\
&\quad 4\|p_Z - \widehat{p}_Z\|_1 + 2\|p_{XZ} - \widehat{p}_{XZ}\|_1 \\
&\quad 4\|p_Z - \widehat{p}_Z\|_1 + 2\|p_{YZ} - \widehat{p}_{YZ}\|_1 \\
&\quad 8\|p_Z - \widehat{p}_Z\|_1 \\
&= 20\|p_Z - \widehat{p}_Z\|_1 + \|p_{XYZ} - \widehat{p}_{XYZ}\|_1 + 2\|p_{XZ} - \widehat{p}_{XZ}\|_1 + +2\|p_{YZ} - \widehat{p}_{YZ}\|_1 \\
&\leq 25\epsilon \,.
\end{aligned}
$$

Choosing $\epsilon = c/50$, we have with probability at least

$$1 - 3\exp(-C_0 n c^2)\,,$$

$|\widehat{m}^{Ber} - m^{Ber}| \leq c/2$ and ψ^{Ber} is correct, where the constant $C_0 = 50^2 \times 25$. □

B.3.2 Proof of Lower Bound for Bernoulli Distribution

Proof. We proceed to verify the validity of p_0^{Ber} and p_1^{Ber} by showing they satisfy (C2) and (C3). We can compute the KL divergence between them:

$$\mathbf{KL}(p_1^{Ber} \| p_0^{Ber}) = \log(1 - 4c^2) + 2c \log(1 + \frac{4c}{1 - 2c}) \leq 100c^2\,,$$

for c small enough. Then (C2) holds. Moreover, it suffices to show for p_1^{Ber} is c-strong tree-faithfulness to the chain graph $Z \to X \to Y$. To achieve this, we can compute

$$
\begin{aligned}
&m^{Ber}(X, Y) = 2c, m^{Ber}(X, Z) = 2c, \\
&m^{Ber}(X, Y \mid Z) \geq 2c(1 - 4c^2) \geq c, \\
&m^{Ber}(Z, X \mid Y) \geq 2c(1 - 4c^2) \geq c\,,
\end{aligned}
$$

for c small enough. Then (C3) holds. □

B.4 Proof of Theorem 3.3.2 (Gaussian distribution)

Proof. The upper bound (C1) of ψ^{Gauss} is given by Lemma C.1 in Wang et al. [2024]. We proceed to verify the validity of p_0^{Gauss} and p_1^{Gauss} by showing they satisfy (C2) and (C3). We can compute the KL divergence between them:

$$
\mathbf{KL}(p_1^{Gauss} \| p_0^{Gauss}) = -\log(1 - 4c^2) \leq 100c^2\,,
$$

for c small enough. Then (C2) holds. Moreover, it suffices to show for p_1^{Gauss} is c-strong tree-faithfulness to the chain graph $Z \to X \to Y$. To achieve this, we can compute

$$
\begin{aligned}
&m^{Gauss}(X, Y) = 2c, m^{Gauss}(X, Z) = 2c, \\
&m^{Gauss}(X, Y \mid Z) \geq 2c(1 - 4c^2) \geq c, \\
&m^{Gauss}(Z, X \mid Y) \geq 2c(1 - 4c^2) \geq c\,,
\end{aligned}
$$

for c small enough. Then (C3) holds. □

B.5 Proof of Theorem 3.3.3 (Poly-tree Learning)

Proof. The optimality of Gaussian poly-tree learning is established in Wang et al. [2024]. For Bernoulli poly-tree learning, the upper bound using ψ^B follows the proof of Theorem 4.3 in Wang et al. [2024] combined with results in Appendix B.3.1. It suffices to prove the validity of the lower bound construction (3.4), which is given by the following two lemmas:

Lemma B.5.1. *Assuming $c \leq 1/4$, for any two Markov chains T_1, T_2, $\mathbf{KL}(P_{T_1} \| P_{T_2}) \leq 16dc^2$.*

Lemma B.5.2. *Assuming $c \leq 1/4$, for any Markov chain T, P_T satisfies c-strong tree-faithfulness with respect to m^B.*

Since the size of all directed Markov chains is

$$\log d! \geq \frac{1}{2} d \log d \,,$$

Corollary D.2.3 implies the lower bound of $\log d / c^2$. □

We proceed to prove Lemma B.5.1 and B.5.2. We start with three facts about the construction: the first states the marginal of any variable is a "coin flip"; the second states the positions of conditional probability can be switched; the third states the conditional probability of each variable is close to "coin flip".

Lemma B.5.3. *For P_T, we have the following:*

1. *For any $k \in [d]$, $\mathbb{P}(X_k = x_k) = 1/2$ for $x_k = 1$ or 0.*
2. *For any $j \neq k \in [d]$, $\mathbb{P}(X_k = x_k \,|\, X_j = x_j) = \mathbb{P}(X_j = x_j \,|\, X_k = x_k)$.*
3. *For any $j \neq k \in [d]$, $\mathbb{P}(X_k = x_k \,|\, X_j = x_j) = 1/2 + \delta$ with $|\delta| \leq c$, for $x_k, x_j \in \{0, 1\}$.*

Proof. Without loss of generality, fix T to be the natural ordering: $1 \to 2 \to \cdots \to d$. For the first fact, since X_1 is a Bernoulli random variable, for X_2,

$$\mathbb{P}(X_2 = x_2) = \sum_{x_1} \mathbb{P}(X_2 = x_2 \,|\, X_1 = x_1)\mathbb{P}(X_1 = x_1) = \frac{1}{2} \times (\frac{1}{2} + c) + \frac{1}{2} \times (\frac{1}{2} - c) = \frac{1}{2}\,.$$

So marginally X_2 is a Bernoulli random variable. By induction, $X_3, X_4, \ldots, X_d$ are all marginally a Bernoulli random variable.

For the second fact, because of the first fact,

$$\mathbb{P}(X_k \,|\, X_j) = \frac{\mathbb{P}(X_j \,|\, X_k)\mathbb{P}(X_k)}{\mathbb{P}(X_j)} = \mathbb{P}(X_j \,|\, X_k)\,.$$

For the last fact, due to the second fact, we look at the case where $j < k$, otherwise we can switch the positions. By Markov property,

$$\begin{aligned}\mathbb{P}(X_k = x_k \,|\, X_j = x_j) = \mathbb{P}(X_k = x_k \,|\, X_{k-1} = 1)\mathbb{P}(X_{k-1} = 1 \,|\, X_j = x_j) \\ + \mathbb{P}(X_k = x_k \,|\, X_{k-1} = 0)\mathbb{P}(X_{k-1} = 0 \,|\, X_j = x_j)\,,\end{aligned}$$

which is a convex combination of $\frac{1}{2} \pm c$, then $\mathbb{P}(X_k = x_k \,|\, X_j = x_j) = \frac{1}{2} + \delta$ for some $|\delta| \le c$. □

Proof of Lemma B.5.1. Without loss of generality, let T_1 be the natural ordering: $1 \to 2 \to \cdots \to d$, and T_2 is some other ordering π: $\pi(1) \to \pi(2) \to \cdots \to \pi(d)$. Then we have the KL divergence

$$\begin{aligned}&\mathbf{KL}(P_{T_1} \| P_{T_2}) = \mathbb{E}_{P_{T_1}} \log \frac{P_{T_1}}{P_{T_2}} \\ &= \sum_{k=1}^{d} \left(\mathbb{E}_{T_1} \log 2P_{T_1}(X_k \,|\, X_{k-1}) - \mathbb{E}_{T_1} \log 2P_{T_2}(X_{\pi(k)} \,|\, X_{\pi(k-1)}) \right).\end{aligned}$$

Inside the summation, for the first term,

$$
\begin{aligned}
&\mathbb{E}_{T_1} \log 2P_{T_1}(X_k \mid X_{k-1}) \\
=&\sum_{x_{k-1}} \mathbb{P}(X_{k-1} = x_{k-1}) \sum_{x_k} \mathbb{P}(X_k = x_k \mid X_{k-1} = x_{k-1}) \log 2\mathbb{P}(X_k = x_k \mid X_{k-1} = x_{k-1}) \\
=&2 \times \frac{1}{2} \times \left((\frac{1}{2} + c) \log(1 + 2c) + (\frac{1}{2} - c) \log(1 - 2c) \right) \\
=&\frac{1}{2} \log(1 - 4c^2) + c \log(1 + \frac{4c}{1 - 2c}) .
\end{aligned}
$$

For the second term, for simplicity, let $(\pi(k), \pi(k-1)) = (\ell, j)$. By Lemma B.5.3, suppose $\mathbb{P}(X_\ell = x_\ell \mid X_j = x_j) = 1/2 \pm \delta$. Then

$$
\begin{aligned}
&\mathbb{E}_{T_1} \log 2P_{T_2}(X_{\pi(k)} \mid X_{\pi(k-1)}) \\
=&\sum_{x_j} \mathbb{P}(X_j = x_j) \sum_{x_\ell} \mathbb{P}(X_\ell = x_\ell \mid X_j = x_j) \log 2\mathbb{P}_{T_2}(X_\ell = x_\ell \mid X_j = x_j) \\
=&2 \times \frac{1}{2} \times \left((\frac{1}{2} + \delta) \log(1 + 2c) + (\frac{1}{2} - \delta) \log(1 - 2c) \right) \\
=&\frac{1}{2} \log(1 - 4c^2) + \delta \log(1 + \frac{4c}{1 - 2c}) .
\end{aligned}
$$

Then the difference between the two terms is

$$
(c - \delta) \log(1 + \frac{4c}{1 - 2c}) \leq (c - \delta)c \times \frac{4}{1 - 2c} \leq 16c^2 ,
$$

when $c \leq 1/4$. Therefore, $\mathbf{KL}(P_{T_1} \| P_{T_2}) \leq 16dc^2$. □

Proof of Lemma B.5.2. Without loss of generality, let T be the natural ordering: $1 \to 2 \to \cdots \to d$. Since there is no V-structure in Markov chain, we only need to show the first requirement in Definition 6 holds, i.e. for any k and $j \neq k, k+1$, $m^B(X_k; X_{k+1} \mid X_j) \geq c$. There are two cases to look at: (1) $k > j$; (2) $j > k + 1$.

For the first case, we omit the capital letter when writing the probability.

$$\begin{aligned} m^B(X_k; X_{k+1} \mid X_j) &= \sum_{x_k, x_{k+1}, x_j} P(x_j) \Big| P(x_{k+1}, x_k \mid x_j) - P(x_{k+1} \mid x_j) P(x_k \mid x_j) \Big| \\ &= \sum_{x_j, x_k} P(x_j) P(x_k \mid x_j) \sum_{x_{k+1}} \Big| P(x_{k+1} \mid x_j, x_k) - P(x_{k+1} \mid x_j) \Big| \\ &= \sum_{x_j, x_k} P(x_j) P(x_k \mid x_j) \sum_{x_{k+1}} \Big| P(x_{k+1} \mid x_k) - \sum_{x'_k} P(x_{k+1} \mid x'_k) P(x'_k \mid x_j) \Big| . \end{aligned}$$

By Lemma B.5.3, suppose $P(x_k \mid x_j) = 1/2 \pm \delta$. The calculation for the equation above is summarized below:

(x_j, x_k)	$P(x_j)P(x_k \mid x_j)$	$\sum_{x_{k+1}} \Big\lvert P(x_{k+1} \mid x_k) - \sum_{x'_k} P(x_{k+1} \mid x'_k) P(x'_k \mid x_j) \Big\rvert$
$(1,1)$	$\frac{1}{2}(\frac{1}{2} + \delta)$	$2c(1 - 2\delta)$
$(1,0)$	$\frac{1}{2}(\frac{1}{2} - \delta)$	$2c(1 + 2\delta)$
$(0,1)$	$\frac{1}{2}(\frac{1}{2} - \delta)$	$2c(1 + 2\delta)$
$(0,0)$	$\frac{1}{2}(\frac{1}{2} + \delta)$	$2c(1 - 2\delta)$

Therefore, we have

$$m^B(X_k; X_{k+1} \mid X_j) = 2c(1 - 4\delta^2) \geq 2c(1 - 4c^2) \geq c \,,$$

when $2(1 - 4c^2) \geq 1 \iff c^2 \leq 1/8$, which is implied by $c \leq 1/4$.

For the second case, we can rewrite

$$\begin{aligned} & m^B(X_k; X_{k+1} \mid X_j) \\ &= \sum_{x_k, x_{k+1}, x_j} \Big| P(x_{k+1}, x_k, x_j) - P(x_{k+1} \mid x_j) P(x_k, x_j) \Big| \\ &= \sum_{x_k, x_{k+1}, x_j} \Big| P(x_j \mid x_{k+1}) P(x_{k+1} \mid x_k) P(x_k) - P(x_j \mid x_k) P(x_k) P(x_{k+1} \mid x_j) \Big| \\ &= \sum_{x_k, x_{k+1}, x_j} P(x_k) P(x_j \mid x_{k+1}) \Big| P(x_{k+1} \mid x_k) - \sum_{x'_{k+1}} P(x_j \mid x'_{k+1}) P(x'_{k+1} \mid x_k) \Big| . \end{aligned}$$

Again, by Lemma B.5.3, suppose $P(x_j \mid x_{k+1}) = 1/2 \pm \delta$. The calculation is summarized in the table below:

(x_k, x_{k+1}, x_j)	$P(x_j \mid x_{k+1})$	$\left\lvert P(x_{k+1} \mid x_k) - \sum_{x'_{k+1}} P(x_j \mid x'_{k+1}) P(x'_{k+1} \mid x_k) \right\rvert$
$(1,1,1)$	$\frac{1}{2} + \delta$	$c(1+2\delta)$
$(1,1,0)$	$\frac{1}{2} - \delta$	$c(1-2\delta)$
$(1,0,1)$	$\frac{1}{2} - \delta$	$c(1+2\delta)$
$(1,0,0)$	$\frac{1}{2} + \delta$	$c(1-2\delta)$
$(0,1,1)$	$\frac{1}{2} + \delta$	$c(1-2\delta)$
$(0,1,0)$	$\frac{1}{2} - \delta$	$c(1+2\delta)$
$(0,0,1)$	$\frac{1}{2} - \delta$	$c(1-2\delta)$
$(0,0,0)$	$\frac{1}{2} + \delta$	$c(1+2\delta)$

Therefore, we have

$$m^B(X_k; X_{k+1} \mid X_j) = 4c > c\,,$$

and P_T satisfies c-strong tree-faithfulness. □

B.6 Proof of Theorem 3.3.4 (Nonparametric Continuous Distribution)

Proof. The upper bound (C1) of ψ^{NP} is given by Theorem 5.6 in Neykov et al. [2021]. We proceed to specify the constructions of p_0^{NP} and p_1^{NP} and show they satisfy (C2) and (C3).

Let p_0^{NP} be independent uniform distributions $Unif^3[0,1]$. Thus, p_0^{NP} is Markov to an empty graph, and satisfies Lipschitz and smoothness condition. For p_1^{NP}, we design it to be Markov to a V-structure $X \to Z \leftarrow Y$, which is a three-node poly-forest. Under this graph, a faithful distribution is supposed to have $X \perp\!\!\!\perp Y$ while $X \not\perp\!\!\!\perp Y \mid Z$ and $X \not\perp\!\!\!\perp Z, Y \not\perp\!\!\!\perp Z$. In particular, p_1^{NP} is specified as follows (we will suppress the superscript and subscript by writing it as p to avoid notation clutter): $X, Y \sim Unif[0,1]$ and $p_{Z \mid X,Y}$ is a mixture of perturbation to uniform distribution, whose component depends on indepen-

dent multi-dimensional Radmacher random variables $\Delta \in \{-1,1\}^{m' \times m'}, \nu \in \{-1,1\}^m$ for some positive integers m, m' which are specified later. Then

$$p(Z \mid X, Y) = \mathbb{E}_{\Delta,\nu}\left[1+\gamma_\Delta(X,Y)\eta_\nu(Z)\right],$$

where

$$\begin{aligned}
\gamma_\Delta(x,y) &= \rho^2 \sum_{i\in[m']} \sum_{j\in[m']} \Delta_{ij} h_{ij,m'}(x,y) \\
\eta_\nu(z) &= \rho \sum_{j\in[m]} \nu_j h_{j,m}(z) \\
h_{ij,m'}(x,y) &= \begin{cases} \sqrt{m'}^2 \widetilde{h}(m'x-i+1, m'y-j+1) & \forall (x,y) \in [\frac{i-1}{m}, \frac{i}{m}] \otimes [\frac{j-1}{m}, \frac{j}{m}] \\ 0 & \text{otherwise} \end{cases} \\
h_{j,m}(z) &= \begin{cases} \sqrt{m} h(mz-j+1) & \forall z \in [\frac{j-1}{m}, \frac{j}{m}] \\ 0 & \text{otherwise} \end{cases} \\
\int \widetilde{h}(x,y)dx &= \sqrt{m'}h(y) \quad \int \widetilde{h}(x,y)dy = \sqrt{m'}h(x)\,.
\end{aligned}$$

for some function $h(x)$ infinitely differentiable on $[0,1]$ such that

$$\int h(x)dx = 0, \int h^2(x)dx = 1, \int |h(x)|dx = b_1, \int |\widetilde{h}(x,y)|dxdy = b_2, \|h\|_\infty \vee \|h'\|_\infty \le a\,,$$

for some $\rho > 0$ that will be specified later, and some constants $a, b_1, b_2 > 0$.

Now we show that each p satisfy the c-strong tree-faithfulness and smoothness conditions. Since the operation $\mathbb{E}_{\Delta,\nu}$ is linear, we consider one instance of (Δ, ν) in the following discussion.

c-**strong Tree-faithfulness** We highlight several important observations:

$$
\begin{aligned}
p(z) &= \int_{x,y} p(z \mid x,y)p(x)p(y) = \int_{x,y} p(z \mid x,y) = 1 \\
p(x \mid z) &= \frac{p(x,z)}{p(z)} = p(x,z) = \int_{y} p(z \mid x,y) = 1 + \Big[\rho^2 \sum_i \Big(\sum_j \Delta_{ij}\Big) h_{i,m'}(x)\Big]\eta_\nu(z) \\
p(y \mid z) &= \frac{p(y,z)}{p(z)} = p(y,z) = \int_{x} p(z \mid x,y) = 1 + \Big[\rho^2 \sum_j \Big(\sum_i \Delta_{ij}\Big) h_{j,m'}(y)\Big]\eta_\nu(z)\,.
\end{aligned}
$$

Therefore, tree-faithfulness is satisfied (while strong version still needs to be shown):

$$
\begin{aligned}
p(x,z) - p(x)p(z) &= \Big[\rho^2 \sum_i \Big(\sum_j \Delta_{ij}\Big) h_{i,m'}(x)\Big]\eta_\nu(z) \neq 0 \\
p(y,z) - p(y)p(z) &= \Big[\rho^2 \sum_j \Big(\sum_i \Delta_{ij}\Big) h_{j,m'}(y)\Big]\eta_\nu(z) \neq 0 \\
p(x,y \mid z) - p(x \mid z)p(y \mid z) &= \Big\{\gamma_\Delta(x,y) \\
&\quad - \Big[\rho^2 \sum_j \Big(\sum_i \Delta_{ij}\Big) h_{j,m'}(y)\Big] - \Big[\rho^2 \sum_i \Big(\sum_j \Delta_{ij}\Big) h_{i,m'}(x)\Big]\Big\}\eta_\nu(z) \\
&\quad - \rho^4 \Big[\sum_i \Big(\sum_j \Delta_{ij}\Big) h_{i,m'}(x)\Big]\Big[\sum_j \Big(\sum_i \Delta_{ij}\Big) h_{j,m'}(y)\Big]\eta_\nu^2(z) \neq 0\,.
\end{aligned}
$$

By Lemma B.4 in Neykov et al. [2021], it suffices to show the above three nonzero quantities are bounded away from zero in L_1. Specifically, because we have for any i or j,

$$
0.7\sqrt{m'} \leq \mathbb{E}_\Delta |\sum_j \Delta_{ij}| = \mathbb{E}_\Delta |\sum_i \Delta_{ij}| \leq \sqrt{m'}\,.
$$

Then

$$
\begin{aligned}
&\|p(x,y\,|\,z)-p(x\,|\,z)p(y\,|\,z)\|_1\\
\geq&\rho^2\int\Big|\sum_{i,j}\Delta_{ij}\Big[h_{ij,m'}(x,y)-h_{i,m'}(x)-h_{j,m'}(y)\Big]\Big|\Big|\eta_v(z)\Big|\\
&-\rho^4\int\Big|\Big[\sum_i\Big(\sum_j\Delta_{ij}\Big)h_{i,m'}(x)\Big]\Big[\sum_j\Big(\sum_i\Delta_{ij}\Big)h_{j,m'}(y)\Big]\Big|\Big|\eta_v^2(z)\Big|\\
\geq&\Big(\rho^2(m')^2\times\frac{1}{m'}g\Big)\times\Big(\rho\sqrt{m}a\Big)-\rho^4\Big(m'\times\sqrt{m'}\times\frac{1}{\sqrt{m'}}a\Big)^2\times\Big(\rho^2m\Big)\\
=&\rho^3m'\sqrt{m}\times ag-\rho^6(m')^2m\times a^2\,,
\end{aligned}
$$

where $g=\int|\widetilde{h}(x,y)-\frac{1}{\sqrt{m'}}h(x)-\frac{1}{\sqrt{m'}}h(y)|dxdy$ being a constant. And we need for either $w=x$ or y,

$$
\begin{aligned}
\|p(w,z)-p(w)p(z)\|_1&=\int\rho^2\Big|\sum_i\Big(\sum_j\Delta_{ij}\Big)h_{i,m'}(w)\Big|\Big|\eta_\nu(z)\Big|\\
&\geq\Big(\rho^2m'\times 0.7\sqrt{m'}\times\frac{1}{\sqrt{m'}}a\Big)\times\Big(\rho\sqrt{m}a\Big)\\
&=\rho^3m'\sqrt{m}a^2\,.
\end{aligned}
$$

We will lower bound both of them above by the order of c when specifying the parameters.

Lipschitz Condition We then check the TV distance between $p(x,y\,|\,z)$ and $p(x,y\,|\,z')$:

$$
\begin{aligned}
\|p(x,y\,|\,z)-p(x,y\,|\,z')\|_1&\leq\int\Big|\gamma_\Delta(x,y)\Big|\Big|\eta_\nu(z)-\eta_\nu(z')\Big|\\
&\leq b_2m'\rho^2\times\Big|\eta_\nu(z)-\eta_\nu(z')\Big|\\
&\leq\Big[(b_2m'\rho^2)\times(m^{1/2}m\rho\|h'\|_\infty)\Big]|z-z'|\,.
\end{aligned}
$$

We will need the term in the bracket to be smaller than some constant.

Smoothness Condition We check the Hölder smoothness of $p(x, y \mid z)$ in (x, y). Following the proof Theorem 4.2 in Neykov et al. [2021], for any $k \leq \lfloor s \rfloor$

$$\begin{aligned}
&\left| \frac{\partial^k}{\partial x^k} \frac{\partial^{\lfloor s \rfloor - k}}{\partial y^{\lfloor s \rfloor - k}} p(x, y \mid z) - \frac{\partial^k}{\partial x^k} \frac{\partial^{\lfloor s \rfloor - k}}{\partial y^{\lfloor s \rfloor - k}} p(x', y' \mid z) \right| \\
\leq &\left| \frac{\partial^k}{\partial x^k} \frac{\partial^{\lfloor s \rfloor - k}}{\partial y^{\lfloor s \rfloor - k}} \gamma_\Delta(x, y) \eta_\nu(z) - \frac{\partial^k}{\partial x^k} \frac{\partial^{\lfloor s \rfloor - k}}{\partial y^{\lfloor s \rfloor - k}} \gamma_\Delta(x', y') \eta_\nu(z) \right| \\
\leq &m^{1/2} \rho \|h\|_\infty \left| \frac{\partial^k}{\partial x^k} \frac{\partial^{\lfloor s \rfloor - k}}{\partial y^{\lfloor s \rfloor - k}} \gamma_\Delta(x, y) - \frac{\partial^k}{\partial x^k} \frac{\partial^{\lfloor s \rfloor - k}}{\partial y^{\lfloor s \rfloor - k}} \gamma_\Delta(x', y') \right| \\
\leq &(\rho m^{1/2} a) \times \left(\rho^2 (m')^s \sqrt{m'^2} \left\| \frac{\partial^k}{\partial x^k} \frac{\partial^{\lfloor s \rfloor - k}}{\partial y^{\lfloor s \rfloor - k}} \widetilde{h}(x, y) \right\|_\infty \right).
\end{aligned}$$

Thus, the quantity that need to be bounded above by constant is

$$\rho^3 m^{1/2} {m'}^{1+s}.$$

Parameter Choice All in all, we need the choice of m, m', ρ to satisfy the following requirements:

- c-strong tree-faithfulness:

$$\begin{aligned}
\rho^3 m^{1/2} m' - \rho^6 m (m')^2 &\gtrsim c \\
\rho^3 m^{1/2} m' &\gtrsim c
\end{aligned}$$

- Lipschitz condition:

$$\rho^3 m^{3/2} m' \lesssim 1$$

- Smoothness condition:

$$\rho^3 m^{1/2} {m'}^{1+s} \lesssim 1$$

By setting $m' = m^{1/s}, \rho^3 \asymp m^{-(3/2+1/s)}, m \asymp c^{-1}$ and assuming c is sufficiently small, above requirements are satisfied. Thus, p falls into the considered model class.

KL Divergence We start by bounding the χ^2 divergence then upper bound KL divergence by χ^2 divergence:

$$\begin{aligned}\chi^2(p_1^{NP} \| p_0^{NP}) + 1 = & \int \frac{(p_1^{NP})^2}{p_0^{NP}} \\ = & \mathbb{E}_{\Delta,\Delta',\nu,\nu'} \int \left[1 + \gamma_\Delta(x,y)\eta_\nu(z)\right] \times \left[1 + \gamma_{\Delta'}(x,y)\eta_{\nu'}(z)\right].\end{aligned}$$

The integral is

$$\int \left[1 + \gamma_\Delta(x,y)\gamma_{\Delta'}(x,y)\eta_\nu(z)\eta_{\nu'}(z)\right] = 1 + \rho^6 \langle \Delta, \Delta' \rangle \langle \nu, \nu' \rangle .$$

Therefore,

$$\begin{aligned}\chi^2(p_1^{NP} \| p_0^{NP}) + 1 &= \mathbb{E}_{\Delta,\Delta',\nu,\nu'} \left\{ 1 + \rho^6 \langle \Delta, \Delta' \rangle \langle \nu, \nu' \rangle \right\} \\ &\leq \mathbb{E}_{\Delta,\Delta',\nu,\nu'} \left\{ \exp\left(\rho^6 \langle \Delta, \Delta' \rangle \langle \nu, \nu' \rangle \right) \right\}.\end{aligned}$$

Following the proof Theorem 4.2 in Neykov et al. [2021], we can upper bound the right hand side above and obtain

$$\chi^2(p_1^{NP} \| p_0^{NP}) + 1 \leq \sqrt{\frac{1}{1 - (\rho^6)^2 m m'^2}}\,.$$

Since the function $f(t) = 1/\sqrt{1-t^2} \leq 2t + 1$ for t sufficiently small, with the choice of ρ, m, m', we have for sufficiently small c,

$$\chi^2(p_1^{NP} \| p_0^{NP}) + 1 \leq C_0 \times c^{\frac{5s+2}{2s}} + 1\,,$$

for some constant C_0. Therefore, we arrive at

$$\mathbf{KL}(p_1^{NP} \| p_0^{NP}) \leq \chi^2(p_1^{NP} \| p_0^{NP}) \lesssim c^{\frac{5s+2}{2s}}$$

Application of Corollary D.2.4 completes the proof. □

APPENDIX C

PROOFS IN CHAPTER 4

C.1 Proof of Lemma 4.2.1 (Variable selection Upper Bound)

Proof. It is easy to see that the estimator succeeds when $\|\Pi_{S_*}^{\perp} Y\|^2$ is the smallest, i.e.

$$\begin{aligned}\mathbb{P}(\widehat{S}^{\mathrm{BSS}} \neq S_*) &= \mathbb{P}\Big[\bigcup_{T \in \mathcal{T}_{d,s} \setminus \{S_*\}} \Big\{\|\Pi_T^{\perp} Y\|^2 - \|\Pi_{S_*}^{\perp} Y\|^2 < 0\Big\}\Big] \\ &\leq \sum_{\ell=1}^{s} \sum_{\substack{T \in \mathcal{T}_{d,s} \setminus \{S_*\} \\ |\mathcal{T}_{d,s} \setminus T| = \ell}} \mathbb{P}\Big[\|\Pi_T^{\perp} Y\|^2 - \|\Pi_{S_*}^{\perp} Y\|^2 < 0\Big].\end{aligned}$$

Now we introduce a deviation bound for the error probability, whose proof can be found below. For any $S, T \in \mathcal{T}_{d,s}$, we define the signal to distinguish S and T to be

$$\Delta_1(S, T) := \beta_{S \setminus T}^{\top} \Sigma_{S \setminus T \,|\, T} \beta_{S \setminus T} / \sigma^2 .$$

Then if $|S_* \setminus T| = \ell$, it is easy to see that $\Delta_1(S_*, T) \geq \ell \Delta_1$.

Lemma C.1.1. *If $n - s \geq \frac{32}{\Delta_1}$, then for any $T \in \mathcal{T}_{d,s} \setminus S_*$,*

$$\mathbb{P}\Big[\|\Pi_T^{\perp} Y\|^2 - \|\Pi_{S_*}^{\perp} Y\|^2 < 0\Big] \leq 5 \exp\Big(-(n-s)\frac{\min(\Delta_1(S_*, T), 1)}{1024}\Big).$$

Applying Lemma C.1.1, we have

$$\begin{aligned}
\mathbb{P}(\widehat{S}^{\mathrm{BSS}} \neq S_*) &\leq \sum_{\ell=1}^{s} \sum_{\substack{T \in \mathcal{T}_{d,s} \setminus \{S_*\} \\ |S_* \setminus T| = \ell}} \mathbb{P}\left[\|\Pi_T^{\perp} Y\|^2 - \|\Pi_{S_*}^{\perp} Y\|^2 < 0 \right] \\
&\leq \sum_{\ell=1}^{s} \sum_{\substack{T \in \mathcal{T}_{d,s} \setminus \{S_*\} \\ |S_* \setminus T| = \ell}} 5 \exp\left(- (n-s) \frac{\min(\Delta_1(S_*, T), 1)}{1024} \right) \\
&\leq \max_{\ell} s \times \binom{s}{\ell} \binom{d-s}{\ell} \times 5 \exp\left(- (n-s) \frac{\min(\ell \Delta_1, 1)}{1024} \right).
\end{aligned}$$

Since $s \leq d/2$, $\binom{s}{\ell} \leq \binom{d-s}{\ell}$ and

$$\log 5s \leq \max_{\ell} \log 5 \binom{s}{\ell} \leq \max_{\ell} 2 \log \binom{s}{\ell}.$$

Therefore, the error probability

$$\begin{aligned}
\mathbb{P}(\widehat{S}^{\mathrm{BSS}} \neq S_*) &\leq \max_{\ell} \exp\left(\log 5s + \log \binom{s}{\ell} + \log \binom{d-s}{\ell} - (n-s) \frac{\min(\ell \Delta_1, 1)}{1024} \right) \\
&\leq \max_{\ell} \exp\left(4 \log \binom{d-s}{\ell} - (n-s) \frac{\min(\ell \Delta_1, 1)}{1024} \right).
\end{aligned}$$

Setting the RHS to be smaller than δ for all $\ell \in [s]$, we have desired sample complexity. □

Proof of Lemma C.1.1.

$$\begin{aligned}
\mathbb{P}\left(\|\Pi_T^{\perp} Y\|^2 - \|\Pi_{S_*}^{\perp} Y\|^2 < 0 \right) &= \mathbb{P}\left(\frac{\|\Pi_T^{\perp} Y\|^2 - \|\Pi_{S_*}^{\perp} Y\|^2}{(n-s)\sigma^2} < 0 \right) \\
&\leq \mathbb{P}\left(\frac{\|\Pi_T^{\perp} Y\|^2 - \|\Pi_{S_*}^{\perp} Y\|^2}{(n-s)\sigma^2} \leq \Delta_1(S_*, T)/4 \right) \\
&\leq \mathbb{P}\left(\frac{\left| \|\Pi_T^{\perp} Y\|^2 - \|\Pi_T^{\perp} \epsilon\|^2 \right|}{(n-s)\sigma^2} \leq \Delta_1(S_*, T)/2 \right) \\
&\quad + \mathbb{P}\left(\frac{\left| \|\Pi_T^{\perp} \epsilon\|^2 - \|\Pi_{S_*}^{\perp} \epsilon\|^2 \right|}{(n-s)\sigma^2} \geq \Delta_1(S_*, T)/4 \right).
\end{aligned}$$

We bound these two terms separately using the lemmas below.

Lemma C.1.2. *If* $n - s \geq \frac{32}{\Delta_1}$, *then*

$$\mathbb{P}\bigg(\frac{\big|\|\Pi_T^{\perp}\epsilon\|^2 - \|\Pi_{S_*}^{\perp}\epsilon\|^2\big|}{(n-s)\sigma^2} \geq \Delta_1(S_*, T)/4\bigg) \leq 2\exp\bigg(-(n-s)\frac{\Delta_1(S_*, T)}{32}\bigg).$$

Lemma C.1.3.

$$\begin{aligned}
&\mathbb{P}\bigg(\frac{\big|\|\Pi_T^{\perp}Y\|^2 - \|\Pi_T^{\perp}\epsilon\|^2\big|}{(n-s)\sigma^2} \leq \Delta_1(S_*, T)/2\bigg) \\
&\leq 2\exp\bigg(-(n-s)\frac{\min(\Delta_1(S_*, T), \sqrt{\Delta_1(S_*, T)})}{1024}\bigg) \\
&+ \exp\bigg(-(n-s)/256\bigg).
\end{aligned}$$

Combining Lemma C.1.2 and C.1.3, we complete the proof. □

Proof of Lemma C.1.2.

$$\begin{aligned}
&\|\Pi_T^{\perp}\epsilon\|^2/\sigma^2 - \|\Pi_{S_*}^{\perp}\epsilon\|^2/\sigma^2 \\
&= \epsilon^{\top}\bigg((I_n - X_T(X_T^{\top}X_T)^{-1}X_T) - (I_n - X_{S_*}(X_{S_*}^{\top}X_{S_*})^{-1}X_{S_*})\bigg)\epsilon/\sigma^2 \\
&= \epsilon^{\top}\bigg(X_{S_*}(X_{S_*}^{\top}X_{S_*})^{-1}X_{S_*} - X_T(X_T^{\top}X_T)^{-1}X_T)\bigg)\epsilon/\sigma^2 \\
&= \|(\Pi_{S_*} - \Pi_{S_*\cap T})\epsilon\|^2/\sigma^2 - \|(\Pi_T - \Pi_{S_*\cap T})\epsilon\|^2/\sigma^2 \\
&= Z - \widetilde{Z},
\end{aligned}$$

where $Z, \widetilde{Z} \sim \chi^2_\ell$ given X_T and X_{S_*} because both $\Pi_{S_*} - \Pi_{S_*\cap T}$ and $\Pi_T - \Pi_{S_*\cap T}$ are

projection matrix with trace equal to $|S_* \setminus T| = |T \setminus S_*| = \ell$. Therefore,

$$\begin{aligned}
\mathbb{P}\left(\frac{\left|\|\Pi_T^{\perp}\epsilon\|^2 - \|\Pi_{S_*}^{\perp}\epsilon\|^2\right|}{(n-s)\sigma^2} \geq \Delta_1(S_*,T)/4\right) &= \mathbb{P}\left(\frac{|Z-\widetilde{Z}|}{(n-s)} \geq \Delta_1(S_*,T)/4\right) \\
&\leq \mathbb{P}\left(\frac{|Z-\ell|}{\ell} \geq \frac{(n-s)\Delta_1(S_*,T)}{8\ell}\right) \\
&\quad + \mathbb{P}\left(\frac{|\widetilde{Z}-\ell|}{\ell} \geq \frac{(n-s)\Delta_1(S_*,T)}{8\ell}\right) \\
&= 2\mathbb{P}\left(\frac{|Z-\ell|}{\ell} \geq \frac{(n-s)\Delta_1(S_*,T)}{8\ell}\right) \\
&\leq 2\exp\left(-(n-s)\frac{\Delta_1(S_*,T)}{32}\right).
\end{aligned}$$

We apply Lemma D.1.2 for the last inequality since $n - s \geq \frac{32}{\Delta_1} = \frac{32\ell}{\ell\Delta_1} \geq \frac{32}{\Delta_1(S_*,T)}$. □

Proof of Lemma C.1.3. We can decompose the Gaussian variables

$$X_{S_*\setminus T} = X_T(\Sigma_{TT})^{-1}\Sigma_{T(S_*\setminus T)} + E_{S_*\setminus T},$$

where each row of $E_{S_*\setminus T}$ is i.i.d. from $\mathcal{N}(0, \Sigma_{S_*\setminus T\,|\,T})$ and $E_{S_*\setminus T} \perp\!\!\!\perp X_T$. Thus

$$\begin{aligned}
\Pi_T^{\perp}Y &= \Pi_T^{\perp}(X_{S_*}\beta_{S_*} + \epsilon) \\
&= \Pi_T^{\perp}(X_{S_*\cap T}\beta_{S_*\cap T} + X_{S_*\setminus T}\beta_{S_*\setminus T} + \epsilon) \\
&= \Pi_T^{\perp}(X_{S_*\setminus T}\beta_{S_*\setminus T} + \epsilon) \\
&= \Pi_T^{\perp}\left((X_T(\Sigma_{TT})^{-1}\Sigma_{T(S_*\setminus T)} + E_{S_*\setminus T})\beta_{S_*\setminus T} + \epsilon\right) \\
&= \Pi_T^{\perp}(E_{S_*\setminus T}\beta_{S_*\setminus T} + \epsilon),
\end{aligned}$$

where $E_{S_*\setminus T}\beta_{S_*\setminus T}$ is a vector and each entry is i.i.d. from $\mathcal{N}(0, \Delta_1(S_*, T)\sigma^2)$. Therefore,

$$\begin{aligned}\|\Pi_T^\perp Y\|^2 - \|\Pi_T^\perp \epsilon\|^2 &= \|\Pi_T^\perp E_{S_*\setminus T}\beta_{S_*\setminus T}\|^2 + 2\langle \Pi_T^\perp E_{S_*\setminus T}\beta_{S_*\setminus T}, \epsilon\rangle \\ &= \sigma^2\Delta_1(S_*, T)\|\Pi_T^\perp U\|^2 + 2\sigma^2\sqrt{\Delta_1(S_*, T)}\langle \Pi_T^\perp U, \Pi_T^\perp U_\epsilon\rangle\,,\end{aligned}$$

where $U, U_\epsilon \sim \mathcal{N}(0, I_n)$ and $U \perp\!\!\!\perp U_\epsilon$. Then

$$\begin{aligned}&\mathbb{P}\bigg(\frac{\big|\|\Pi_T^\perp Y\|^2 - \|\Pi_T^\perp \epsilon\|^2\big|}{(n-s)\sigma^2} \le \Delta_1(S_*, T)/2\bigg) \\ \le\ &\mathbb{P}\bigg(\frac{\sigma^2\Delta_1(S_*, T)\|\Pi_T^\perp U\|^2}{\sigma^2(n-s)} \le \frac{3}{4}\Delta_1(S_*, T)\bigg) \\ +\ &\mathbb{P}\bigg(\frac{2\sigma^2\sqrt{\Delta_1(S_*, T)}\langle \Pi_T^\perp U, \Pi_T^\perp U_\epsilon\rangle}{\sigma^2(n-s)} \le -\frac{1}{4}\Delta_1(S_*, T)\bigg)\,.\end{aligned}$$

For the first term, apply Lemma D.1.2,

$$\begin{aligned}\mathbb{P}\bigg(\frac{\sigma^2\Delta_1(S_*, T)\|\Pi_T^\perp U\|^2}{\sigma^2(n-s)} \le \frac{3}{4}\Delta_1(S_*, T)\bigg) &= \mathbb{P}\bigg(\frac{\chi^2_{n-s}}{n-s} \le \frac{3}{4}\bigg) \\ &= \mathbb{P}\bigg(\frac{\chi^2_{n-s}}{n-s} - 1 \le -\frac{1}{4}\bigg) \\ &\le \exp(-(n-s)/256)\,.\end{aligned}$$

For the second term, since

$$\begin{aligned}2\langle \Pi_T^\perp U, \Pi_T^\perp U_\epsilon\rangle &= \frac{1}{2}\Big(\|\Pi_T^\perp(U+U_\epsilon)\|^2 - \|\Pi_T^\perp(U-U_\epsilon)\|^2\Big) \\ &= \|\Pi_T^\perp \frac{U+U_\epsilon}{\sqrt{2}}\|^2 - \|\Pi_T^\perp \frac{U-U_\epsilon}{\sqrt{2}}\|^2 \\ &= W - \widetilde{W}\,,\end{aligned}$$

where $W, \widetilde{W} \sim \chi^2_{n-s}$. Therefore, apply Lemma D.1.2,

$$\begin{aligned}
\mathbb{P}\bigg(\frac{2\sigma^2\sqrt{\Delta_1(S_*,T)}\langle\Pi_T^{\perp}U,\Pi_T^{\perp}U_\epsilon\rangle}{\sigma^2(n-s)} &\leq -\frac{1}{4}\Delta_1(S_*,T)\bigg) \\
&= \mathbb{P}\bigg(\frac{W-\widetilde{W}}{n-s} \leq -\frac{\sqrt{\Delta_1(S_*,T)}}{4}\bigg) \\
&\leq \mathbb{P}\bigg(\frac{|W-(n-s)|}{n-s} \geq \frac{\sqrt{\Delta_1(S_*,T)}}{8}\bigg) \\
&\quad + \mathbb{P}\bigg(\frac{|\widetilde{W}-(n-s)|}{n-s} \geq \frac{\sqrt{\Delta_1(S_*,T)}}{8}\bigg) \\
&\leq 2\exp\bigg(-(n-s)\frac{\min(\Delta_1(S_*,T),\sqrt{\Delta_1(S_*,T)})}{1024}\bigg).
\end{aligned}$$

□

C.2 Proof of Theorem 4.2.3 (Variable Selection Lower Bound)

Proof. Consider a covariance matrix of X:

$$\Sigma = (1-\rho)I_d + \rho\mathbf{1}_d\mathbf{1}_d^\top$$

where $\rho = 1-\omega$. Then for any distinct $S, T \in \mathcal{T}_{d,s}$ with $|S \setminus T| = r$, we can calculate the conditional covariance matrix

$$\begin{aligned}
\Sigma_{S\setminus T\,|\,T} &= \Sigma_{(S\setminus T)(S\setminus T)} - \Sigma_{(S\setminus T)T}\Sigma_{TT}^{-1}\Sigma_{T(S\setminus T)} \\
&= (1-\rho)I_r + \rho\mathbf{1}_r\mathbf{1}_r^\top - \rho\mathbf{1}_r\mathbf{1}_s^\top \times \left((1-\rho)I_s + \rho\mathbf{1}_s\mathbf{1}_s^\top\right)^{-1} \times \rho\mathbf{1}_s\mathbf{1}_r^\top \\
&= (1-\rho)I_r + \rho\mathbf{1}_r\mathbf{1}_r^\top - \rho\mathbf{1}_r\mathbf{1}_s^\top \times \left(\frac{1}{1-\rho}\big(I_s - \frac{\rho}{1-\rho+s\rho}\mathbf{1}_s\mathbf{1}_s^\top\big)\right) \times \rho\mathbf{1}_s\mathbf{1}_r^\top \\
&= (1-\rho)\left(I_r + \frac{\rho}{1-\rho+s\rho}\mathbf{1}_r\mathbf{1}_r^\top\right),
\end{aligned}$$

then the minimum eigenvalue is

$$\lambda_{\min}(\Sigma_{S\setminus T\,|\,T}) = \begin{cases} (1-\rho)\times(1+\frac{\rho}{1-\rho+s\rho}) & r=1 \\ 1-\rho & r\geq 2 \end{cases}.$$

Since $\Sigma_{S\setminus T\,|\,T}$ is independent with the choice of (S,T) but only depends on $|S\setminus T|$, this covariance matrix Σ satisfies the requirement on $\Omega_{d,s}(\omega)$:

$$\min_{S\in\mathcal{T}_{d,s}}\min_{T\in\mathcal{T}_{d,s}\setminus S}\lambda_{\min}(\Sigma_{S\setminus T\,|\,T}) = 1-\rho = \omega\,.$$

Now we can construct s many ensembles to establish s lower bounds, while they will lead to only one in the end. For each of them, we fix the covariance matrix Σ and coefficient vector $\beta = \beta_{\min}\mathbf{1}_d$, construct the ensemble solely by varying support. For the ℓ-th ensemble ($\ell = 1,2,\cdots,s$), let

$$\mathcal{S}'_\ell := \left\{S'\subseteq\{s,s+1,\cdots,d\} : |S'| = \ell\right\}.$$

Thus $|\mathcal{S}'_\ell| = \binom{d-s}{\ell}$. We consider set of supports:

$$\mathcal{S}_\ell := \left\{S : S = \{1,2,\cdots,s-\ell\}\cup S', S'\in\mathcal{S}'_\ell\right\}.$$

Thus $|\mathcal{S}_\ell| = |\mathcal{S}'_\ell| = \binom{d-s}{\ell}$ and each element $S\in\mathcal{S}_\ell$ determines a model $Y = X_S^\top\beta_S + \epsilon$. For any two supports $S,T\in\mathcal{S}_\ell$, write

$$S = \{1,2,\cdots,s-\ell\}\cup S'$$
$$T = \{1,2,\cdots,s-\ell\}\cup T'\,,$$

with $S', T' \in \mathcal{S}'_\ell$. Now we try to calculate the KL divergence between two models specified by S and T. We further denote $S'' = S' \setminus T'$, and $T'' = T' \setminus S'$ with $|S''| = |T''| = r \leq \ell$, and the models determined by S and T to be P_S and P_T. Therefore,

$$\begin{aligned}
\mathbf{KL}(P_S \| P_T) &= \mathbb{E}_{P_S} \log \frac{P_S}{P_T} \\
&= \mathbb{E}_{P_S} \log \frac{\exp\Big(-(Y - X_S^\top \beta_S)^2/2\sigma^2 \Big)}{\exp\Big(-(Y - X_T^\top \beta_T)^2/2\sigma^2 \Big)} \\
&= \mathbb{E}_X \mathbb{E}_\epsilon \frac{1}{2\sigma^2} \Big((X_S^\top \beta_S - X_T^\top \beta_T + \epsilon)^2 - \epsilon^2 \Big) \\
&= \mathbb{E}_X (X_S^\top \beta_S - X_T^\top \beta_T)^2/2\sigma^2 \\
&= \mathbb{E}_X (X_{S''}^\top \beta_{S''} - X_{T''}^\top \beta_{T''})^2/2\sigma^2 \\
&= \mathbf{1}_r^\top (\Sigma_s + \Sigma_t - \Sigma_{st} - \Sigma_{ts}) \mathbf{1}_r \times \frac{\beta_{\min}^2}{2\sigma^2} \,,
\end{aligned}$$

where $\Sigma_s := \Sigma_{S''S''}, \Sigma_t := \Sigma_{T''T''}, \Sigma_{st} := \Sigma_{S''T''}, \Sigma_{ts} := \Sigma_{T''S''}$. Then

$$\begin{aligned}
\mathbf{1}_r^\top (\Sigma_s + \Sigma_t - \Sigma_{st} - \Sigma_{ts}) \mathbf{1}_r = \mathbf{1}_r^\top (\Sigma_s - \Sigma_{st}\Sigma_t^{-1}\Sigma_{ts}) \mathbf{1}_r + \mathbf{1}_r^\top (\Sigma_t - \Sigma_{ts}\Sigma_s^{-1}\Sigma_{st}) \mathbf{1}_r \\
+ \mathbf{1}_r^\top (\Sigma_{st}\Sigma_t^{-1}(\Sigma_{ts} - \Sigma_t)) \mathbf{1}_r + \mathbf{1}_r^\top (\Sigma_{ts}\Sigma_s^{-1}(\Sigma_{st} - \Sigma_s)) \mathbf{1}_r \,.
\end{aligned}$$

The first two terms are the same, which are

$$\begin{aligned}
\mathbf{1}_r^\top (\Sigma_s - \Sigma_{st}\Sigma_t^{-1}\Sigma_{ts}) \mathbf{1}_r &= \mathbf{1}_r^\top \Big[(1-\rho)\big(I_r + \frac{\rho}{1-\rho+r\rho} \mathbf{1}_r \mathbf{1}_r^\top \big) \Big] \mathbf{1}_r \\
&= r(1-\rho) \times \Big(1 + \frac{r\rho}{1-\rho+r\rho} \Big) \\
&\leq 2r(1-\rho) \\
&= 2r\omega \leq 2\ell\omega \,.
\end{aligned}$$

The last two terms are the same, which are

$$\mathbf{1}_r^\top(\Sigma_{st}\Sigma_t^{-1}(\Sigma_{ts}-\Sigma_s))\mathbf{1}_r = \mathbf{1}_r^\top\left[\rho\mathbf{1}_r\mathbf{1}_r^\top \times \left(\frac{1}{1-\rho}\big(I_r - \frac{\rho}{1-\rho+r\rho}\mathbf{1}_r\mathbf{1}_r^\top\big)\right) \times (\rho-1)I_r\right]\mathbf{1}_r$$
$$= \frac{-\rho(1-\rho)}{1-\rho+r\rho} \times r^2 \leq 0\,.$$

Thus $\mathbf{KL}(P_S\|P_T) \leq 2\ell\beta_{\min}^2\omega/\sigma^2$, which holds for any two $S,T \in \mathcal{S}_\ell$ and leads to a upper bound for any two models in the ℓ-th ensemble. Finally, for the ℓ-th ensemble, we apply Fano's inequality Corollary D.2.3 with KL divergence upper bound $2\ell\beta_{\min}^2\omega/\sigma^2$ and ensemble cardinality $\binom{d-s}{\ell}$, which completes the proof. □

C.3 Proof of Theorem 4.2.5 (Variable Selection Upper Bound)

Proof. Recall that $|S_*| = s \leq \bar{s}$. Denote the event that S_* beats an opponent T with $|T| = j$:

$$\mathcal{E}(T,j) = \left\{\frac{\|\Pi_{S_*}^\perp Y\|^2}{n-s} + \frac{s}{4}\beta_{\min}^2\omega \leq \frac{\|\Pi_T^\perp Y\|^2}{n-j} + \frac{j}{4}\beta_{\min}^2\omega\right\},$$

then the estimator succeeds with

$$\mathbb{P}(\widehat{S}^{\text{BSSu}} = S_*) = \mathbb{P}\bigg(\bigcap_{j\in\{1,2,\ldots,\bar{s}\}}\ \bigcap_{T\in\mathcal{T}_{d,j}\setminus\{S_*\}} \mathcal{E}(T,j)\bigg).$$

Therefore, use ℓ for the distance from j to s, i.e. $\ell = |j - s|$,

$$\begin{aligned}
\mathbb{P}(\widehat{S}^{\mathrm{BSSu}} \neq S_*) &= \mathbb{P}\Big(\bigcup_{j \in [\bar{s}]} \bigcup_{T \in \mathcal{T}_{d,j} \setminus \{S_*\}} \overline{\mathcal{E}(T,j)} \Big) \\
&\leq \sum_{T \in \mathcal{T}_{d,s} \setminus \{S_*\}} \mathbb{P}(\overline{\mathcal{E}(T,s)}) + \sum_{j \neq s} \sum_{T \in \mathcal{T}_{d,j}} \mathbb{P}(\overline{\mathcal{E}(T,j)}) \\
&= \sum_{T \in \mathcal{T}_{d,s} \setminus \{S_*\}} \mathbb{P}(\overline{\mathcal{E}(T,s)}) \\
&\quad + \sum_{\ell=1}^{s} \sum_{T \in \mathcal{T}_{d,s-\ell}} \mathbb{P}(\overline{\mathcal{E}(T,s-\ell)}) + \sum_{\ell=1}^{\bar{s}-s} \sum_{T \in \mathcal{T}_{d,s+\ell}} \mathbb{P}(\overline{\mathcal{E}(T,s+\ell)}) \,.
\end{aligned}$$

The first term is controlled by Theorem 4.2.2, now let's look at remaining two. Use $k := |S_* \cap T|$ for the overlap between T and S_*, define

$$\begin{aligned}
A_1 &:= \sum_{\ell=1}^{s} \sum_{T \in \mathcal{T}_{d,s-\ell}} \mathbb{P}(\overline{\mathcal{E}(T,s-\ell)}) = \sum_{\ell=1}^{s} \sum_{k=0}^{s-\ell} \sum_{\substack{T \in \mathcal{T}_{d,s-\ell} \\ |T \cap S_*| = k}} \mathbb{P}(\overline{\mathcal{E}(T,s-\ell)}) \\
A_2 &:= \sum_{\ell=1}^{\bar{s}-s} \sum_{T \in \mathcal{T}_{d,s+\ell}} \mathbb{P}(\overline{\mathcal{E}(T,s+\ell)}) = \sum_{\ell=1}^{\bar{s}-s} \sum_{k=0}^{s} \sum_{\substack{T \in \mathcal{T}_{d,s+\ell} \\ |T \cap S_*| = k}} \mathbb{P}(\overline{\mathcal{E}(T,s+\ell)}) \,.
\end{aligned}$$

The cardinality of the innermost sums of A_1 and A_2 are

$$\begin{aligned}
\binom{s}{k}\binom{d-s}{s-k-\ell} &= \binom{s}{s-k}\binom{d-s}{s-k-\ell} \leq \binom{d-s}{s-k}^2 \\
\binom{s}{k}\binom{d-s}{s-k+\ell} &= \binom{s}{s-k}\binom{d-s}{s-k+\ell} \leq \binom{d-s}{s-k+\ell}^2
\end{aligned}$$

respectively. The last inequality is because $s \leq \bar{s} \leq d/2$, $s - k + \ell \leq \bar{s}$, then $\binom{s}{a} \leq \binom{d-s}{a}$ for $a \leq s$. Now we analyze the error probability respectively. Denote $v := \omega\beta_{\min}^2/\sigma^2$ as a short hand.

For $|T| = s - \ell$ and $|T \cap S_*| = k$, we have $|S_* \setminus T| = s - k \geq \ell$, $|T \setminus S_*| = s - \ell - k$.

$$\begin{aligned}
\mathbb{P}(\overline{\mathcal{E}(T, s-\ell)}) &= \mathbb{P}\bigg(\frac{\|\Pi_T^{\perp} Y\|^2/\sigma^2}{n-(s-\ell)} - \frac{\|\Pi_{S_*}^{\perp} Y\|^2/\sigma^2}{n-s} \leq \frac{\ell}{4}\omega\beta_{\min}^2/\sigma^2 \bigg) \\
&\leq \mathbb{P}\bigg(\frac{\|\Pi_T^{\perp} Y\|^2/\sigma^2}{n-(s-\ell)} - \frac{\|\Pi_{S_*}^{\perp} Y\|^2/\sigma^2}{n-s} \leq \frac{|S_* \setminus T|}{4} v \bigg) \\
&\leq \mathbb{P}\bigg(\frac{\|\Pi_T^{\perp} Y\|^2/\sigma^2}{n-(s-\ell)} - \frac{\|\Pi_{S_*}^{\perp} Y\|^2/\sigma^2}{n-s} \leq \frac{1}{4}\Delta_1(S_*, T) \bigg).
\end{aligned}$$

For $|T| = s + \ell$ and $|T \cap S_*| = k$, we have $|S_* \setminus T| = s - k$, $|T \setminus S_*| = s + \ell - k$.

$$\begin{aligned}
\mathbb{P}(\overline{\mathcal{E}(T, s+\ell)}) &= \mathbb{P}\bigg(\frac{\|\Pi_T^{\perp} Y\|^2/\sigma^2}{n-(s+\ell)} - \frac{\|\Pi_{S_*}^{\perp} Y\|^2/\sigma^2}{n-s} \leq -\frac{\ell}{4}\omega\beta_{\min}^2/\sigma^2 \bigg) \\
&\leq \mathbb{P}\bigg(\frac{\|\Pi_T^{\perp} Y\|^2/\sigma^2}{n-(s+\ell)} - \frac{\|\Pi_{S_*}^{\perp} Y\|^2/\sigma^2}{n-s} \leq \frac{1}{4}\Big(\Delta_1(S_*, T) - \ell v\Big) \bigg).
\end{aligned}$$

Now we introduce following lemma to control the error probability. The proof will be given in the sequel.

Lemma C.3.1. *If $n - \bar{s} \geq \frac{96}{\beta_{\min}^2 \omega/\sigma^2}$, then for any $T \in \mathcal{T}_d^{\bar{s}} \setminus \{S_*\}$, let $\ell' := \max\{|T| - |S_*|, 0\}$,*

$$\begin{aligned}
&\mathbb{P}\bigg[\frac{\|\Pi_T^{\perp} Y\|^2}{(n-|T|)\sigma^2} - \frac{\|\Pi_{S_*}^{\perp} Y\|^2}{(n-s)\sigma^2} \leq \frac{1}{4}\Big(\Delta_1(S_*, T) - \ell'\omega\beta_{\min}^2/\sigma^2\Big) \bigg] \\
\leq & 5\exp\bigg(-(n-\bar{s}) \frac{\min\Big((|S_* \setminus T| + \ell')\omega\beta_{\min}^2/\sigma^2, 1\Big)}{9216} + \frac{|T \setminus S_*|}{4} \bigg).
\end{aligned}$$

For each error probability in A_1, $s < j$ thus $\ell' = 0$. While for each error probability in

A_2, $\ell' = s - j = \ell$, we apply Lemma C.1.1 correspondingly. For A_1, let $t := s - k \in [s]$,

$$
\begin{aligned}
A_1 &= \sum_{\ell=1}^{s} \sum_{k=0}^{s-\ell} \sum_{\substack{T \in \mathcal{T}_{d,s-\ell} \\ |T \cap S_*| = k}} \mathbb{P}(\overline{\mathcal{E}(T, s-\ell)}) \\
&\leq s(s-\ell+1) \max_{\substack{1 \leq \ell \leq s \\ 0 \leq k \leq s-\ell}} \binom{d-s}{s-k}^2 \max_{\substack{T \in \mathcal{T}_{d,s-\ell} \\ |T \cap S_*| = k}} \mathbb{P}(\overline{\mathcal{E}(T, s-\ell)}) \\
&\leq s\bar{s} \max_{\substack{1 \leq \ell \leq s \\ 0 \leq k \leq s-\ell}} 5 \exp\Bigg(- (n-\bar{s}) \frac{\min\left((s-k)\omega\beta_{\min}^2/\sigma^2, 1\right)}{9216} \\
&\qquad\qquad + \frac{s-k-\ell}{4} + 2\log\binom{d-s}{s-k}\Bigg) \\
&\leq s\bar{s} \max_{\substack{1 \leq \ell \leq s \\ 0 \leq k \leq s-\ell}} 5 \exp\left(- (n-\bar{s}) \frac{\min\left((s-k)\omega\beta_{\min}^2/\sigma^2, 1\right)}{9216} + 3\log\binom{d-s}{s-k}\right) \\
&= s\bar{s} \max_{t \in [s]} 5 \exp\left(- (n-\bar{s}) \frac{\min\left(t\omega\beta_{\min}^2/\sigma^2, 1\right)}{9216} + 3\log\binom{d-s}{t}\right).
\end{aligned}
$$

For A_2, which is positive only when $s < \bar{s}$, let $t := s - k + \ell \in [\bar{s}]$,

$$
\begin{aligned}
A_2 &= \sum_{\ell=1}^{\bar{s}-s} \sum_{k=0}^{s} \sum_{\substack{T \in \mathcal{T}_{d,s+\ell} \\ |T \cap S_*| = k}} \mathbb{P}(\overline{\mathcal{E}(T, s+\ell)}) \\
&\le (\bar{s}-s)(s+1) \max_{\substack{1 \le \ell \le \bar{s}-s \\ 0 \le k \le s}} \binom{d-s}{s-k+\ell}^2 \max_{\substack{T \in \mathcal{T}_{d,s+\ell} \\ |T \cap S_*| = k}} \mathbb{P}(\overline{\mathcal{E}(T, s+\ell)}) \\
&\le (\bar{s}-s)\bar{s} \max_{\substack{1 \le \ell \le \bar{s}-s \\ 0 \le k \le s}} 5 \exp\Bigg(- (n-\bar{s}) \frac{\min\left((s-k+\ell)\omega\beta_{\min}^2/\sigma^2, 1\right)}{9216} \\
&\qquad\qquad + \frac{s-k+\ell}{4} + 2\log\binom{d-s}{s-k+\ell}\Bigg) \\
&\le (\bar{s}-s)\bar{s} \max_{\substack{1 \le \ell \le \bar{s}-s \\ 0 \le k \le s}} 5 \exp\Bigg(- (n-\bar{s}) \frac{\min\left((s-k+\ell)\omega\beta_{\min}^2/\sigma^2, 1\right)}{9216} \\
&\qquad\qquad + 3\log\binom{d-s}{s-k+\ell}\Bigg) \\
&= (\bar{s}-s)\bar{s} \max_{t \in [\bar{s}]} 5 \exp\left(- (n-\bar{s}) \frac{\min\left(t\omega\beta_{\min}^2/\sigma^2, 1\right)}{9216} + 3\log\binom{d-s}{t}\right).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
A_1 + A_2 &\le 5\bar{s}^2 \max_{t \in [\bar{s}]} \exp\left(- (n-\bar{s}) \frac{\min\left(t\omega\beta_{\min}^2/\sigma^2, 1\right)}{9216} + 3\log\binom{d-s}{t}\right) \\
&= \max_{t \in [\bar{s}]} \exp\left(- (n-\bar{s}) \frac{\min\left(t\omega\beta_{\min}^2/\sigma^2, 1\right)}{9216} + 3\log\binom{d-s}{t} + \log(5\bar{s}^2)\right).
\end{aligned}
$$

Since

$$\begin{aligned}\log(5\bar{s}^2) &= \log 5 + 2\log \bar{s} \\ &\leq \log 5 + 2\max_{t\in[\bar{s}]} \log \binom{\bar{s}}{t} \\ &\leq 3\max_{t\in[\bar{s}]} \log \binom{\bar{s}}{t} \\ &\leq 3\max_{t\in[\bar{s}]} \log \binom{d-s}{t},\end{aligned}$$

we have

$$A_1 + A_2 \leq \max_{t\in[\bar{s}]} \exp\left(-(n-\bar{s})\frac{\min\left(t\omega\beta_{\min}^2/\sigma^2, 1\right)}{9216} + 6\log\binom{d-s}{t}\right).$$

Combined with Theorem 4.2.2, we have following error probability,

$$\begin{aligned}\mathbb{P}(\widehat{S}^{\text{BSSu}} \neq S) &\leq 2\max_{t\in[\bar{s}]} \exp\left(-(n-\bar{s})\frac{\min\left(t\omega\beta_{\min}^2/\sigma^2, 1\right)}{9216} + 6\log\binom{d-s}{t}\right) \\ &\leq 2\max_{t\in[\bar{s}]} \exp\left(-(n-\bar{s})\frac{\min\left(t\omega\beta_{\min}^2/\sigma^2, 1\right)}{9216} + 6\log\binom{d}{t}\right).\end{aligned}$$

Setting the RHS to be smaller than δ leads to desired sample complexity. □

Proof of Lemma C.3.1. Let $|T| = j \leq \bar{s}$, then

$$\begin{aligned}\frac{\|\Pi_T^{\perp} Y\|^2/\sigma^2}{n-j} - \frac{\|\Pi_{S_*}^{\perp} Y\|^2/\sigma^2}{n-s} = &\frac{\|\Pi_T^{\perp} Y\|^2/\sigma^2 - \|\Pi_T^{\perp}\epsilon\|^2/\sigma^2}{n-j} \\ &+ \frac{\|\Pi_T^{\perp}\epsilon\|^2/\sigma^2}{n-j} - \frac{\|\Pi_{S_*}^{\perp}\epsilon\|^2/\sigma^2}{n-s}\end{aligned}$$

Similar to the proof of Lemma C.1.3, we can write

$$\Pi_T^{\perp} Y = \Pi_T^{\perp}(E_{S_*\setminus T}\beta_{S_*\setminus T} + \epsilon)\,,$$

where $E_{S_*\setminus T}\beta_{S_*\setminus T}$ is a random vector vector independent with T and each entry is i.i.d. from $\mathcal{N}(0, \Delta_1(S_*, T)\sigma^2)$. Therefore,

$$\begin{aligned}
\frac{\|\Pi_T^{\perp} Y\|^2/\sigma^2 - \|\Pi_T^{\perp}\epsilon\|^2/\sigma^2}{n-j} &= \frac{\|\Pi_T^{\perp} E_{S_*\setminus T}\beta_{S_*\setminus T}\|^2 + 2\langle \Pi_T^{\perp} E_{S_*\setminus T}\beta_{S_*\setminus T}, \epsilon\rangle}{(n-j)\sigma^2} \\
&= \frac{\Delta_1(S_*, T)\|\Pi_T^{\perp} U\|^2}{n-j} + \frac{2\sqrt{\Delta_1(S_*, T)}\langle \Pi_T^{\perp} U, \Pi_T^{\perp} U_\epsilon\rangle}{n-j} \\
&=: B_1 + B_2\,,
\end{aligned}$$

where $U, U_\epsilon \sim \mathcal{N}(0, I_n)$ and $U \perp\!\!\!\perp U_\epsilon$.

$$\begin{aligned}
&\frac{\|\Pi_T^{\perp}\epsilon\|^2/\sigma^2}{n-j} - \frac{\|\Pi_{S_*}^{\perp}\epsilon\|^2/\sigma^2}{n-s} \\
&= \frac{\|\Pi_T^{\perp}\epsilon\|^2/\sigma^2 - \|\Pi_{S_*}^{\perp}\epsilon\|^2/\sigma^2}{n-j} + \Big(\frac{1}{n-j} - \frac{1}{n-s}\Big)\|\Pi_{S_*}^{\perp}\epsilon\|^2/\sigma^2 \\
&= \frac{\|(\Pi_{S_*} - \Pi_{S_*\cap T})\epsilon\|^2/\sigma^2}{n-j} - \frac{\|(\Pi_T - \Pi_{S_*\cap T})\epsilon\|^2/\sigma^2}{n-j} \\
&+ \frac{j-s}{n-j} \times \frac{\|\Pi_{S_*}^{\perp}\epsilon\|^2/\sigma^2}{n-s} \\
&\geq -\frac{\|(\Pi_T - \Pi_{S_*\cap T})\epsilon\|^2/\sigma^2}{n-j} - \frac{s-j}{n-j} \times \frac{\|\Pi_{S_*}^{\perp}\epsilon\|^2/\sigma^2}{n-s} \\
&=: B_3 + B_4\,,
\end{aligned}$$

Recall our short hand notation $\nu = \omega\beta_{\min}^2/\sigma^2$, then

$$\begin{aligned}
&\mathbb{P}\Big[\frac{\|\Pi_T^{\perp} Y\|^2/\sigma^2}{n-|T|} - \frac{\|\Pi_{S_*}^{\perp} Y\|^2/\sigma^2}{n-s} \le \frac{1}{4}\Big(\Delta_1(S_*,T) - \ell'\nu\Big)\Big] \\
\le &\mathbb{P}\Big(B_1 \le \frac{1}{2}\Delta_1(S_*,T)\Big) + \sum_{k=2}^{4}\mathbb{P}\Big(B_k \le -\frac{1}{12}(\Delta_1(S_*,T) + \ell'\nu)\Big)\,.
\end{aligned}$$

We deal with these 4 error probabilities individually.

For B_1, analogous to the first part of proof of Lemma C.1.3, we can conclude

$$\mathbb{P}\Big(B_1 \le \frac{1}{2}\Delta_1(S_*,T)\Big) \le \exp(-(n-\bar{s})/64)\,.$$

For B_2, analogous to the second part of proof of Lemma C.1.3, we firstly condition on X_T, then $2\langle \Pi_T^{\perp} U, \Pi_T^{\perp} U_\epsilon\rangle = W - \widetilde{W}$ where $W, \widetilde{W} \sim \chi^2_{n-j}$. Thus,

$$\begin{aligned}
&\mathbb{P}\Big(B_2 \le -\frac{1}{12}(\Delta_1(S_*,T) + \ell'\nu)\Big) \\
= &\mathbb{P}\Big(\frac{W-\widetilde{W}}{n-j} \le -\frac{1}{12}\frac{\ell'\nu + \Delta_1(S_*,T)}{\sqrt{\Delta_1(S_*,T)}}\Big) \\
\le &2\mathbb{P}\Big(\frac{|W-(n-j)|}{n-j} \le \frac{1}{24}\frac{\ell'\nu + \Delta_1(S_*,T)}{\sqrt{\Delta_1(S_*,T)}}\Big) \\
\le &2\exp\Big(-(n-\bar{s})\min\Big(\frac{\ell'\nu + \Delta_1(S_*,T)}{\sqrt{\Delta_1(S_*,T)}}, \frac{(\ell'\nu + \Delta_1(S_*,T))^2}{\Delta_1(S_*,T)}\Big)/9216\Big) \\
\le &2\exp\Big(-(n-\bar{s})\min\Big(\ell'\nu + \Delta_1(S_*,T), \sqrt{\ell'\nu + \Delta_1(S_*,T)}\Big)/9216\Big)\,.
\end{aligned}$$

The last inequality is because

$$\begin{aligned}
\frac{(\ell'\nu + \Delta_1(S_*,T))^2}{\Delta_1(S_*,T)} &= \frac{(\ell'\nu)^2 + \Delta_1^2(S_*,T) + 2\ell'\nu\Delta_1(S_*,T)}{\Delta_1(S_*,T)} \\
&= \Delta_1(S_*,T) + 2\ell'\nu + \frac{(\ell'\nu)^2}{\Delta_1(S_*,T)} \\
&\ge \Delta_1(S_*,T) + \ell'\nu\,.
\end{aligned}$$

For B_3, we first condition on X_T and X_{S_*}, $\|(\Pi_T - \Pi_{S_* \cap T})\epsilon\|^2/\sigma^2 = Z \sim \chi^2_{|T \setminus S_*|}$. Then

$$\begin{aligned}
\mathbb{P}\Big(B_3 \le -\frac{1}{12}(\Delta_1(S_*, T) + \ell'\nu)\Big) &= \mathbb{P}\Big(\frac{Z}{n-j} \ge \frac{1}{12}(\Delta_1(S_*, T) + \ell'\nu)\Big) \\
&= \mathbb{P}\Big(\frac{\chi^2_{|T\setminus S_*|}}{|T \setminus S_*|} - 1 \ge \frac{(n-j)(\Delta_1(S_*, T) + \ell'\nu)}{12|T \setminus S_*|} - 1\Big) \\
&\le \exp\bigg(-(n-j)\frac{\Delta_1(S_*, T) + \ell'\nu}{48} + \frac{|T \setminus S_*|}{4}\bigg) \\
&\le \exp\bigg(-(n-\bar{s})\frac{(|S_* \setminus T| + \ell')\nu}{48} + \frac{|T \setminus S_*|}{4}\bigg).
\end{aligned}$$

The second to the last inequality holds when

$$\frac{(n-\bar{s})(\Delta_1(S_*, T) + \ell'\nu)}{48|T \setminus S_*|} - \frac{1}{4} \ge 1 \Leftarrow n - j \ge \frac{96|T \setminus S_*|}{(|S_* \setminus T| + \ell')\nu},$$

which is ensured by $n - \bar{s} \ge 96/\nu$ because $|S_* \setminus T| + \ell' \ge |T \setminus S_*|$ by definition of ℓ'.

For B_4, when $j > s$, $B_4 \ge 0$. When $s > j$, we first condition on X_{S_*}, $\|\Pi^{\perp}_{S_*}\epsilon\|^2/\sigma^2 = \widetilde{Z} \sim \chi^2_{n-s}$. Then

$$\begin{aligned}
\mathbb{P}\Big(B_4 \le -\frac{1}{12}(\Delta_1(S_*, T) + \ell'\nu)\Big) &= \mathbb{P}\bigg(\frac{\widetilde{Z}}{n-s} \ge \frac{(n-j)(\Delta_1(S_*, T) + \ell'\nu)}{12(s-j)}\bigg) \\
&= \mathbb{P}\bigg(\frac{\chi^2_{n-s}}{n-s} - 1 \ge \frac{(n-j)(\Delta_1(S_*, T) + \ell'\nu)}{12(s-j)} - 1\bigg) \\
&\le \exp\bigg(-(n-s)\Big(\frac{\Delta_1(S_*, T) + \ell'\nu}{48} \times \frac{n-j}{s-j} - \frac{1}{4}\Big)\bigg) \\
&\le \exp\bigg(-(n-s)\frac{\Delta_1(S_*, T) + \ell'\nu}{48}\bigg) \\
&\le \exp\bigg(-(n-\bar{s})\frac{(|S_* \setminus T| + \ell')\nu}{48}\bigg).
\end{aligned}$$

The first inequality holds when

$$\frac{(n-j)(\Delta_1(S_*, T) + \ell'\nu)}{48(s-j)} - \frac{1}{4} \ge 1 \Leftarrow n - j \ge \frac{96(s-j)}{(|S_* \setminus T| + \ell')\nu},$$

which is ensure by $n - \bar{s} \geq 96/\nu$ since $|S_* \setminus T| \geq s - j$. The second inequality holds because

$$\begin{aligned}\frac{(n-j)(\Delta_1(S_*, T) + \ell'\nu)}{48(s-j)} - \frac{1}{4} &\geq \frac{(|S_* \setminus T| + \ell')\nu}{48} \times \frac{n-j}{s-j} - \frac{1}{4} \\ &= \frac{(|S_* \setminus T| + \ell')\nu}{48}\left(\frac{n-j}{s-j} - \frac{12}{(|S_* \setminus T| + \ell')\nu}\right) \\ &\geq \frac{(|S_* \setminus T| + \ell')\nu}{48}.\end{aligned}$$

The last inequality in equation above holds when

$$\frac{n-j}{s-j} - \frac{12}{(|S_* \setminus T| + \ell')\nu} \geq 1 \Leftrightarrow n - s \geq \frac{12(s-j)}{(|S_* \setminus T| + \ell')\nu},$$

which is ensured by $n - \bar{s} \geq 96/\nu$.

Finally, combining these error probability bounds, we conclude

$$\begin{aligned}&\mathbb{P}\left[\frac{\|\Pi_T^{\perp} Y\|^2/\sigma^2}{n - |T|} - \frac{\|\Pi_{S_*}^{\perp} Y\|^2/\sigma^2}{n-s} \leq \frac{1}{4}\Big(\Delta_1(S_*, T) - \ell'\nu\Big)\right] \\ \leq &\mathbb{P}\Big(B_1 \leq \frac{1}{2}\Delta_1(S_*, T)\Big) + \sum_{k=2}^{4} \mathbb{P}\Big(B_k \leq -\frac{1}{12}(\Delta_1(S_*, T) + \ell'\nu)\Big) \\ \leq &5\exp\left(-(n-\bar{s})\frac{\min\Big((|S_* \setminus T| + \ell')\nu, 1\Big)}{9216} + \frac{|T \setminus S_*|}{4}\right).\end{aligned}$$

□

C.4 Proof of Theorem 4.2.6 (Variable Selection Lower Bound)

Proof. Again, we consider the covariance matrix of X:

$$\Sigma = (1-\rho) I_d + \rho \mathbf{1}_d \mathbf{1}_d^{\top}$$

with $\rho = 1 - \omega$. Then for any $T \in \mathcal{T}_d^{\bar{s}}$ with $|T| = j$, and $|S \setminus T| = r$, we can calculate the conditional covariance matrix

$$\begin{aligned}\Sigma_{S\setminus T\,|\,T} &= \Sigma_{(S\setminus T)(S\setminus T)} - \Sigma_{(S\setminus T)T}\Sigma_{TT}^{-1}\Sigma_{T(S\setminus T)} \\ &= (1-\rho)\left(I_r + \frac{\rho}{1-\rho+j\rho}\mathbf{1}_r\mathbf{1}_r^\top\right),\end{aligned}$$

then the minimum eigenvalue is

$$\lambda_{\min}(\Sigma_{S\setminus T\,|\,T}) = \begin{cases} (1-\rho)\times(1+\frac{\rho}{1-\rho+j\rho}) & r=1 \\ 1-\rho & r\geq 2 \end{cases}.$$

Since $\lambda_{\min}(\Sigma_{S\setminus T\,|\,T})$ is independent with the choice of (S,T), this covariance matrix Σ satisfies the requirement on $\Omega_d^{\bar{s}}(\omega)$:

$$\min_{S\in\mathcal{T}_d^{\bar{s}}}\min_{\substack{T\in\mathcal{T}_d^{\bar{s}}\setminus\{S\}\\ T\not\supseteq S}} \lambda_{\min}(\Sigma_{S\setminus T\,|\,T}) = 1-\rho = \omega\,.$$

Note that we only take $T \not\supseteq S$ to make sure $r \geq 1$. Now we fix the covariance matrix Σ, consider the ensemble with support size one: Each integer $k \in [d]$ determines a model $Y = X_k\beta_{\min} + \epsilon$. Thus the cardinality of this model ensemble is $|\binom{d}{1}| = d$. Now we calculate the KL divergence between two models specified by k and j. We further denote

and the models determined by them to be P_k and P_j. Therefore,

$$\begin{aligned}\mathbf{KL}(P_k \| P_j) &= \mathbb{E}_{P_k} \log \frac{P_k}{P_j} \\ &= \mathbb{E}_X (X_k - X_j)^2 \beta_{\min}^2 / 2\sigma^2 \\ &= 2(1-\rho) \times \frac{\beta_{\min}^2}{2\sigma^2} \\ &= \frac{\omega \beta_{\min}^2}{\sigma^2}\end{aligned}$$

Thus $\mathbf{KL}(P_k \| P_j) \leq \beta_{\min}^2 \omega / \sigma^2$, which holds for any pair of $j, k \in [d]$ and leads to a upper bound for any two models in this ensemble. Finally, we apply Fano's inequality Corollary D.2.3 with KL divergence upper bound $\beta_{\min}^2 \omega / \sigma^2$ and ensemble cardinality d, which completes the proof. □

C.5 Neighbourhood Selection and Variable Selection

For completeness, we recall the connection between neighbourhood selection in SEM and variable selection in linear models here. Recall the definition of the Markov boundary of X_k with respect to a set $A \subseteq V$ is the smallest subset $S \subseteq A$ such that $X_k \perp\!\!\!\perp X_{A \setminus S} \mid X_S$ (i.e. the smallest Markov blanket of X_k in A), denoted by $S(k; A)$. The equivalence of the two problems is given in Proposition C.5.1 below.

Proposition C.5.1. *If $(X_k, X_A) \sim \mathcal{N}(\mathbf{0}, \Gamma)$ and Γ is a positive definite covariance matrix, then for any subset $S \subseteq A$, the following are equivalent:*

1. *$S = S(k; A)$, where $S(k; A)$ is the Markov boundary defined above;*
2. *We have a linear model $X_k = \beta^\top X_A + \epsilon$ with $\epsilon \perp\!\!\!\perp X_A$, $\mathbb{E}[\epsilon] = 0$, and $\operatorname{supp}(\beta) = S$.*

Moreover, suppose $p(X)$ is an SEM by (1.1) *with $b_{jk} \neq 0, \forall j \in pa(k)$. If $A = nd(k)$, then $S(k; A) = pa(k) = \operatorname{supp}(\beta)$.*

Therefore, Markov boundary learning reduces to a regression problem between X_k and X_A More specifically, the goal is to recover the support set (i.e. nonzero entries) for this regression problem, i.e. variable selection. Thus, in our setting we have:

$$\begin{aligned} \text{neighbourhood selection} &\iff \text{Markov boundary learning} \\ &\iff \text{variable selection in regression.} \end{aligned}$$

Proof of Proposition C.5.1. The Markov blanket of k with respect to subset $A \subseteq X_{-k}$ is a subset $T \subseteq A$ such that $X_k \perp\!\!\!\perp (A \setminus T) \,|\, T$. The Markov boundary is the smallest Markov blanket. In the following proof, we will use regression notation by taking $Y := X_k, X := X_A$.

1)⇒ 2): Denote $A \setminus S = S^c$. We can write

$$Y = \Gamma_{YS}\Gamma_S^{-1}X_S + [Y - \Gamma_{YS}\Gamma_S^{-1}X_S] := \beta_S^\top X_S + \epsilon = \beta^\top X + \epsilon\,,$$

where β is a vector such that $\beta_S := \Gamma_{YS}\Gamma_S^{-1}$, $\beta_j = 0$ if $j \in S^c$. Moreover, $\mathbb{E}[\epsilon] = 0 - \beta_S^\top 0 = 0$. Now we want to show $\epsilon \perp\!\!\!\perp S$ and $\epsilon \perp\!\!\!\perp S^c$. By Gaussianity, it suffices to deal with covariance. For the first one,

$$\begin{aligned} \mathbb{E}[X_S\epsilon] &= \mathbb{E}[X_S Y] + \mathbb{E}[X_S X_S^\top \beta_S] \\ &= \Gamma_{SY} + \Gamma_S\beta_S \\ &= \Gamma_{SY} + \Gamma_S\Gamma_S^{-1}\Gamma_{SY} \\ &= 0\,. \end{aligned}$$

For the second one, since $Y \perp\!\!\!\perp S^c \mid S$, i.e.

$$\begin{aligned}
0 &= \mathbb{E}_S \operatorname{cov}(Y, S^c \mid S) \\
&= \mathbb{E}(X_{S^c} Y) - \mathbb{E}_S[\mathbb{E}(X_{S^c} \mid S)\mathbb{E}(Y \mid S)] \\
&= \mathbb{E}(X_{S^c} Y) - \mathbb{E}_S[\mathbb{E}(X_{S^c} \mid S) X_S^\top \beta_S] \\
&= \mathbb{E}(X_{S^c} Y) - \mathbb{E}(X_{S^c} X_S^\top \beta_S] \\
&= \mathbb{E}[X_{S^c} \epsilon] \,.
\end{aligned}$$

The second equality is by definition of conditional covariance; the third equality is by independence between ϵ and S; the fourth equality is by tower rule; the last equality is by definition of ϵ. Finally, we want to show $\beta_j \neq 0, \forall j \in S$. By way of contradiction, suppose $\beta_j = 0$, denote $R = S \setminus \{j\}$, then $Y = \beta_R^\top X_R + \epsilon$. Notice that

$$\begin{aligned}
&\mathbb{E}_R \operatorname{cov}(Y, S^c \cup \{j\} \mid R) \\
&= \left(\mathbb{E}[Y X_j] - \mathbb{E}_R[\mathbb{E}(Y \mid R)\mathbb{E}(X_j \mid R)], \mathbb{E}[Y X_{S^c}] - \mathbb{E}_R[\mathbb{E}(Y \mid R)\mathbb{E}(X_{S^c} \mid R)] \right)^\top \\
&= \left(\mathbb{E}[X_j X_R^\top]\beta_R - \mathbb{E}_R[X_R^\top \beta_R \mathbb{E}(X_j \mid R)], \mathbb{E}[X_{S^c} X_R^\top]\beta_R - \mathbb{E}_R[X_R^\top \beta_R \mathbb{E}(X_{S^c} \mid R)] \right)^\top \\
&= (0, 0)^\top \,.
\end{aligned}$$

Therefore, $Y \perp\!\!\!\perp S^c \cup \{j\} \mid R$, i.e. R is a Markov blanket of Y in X, which contradicts the minimality of S. This completes the first half of the proof.

2)⇒ 1): We can compactly write $Y = \beta_S^\top X_S + \epsilon$. For S to be $S(Y; X)$, we need to check

whether $Y \perp\!\!\!\perp S^c \mid S$ and S is the minimal subset satisfies it. For the first one,

$$\begin{aligned}
\mathbb{E}_S \operatorname{cov}(Y, S^c \mid S) &= \mathbb{E}[X_{S^c} Y] - \mathbb{E}_S[\mathbb{E}(X_{S^c} \mid S)\mathbb{E}(Y \mid S)] \\
&= \mathbb{E}[X_{S^c} X_S^\top \beta_S] - \mathbb{E}_S[\mathbb{E}(X_{S^c} \mid S) X_S^\top \beta_S] \\
&= \mathbb{E}[X_{S^c} X_S^\top \beta_S] - \mathbb{E}[X_{S^c} X_S^\top \beta_S] \\
&= 0 \,.
\end{aligned}$$

The second equality is by definition of linear model. Now for the second one, by way of contradiction, suppose there is a set $T \subseteq X$ with $|T| < |S|$ such that $Y \perp\!\!\!\perp T^c \mid T$. Let $S = (S \cap T) \cup (S \setminus T) := S_1 \cup S_2$. Thus $Y \perp\!\!\!\perp S_2 \mid T$, i.e.

$$\begin{aligned}
0 &= \mathbb{E}_T \operatorname{cov}(Y, X_{S_2} \mid T) \\
&= \mathbb{E}[Y X_{S_2}] - \mathbb{E}_T[\mathbb{E}(Y \mid T)\mathbb{E}(X_{S_2} \mid T)] \\
&= \mathbb{E}[X_{S_2} X_{S_1}^\top]\beta_{S_1} + \mathbb{E}[X_{S_2} X_{S_2}^\top]\beta_{S_2} - \mathbb{E}_T[(X_{S_1}^\top \beta_{S_1} + \mathbb{E}(X_{S_2}^\top \mid T)\beta_{S_2})\mathbb{E}(X_{S_2} \mid T)] \\
&= \Gamma_{S_2}\beta_{S_2} - \mathbb{E}_T[\mathbb{E}(X_{S_2} \mid T)\mathbb{E}(X_{S_2}^\top \mid T)]\beta_{S_2} \\
&= (\Gamma_{S_2} - \Gamma_{S_2 T}\Gamma_T^{-1}\Gamma_{S_2 T})\beta_{S_2} \,,
\end{aligned}$$

Then $\beta_{S_2}^\top(\Gamma_{S_2} - \Gamma_{S_2 T}\Gamma_T^{-1}\Gamma_{S_2 T})\beta_{S_2} = 0$, since $\beta_{S_2} \neq 0$, which contradicts the assumption that Γ is positive definite, because the principal submatrix and Schur complement of a positive definite matrix are still positive definite.

Finally, if $P(Z)$ is an SEM generated by a DAG G, and $A = \mathrm{nd}(k)$ is the nondescendants, by Markov property, we have $X_k \perp\!\!\!\perp X_{A \setminus \mathrm{pa}(k)} \mid \mathrm{pa}(k)$. Thus $\mathrm{pa}(k)$ is a Markov blanket. Because $b_{jk} \neq 0$ for all $j \in \mathrm{pa}(k)$, then $\mathrm{pa}(k)$ is the Markov boundary. □

C.6 Proof of Lemma 4.2.7

Proof. Given any polynomial time support estimator $\widehat{S} = \widehat{S}(X, Y)$, we construct an estimator for β vector as follows:

1. Split the data (Y, X) into two folds with equal size $(Y^{(1)}, X^{(1)})$ and $(Y^{(2)}, X^{(2)})$;
2. Estimate support using the first fold $\widehat{S} = \widehat{S}(Y^{(1)}, X^{(1)})$;
3. Estimate the β vector by

$$\widehat{\beta} = \begin{pmatrix} \widehat{\beta}_{\widehat{S}} \\ \widehat{\beta}_{\widehat{S}^c} \end{pmatrix} = \begin{pmatrix} ({X^{(2)}}^\top X^{(2)})^{-1} {X^{(2)}}^\top Y^{(2)} \\ \mathbf{0}_{d-s} \end{pmatrix}.$$

Therefore, $\widehat{\beta}$ is a polynomial time estimator for β. We are going to employ the construction in the following lemma.

Lemma C.6.1 (Theorem 1, Zhang et al. [2014])**.** *If* $\mathbf{NP} \not\subset \mathbf{P}\backslash\mathbf{poly}$*, then for any* $\delta \in (0, 1)$*, any* $b \in \mathbb{Z}_+$*, any polynomial functions* $G : (\mathbb{Z}_+)^3 \to \mathbb{R}_+$ *and* $F, H : \mathbb{Z}_+ \to \mathbb{R}_+$*, there exists a sparsity level* $s \geq 1$ *such that for any* $d \in [4s, F(s)]$*,* $n' \in [c_1 s \log d, F(s)]$*, and* $\gamma \in [2^{-G(n,d,s)}, 1/24\sqrt{2})$*, there exists a design matrix* $\widetilde{X} \in \mathbb{R}^{n' \times d}$ *such that:*

1. *The RE constant* $|\gamma(\widetilde{X}) - \gamma| \leq 2^{-G(n,d,s)}$*;*
2. *For any* (b, G, H)*-efficient estimator* $\widehat{\beta}$ *with knowledge of* s*, the mean-squared prediction risk is lower bounded as*

$$\max_{\beta \in \Theta_{d,s}} \mathbb{E} \frac{\|\widetilde{X}(\widehat{\beta} - \beta)\|^2}{n'} \geq \frac{c_2}{\gamma^2} \frac{\sigma^2 s^{1-\delta} \log d}{n'}.$$

With the same δ, polynomial functions F, G, H stated in the theorem, there exists sparsity level $s \geq 1$ such that for the $d, n/2, \gamma$ satisfying the requirement, we have $\widetilde{X} \in$

$\mathbb{R}^{(n/2)\times d}$ such that $|\gamma(\widetilde{X}) - \gamma| \le 2^{-G(n,d,s)}$ and for any (b, G, H)-efficient estimator $\widetilde{\beta}$ with knowledge of s,

$$\max_{\beta\in\Theta_{d,s}} \mathbb{E}\frac{\|\widetilde{X}(\widetilde{\beta}-\beta)\|^2}{n/2} \ge \frac{C'}{\gamma^2}\frac{\sigma^2 s^{1-\delta}\log d}{n/2}.$$

We now construct $X = (\widetilde{X}^\top, \widetilde{X}^\top)^\top$ by stacking two copies of $\widetilde{X}$ and take the corresponding maximizer β to form $Y = (Y^{(1)}, Y^{(2)}) = X\beta + \epsilon$ with $\epsilon = (\epsilon^{(1)}, \epsilon^{(2)})$. Note that $\gamma(X) = \gamma(\widetilde{X})$ by definition, thus $|\gamma(X) - \gamma| = |\gamma(\widetilde{X}) - \gamma| \le 2^{-G(s)}$.

We then analyze the property of $\widehat{\beta}$ on this construction. Note that $\epsilon^{(1)} \perp\!\!\!\perp \epsilon^{(2)}$, and $\widehat{S}$ only depends on $\epsilon^{(1)}$ via $Y^{(1)}$, thus $\widehat{S} \perp\!\!\!\perp \epsilon^{(2)}$. Denote that $\widetilde{\Pi}_T = \widetilde{X}(\widetilde{X}^\top\widetilde{X})^{-1}\widetilde{X}^\top$, and $\widetilde{\Pi}_T^\perp = I_{n/2} - \widetilde{\Pi}_T$. We have

$$\begin{aligned}
&\frac{C'}{\gamma^2}\frac{\sigma^2 s^{1-\delta}\log d}{n/2} \\
&\le \mathbb{E}\frac{\|X(\widehat{\beta}-\beta)\|^2}{n} = \mathbb{E}\frac{\|\widetilde{X}(\widehat{\beta}-\beta)\|^2}{n/2} \\
&= \frac{1}{n/2}\mathbb{E}\left\|\widetilde{X}_{\widehat{S}}(\widetilde{X}_{\widehat{S}}^\top\widetilde{X}_{\widehat{S}})^{-1}\widetilde{X}_{\widehat{S}}^\top(\widetilde{X}_{S_*}\beta_{S_*} + \epsilon^{(2)}) - \widetilde{X}_{S_*}\beta_{S_*}\right\|^2 \\
&= \frac{1}{n/2}\mathbb{E}\left[\left\|\widetilde{X}_{\widehat{S}}(\widetilde{X}_{\widehat{S}}^\top\widetilde{X}_{\widehat{S}})^{-1}\widetilde{X}_{\widehat{S}}^\top(\widetilde{X}_{S_*}\beta_{S_*} + \epsilon^{(2)}) - \widetilde{X}_{S_*}\beta_{S_*}\right\|^2 \mid \widehat{S} = S_*\right]\mathbb{P}(\widehat{S} = S_*) \\
&\quad + \frac{1}{n/2}\mathbb{E}\left[\left\|\widetilde{X}_{\widehat{S}}(\widetilde{X}_{\widehat{S}}^\top\widetilde{X}_{\widehat{S}})^{-1}\widetilde{X}_{\widehat{S}}^\top(\widetilde{X}_{S_*}\beta_{S_*} + \epsilon^{(2)}) - \widetilde{X}_{S_*}\beta_{S_*}\right\|^2 \mid \widehat{S} \ne S_*\right]\mathbb{P}(\widehat{S} \ne S_*) \\
&= \frac{1}{n/2}\mathbb{E}\left[\left\|\widetilde{\Pi}_{S_*}\epsilon^{(2)}\right\|^2 \mid \widehat{S} = S_*\right]\mathbb{P}(\widehat{S} = S_*) \\
&\quad + \frac{1}{n/2}\mathbb{E}\left[\left\|\widetilde{\Pi}_{S_*}\epsilon^{(2)} - \widetilde{\Pi}_{\widehat{S}}^\perp\widetilde{X}_{S_*\setminus\widehat{S}}\beta_{S_*\setminus\widehat{S}}\right\|^2 \mid \widehat{S} \ne S_*\right]\mathbb{P}(\widehat{S} \ne S_*) \\
&= \frac{s\sigma^2}{n/2}\mathbb{P}(\widehat{S} = S_*) + \mathbb{E}\left[\frac{s\sigma^2}{n/2} + \frac{\|\widetilde{\Pi}_{\widehat{S}}^\perp\widetilde{X}_{S_*\setminus\widehat{S}}\beta_{S_*\setminus\widehat{S}}\|^2}{n/2} \mid \widehat{S} \ne S_*\right]\mathbb{P}(\widehat{S} \ne S_*) \\
&\le 2\times\frac{s\sigma^2}{n/2} + \mathbb{P}(\widehat{S} \ne S_*)\times\max_{T\ne S_*}\frac{\|\widetilde{\Pi}_T^\perp\widetilde{X}_{S_*\setminus T}\beta_{S_*\setminus T}\|^2}{n/2} \\
&= 2\times\frac{s\sigma^2}{n/2} + \mathbb{P}(\widehat{S} \ne S_*)\times\max_{T\ne S_*}\frac{\|\Pi_T^\perp X_{S_*\setminus T}\beta_{S_*\setminus T}\|^2}{n}.
\end{aligned} \tag{C.1}$$

Recall that $\Pi_T^{\perp} = I_n - X_T(X_T^\top X_T)^{-1}X_T^\top$. Since $\gamma < s^{-\delta/2}$, then $\frac{1}{\gamma^2} > s^{\delta}$, and

$$\frac{C'}{\gamma^2}\frac{\sigma^2 s^{1-\delta}\log d}{n/2} > \frac{C'\sigma^2 s\log d}{n/2} \gtrsim \frac{2\sigma^2 s}{n/2},$$

for sufficient large d. Therefore, for the inequality (C.1) to hold, we must have

$$\mathbb{P}(\widehat{S} \neq S_*) \times \max_{T\neq S_*} \frac{\|\Pi_T^{\perp} X_{S_*\setminus T}\beta_{S_*\setminus T}\|^2}{n} \geq \frac{C_2}{\gamma^2}\frac{\sigma^2 s^{1-\delta}\log d}{n/2}$$

for some constant C_2. Moving the signal term to the right hand side completes the proof.

□

C.7 Discussion of KL-BSS

This appendix starts by introducing the signals of KL-BSS and BSS, which will be useful in the analysis and proofs (Appendix C.7.1). After this, we collect miscellaneous (optional) discussions for interested readers: analyzing Vanilla KL-BSS, which was introduced for computational reasons (Appendix C.7.2), and non-Gaussian designs (Appendix C.7.3).

C.7.1 Signal and Analysis of KL-BSS in General Variable Selection

We introduce the signals used in the analysis, which explicitly illustrate the deficiency in BSS: There is an additional signal component that is being ignored. To see this, let us first define the signals for distinguishing two supports S and T:

Definition 14. For any $(\beta, \Sigma, \sigma^2) \in \mathcal{M}(\Theta, \Omega, \sigma^2)$, and any two sets $S, T \subseteq [d]$, define

$$\begin{aligned}
\Delta_1(S,T) &:= \frac{1}{\sigma^2}\beta_{S\setminus T}^\top \Sigma_{S\setminus T\,|\,T}\beta_{S\setminus T}, \\
\Delta_2(S,T) &:= \frac{1}{\sigma^2}\min_{\alpha\in\Theta_{T\setminus S}} \Big(\alpha_\beta(S,T) - \alpha\Big)^\top \Sigma_{T\setminus S\,|\,S\cap T}\Big(\alpha_\beta(S,T) - \alpha\Big),
\end{aligned} \tag{C.2}$$

where $\alpha_\beta(S,T) := \Sigma^{-1}_{T\setminus S \,|\, S\cap T}\Sigma_{(T\setminus S)(S\setminus T)\,|\, S\cap T}\beta_{S\setminus T}$ is the partial regression coefficients of $X^\top_{S_*\setminus T}\beta_{S_*\setminus T}$ onto $X_{T\setminus S_*}$.

Although both Δ_1 and Δ_2 depend on the parameters (β,Σ,σ^2), we omit them in the arguments for brevity. Δ_1 is the variance contributed by S_* that is not captured by T, while Δ_2 characterizes the violation of α_β to the beta-min condition. Therefore, Algorithm 3 aims to estimate Δ_1 and Δ_2 by their sample counterparts, while BSS only estimates Δ_1 and ignores the signal conveyed by Δ_2 entirely. By contrast, KL-BSS adapts to both situations where either Δ_1 or Δ_2 is larger.

The relation with the eigenvalues $\lambda_K(\Sigma)$ and $\lambda_B(\Sigma)$ that we focus in the main paper is the latter provide lower bounds on the signals and hence the sample complexities ultimately obtained in Theorem 4.3.3:

$$\begin{aligned}\Delta_1(S_*,T)+\Delta_2(S_*,T) &\geq |S_*\setminus T|\times\beta^2_{\min}\lambda_K(\Sigma)/\sigma^2\\ \Delta_1(S_*,T) &\geq |S_*\setminus T|\times\beta^2_{\min}\lambda_B(\Sigma)/\sigma^2\,.\end{aligned} \tag{C.3}$$

Notice that Δ_2 is large when entries in α_β are close to zero and thus fall outside of Θ. This property that KL-BSS is better at distinguishing alternatives with small regression coefficients is particularly useful in the context of SEM, where the information flows in one direction and accumulates at near-sink nodes. By the definition of α_β, an alternative containing these nodes can lead to small regression coefficients.

Define the (global) signal to be

$$\Delta(\mathcal{M}) := \min_{(\beta,\Sigma,\sigma^2)\in\mathcal{M}}\ \min_{T\in\mathcal{T}_{d,s}\setminus\{S_*\}}\frac{1}{|S_*\setminus T|}\Big(\Delta_1(S_*,T)\vee\Delta_2(S_*,T)\Big)\,. \tag{C.4}$$

For comparison, note that

$$\Delta_1(\mathcal{M}) := \min_{(\beta,\Sigma,\sigma^2)\in\mathcal{M}}\ \min_{T\in\mathcal{T}_{d,s}\setminus\{S_*\}}\frac{1}{|S_*\setminus T|}\Delta_1(S_*,T) \tag{C.5}$$

is the signal for BSS, which is similarly defined in Wainwright [2009a].

We can now state the sample complexity result for KL-BSS below, which will be the main technical machinery used for analyzing the performance of KL-BSS:

Theorem C.7.1. *Assume $s \leq d/2$ and let $(\beta, \Sigma, \sigma^2) \in \mathcal{M} := \mathcal{M}(\Theta, \Omega, \sigma^2)$. Given n i.i.d. samples from $P_{\beta,\Sigma,\sigma^2}$, if $\Delta(\mathcal{M}) > 0$ and the sample size satisfies*

$$\begin{aligned} n - s &\gtrsim \max_{r \in [s]} \frac{\log \binom{d-s}{r} + \log(1/\delta)}{r\Delta(\mathcal{M}) \wedge 1} \\ &\asymp \max\left\{ \frac{\log\left(d-s\right) + \log(1/\delta)}{\Delta(\mathcal{M})}, \log \binom{d-s}{s} + \log(1/\delta) \right\}, \end{aligned} \tag{C.6}$$

then $\mathbb{P}_{\beta,\Sigma,\sigma^2}(\widehat{S} = S_*) \geq 1 - \delta$, where $\widehat{S}$ is given by Algorithm 4.

The detailed proof is postponed to Appendix C.12. For comparison, the sample complexity of BSS [adapted to our setting from Wainwright, 2009a] is

$$\frac{\log(d-s)}{\Delta_1(\mathcal{M})} \vee \log \binom{d-s}{s}. \tag{C.7}$$

Obviously, $\Delta_1(\mathcal{M}) \leq \Delta(\mathcal{M})$, i.e. $\Delta(\mathcal{M})$ captures at least as much signal as BSS.

Remark 12. A pointwise version of Theorem C.7.1 (for any fixed $(\beta, \Sigma, \sigma^2) \in \mathcal{M}$) also holds with $\Delta(\beta, \Sigma, \sigma^2) := \min_{T \in \mathcal{T}_{d,s} \setminus \{S_*\}} (\Delta_1(S_*, T) \vee \Delta_2(S_*, T))/|S_* \setminus T|$ in place of $\Delta(\mathcal{M})$, i.e. without the first minimization in (C.4). This is clear from the proof of Theorem C.7.1, whose analysis is uniform for all $(\beta, \Sigma, \sigma^2) \in \mathcal{M}$.

Remark 13. Instead of using $\alpha_\beta(S, T)$, one could exploit the regression vector of the whole alternative support T without the partialing out step. However, it would be less sample efficient to estimate the relatively small extra signal, which is based on the intuition that the coefficients of $X_{S \cap T}$ barely violate the beta-min condition. This leads to Vanilla KL-BSS (Section 4.3.3).

We end this section by characterizing the signal to distinguish any two supports S and T. Lemma C.7.2 below, whose proof is in Appendix C.14, implicitly supports the idea that it should be easier to distinguish S and T when the discrepancy between them is larger. It shows the signal $\Delta_1(S,T) \vee \Delta_2(S,T)$ is the same order as the conditional variance of a linear combination of $|S \setminus T| + |T \setminus S|$ many random variables. This validates the $|S_* \setminus T|$ scaling factor in the denominator in our definition of the global signal (C.4).

Lemma C.7.2. *For any $(\beta, \Sigma, \sigma^2) \in \mathcal{M}(\Theta, \Omega, \sigma^2)$, and any two sets $S, T \in \mathcal{T}_{d,s}$,*

$$\Delta_1(S,T) \vee \Delta_2(S,T) \asymp \frac{1}{\sigma^2} \min_{\alpha_{T\setminus S} \in \Theta_{T\setminus S}} \operatorname{var}\left[X_{S\setminus T}^\top \beta_{S\setminus T} - X_{T\setminus S}^\top \alpha_{T\setminus S} \,|\, S \cap T \right].$$

Moreover, the constant is within $[1/2, 1]$.

C.7.2 Vanilla KL-BSS

In this appendix, we re-visit Vanilla KL-BSS, which was introduced in Section 4.3.3 to reformulate KL-BSS into an MIP for faster computation. Vanilla KL-BSS is simpler and more natural to exploit a larger signal $\widetilde{\Delta}_2(S,T)$ (defined below) compared to $\Delta_2(S,T)$ used by KL-BSS. However, it leads to an extra factor of s that KL-BSS avoids.

Definition 15. For any $(\beta, \Sigma, \sigma^2) \in \mathcal{M}(\Theta, \Omega, \sigma^2)$, and any two sets $S, T \subseteq [d]$, denote

$$\widetilde{\Delta}_2(S,T) := \frac{1}{\sigma^2} \min_{\alpha \in \Theta_T} \left(\widetilde{\alpha}_\beta(S,T) - \alpha \right)^\top \Sigma_{TT} \left(\widetilde{\alpha}_\beta(S,T) - \alpha \right), \tag{C.8}$$

where $\widetilde{\alpha}_\beta(S,T) := \Sigma_{TT}^{-1} \Sigma_{TS} \beta_S$.

Similarly, $\widetilde{\alpha}_\beta(S,T)$ is the coefficient vector of regressing $X_S^\top \beta_S$ onto X_T and $\widetilde{\Delta}_2(S,T)$ characterizes the violation to the constrained space Θ for T as a whole. Following the same strategy of KL-BSS, we compare the candidate supports using residual variances plus the

Algorithm 7 Algorithm for two candidate case

Input: Data matrix X; response Y; candidate supports $S, T \in \mathcal{T}_{d,s}$; coefficient space Θ.
Output: Estimated support $\widehat{S}$.

1. For $R = S$ or T:
   (a) Compute $\widehat{\gamma} = (X_R^\top X_R)^{-1} X_R^\top Y$;
   (b) Compute $\mathcal{L}(R) = \frac{\|\Pi_R^\perp Y\|^2}{n-s} + \min_{\gamma \in \Theta_R} (\widehat{\gamma} - \gamma)^\top \frac{X_R^\top X_R}{n} (\widehat{\gamma} - \gamma)$;
2. Output $\widehat{S} = \arg\min_{R \in \{S,T\}} \mathcal{L}(R)$.

sample counterpart of $\widetilde{\Delta}_2(S,T)$ (as opposed to $\Delta_2(S,T)$). For completeness, the resulting algorithm is shown in Algorithm 7. Analysis of this procedure leads to the following sample complexity for distinguishing the true support S_* against any other alternative T:

Lemma C.7.3. *For any $(\beta, \Sigma, \sigma^2) \in \mathcal{M}(\Theta, \Omega, \sigma^2)$, let $S_* = \mathrm{supp}(\beta)$ and $|S_*| = s$. Given i.i.d. data $(X, Y) \sim P_{\beta,\Sigma,\sigma^2}$, apply Algorithm 7 to estimate support from S_* and T with output $\widehat{S}$. Let $\Delta_1 := \Delta_1(S_*, T)$ and $\widetilde{\Delta}_2 := \widetilde{\Delta}_2(S_*, T)$. If sample size $n \gtrsim s + \frac{s}{\Delta_1 \vee \widetilde{\Delta}_2}$, we have for some constant C_0,*

$$\mathbb{P}_{\beta,\Sigma,\sigma^2}(\widehat{S} = S_*) \gtrsim 1 - 9\exp\left(- C_0(n-s) \min\left(\Delta_1 \vee \widetilde{\Delta}_2, 1\right) + s \right).$$

Since Algorithm 7 can be viewed as a special case of Algorithm 3 when $S \cap T = \emptyset$, we omit the proof. The difference between Algorithms 3 and 7 is also revealed in the error probability (Lemma C.12.1 vs. Lemma C.7.3): Since the calculation in the second term of $\mathcal{L}$ in Algorithm 7 involves estimating an s-dimensional covariance matrix, Lemma C.7.3 has an additional dependence on s (compared to $r = |S_* \setminus T|$ in Lemma C.12.1), but enjoys a larger signal due to $\Delta_2(S_*, T) \le \widetilde{\Delta}_2(S_*, T)$. This leads to a tradeoff between s and $\widetilde{\Delta}_2(S_*, T)$ that makes Vanilla KL-BSS slightly suboptimal in the worst-case, although our experiments show that it actually outperforms KL-BSS on average.

A straightforward application of Algorithm 7 leads to Algorithm 8, which yields Vanilla

Algorithm 8 Vanilla KL-BSS

Input: Data matrix X; response Y; coefficient space Θ; sparsity s.
Output: Estimated support $\widehat{S}$.

1. For $S \in \mathcal{T}_{d,s}$:
   (a) Compute $\widehat{\gamma} = (X_S^\top X_S)^{-1} X_S^\top Y$;
   (b) Compute $\mathcal{L}(S) := \frac{\|\Pi_S^\perp Y\|^2}{n-s} + \min_{\gamma \in \Theta} (\widehat{\gamma} - \gamma)^\top \frac{X_S^\top X_S}{n} (\widehat{\gamma} - \gamma)$;
2. Output $\widehat{S} = \arg\min_{S \in \mathcal{T}_{d,s}} \mathcal{L}(S)$.

KL-BSS. Algorithm 8 can be equivalently re-cast as the MIP in Section 4.3.3, and so everything below applies equally well to the MIP version of Vanilla KL-BSS. Similar to BSS, it can be written as the following estimator: Instead of using sum of squared residual as score, Vanilla KL-BSS minimizes the score $\mathcal{L}(S)$ defined in Algorithm 8.

$$\widehat{S} = \arg\min_{S \in \mathcal{T}_{d,s}} \mathcal{L}(S)\,.$$

Similarly, we define the uniform signal using $\widetilde{\Delta}_2(S,T)$ instead of $\Delta_2(S,T)$ below. For $\mathcal{M} := \mathcal{M}(\Theta, \Omega, \sigma^2)$ define

$$\widetilde{\Delta}(\mathcal{M}) := \frac{1}{\sigma^2} \min_{(\beta,\Sigma,\sigma^2) \in \mathcal{M}} \min_{T \in \mathcal{T}_{d,s} \setminus \{S_*\}} \frac{1}{|S_* \setminus T|} \Big(\Delta_1(S_*, T) \vee \widetilde{\Delta}_2(S_*, T) \Big)\,. \tag{C.9}$$

Theorem C.7.4 below establishes the sample complexity of Vanilla KL-BSS for successful variable selection:

Theorem C.7.4. *Assuming $s \le d/2$, for any $(\beta, \Sigma, \sigma^2) \in \mathcal{M} := \mathcal{M}(\Theta, \Omega, \sigma^2)$, let $S_* =$*

$\mathrm{supp}(\beta)$ *and* $|S_*| = s$. *Given i.i.d. samples from* $P_{\beta,\Sigma,\sigma^2}$, *if the sample size satisfies*

$$\begin{aligned} n - s &\gtrsim \max_{r \in [s]} \frac{\log \binom{d-s}{r} + s + \log(1/\delta)}{r\widetilde{\Delta}(\mathcal{M}) \wedge 1} \\ &\asymp \max \left\{ \frac{\log\left(d-s\right) + s + \log(1/\delta)}{\widetilde{\Delta}(\mathcal{M})}, \log \binom{d-s}{s} + \log(1/\delta) \right\}, \end{aligned}$$

then $\mathbb{P}_{\beta,\Sigma,\sigma^2}(\widehat{S} = S_*) \geq 1 - \delta$, *where* $\widehat{S}$ *is given by Algorithm 8.*

The proof is the same as Theorem C.7.1 by applying Lemma C.7.3 and is omitted. Compared to Theorem C.7.1 for KL-BSS, there are two differences: The signal is larger since $\widetilde{\Delta}(\mathcal{M}) \geq \Delta(\mathcal{M})$, but there is an additional factor of s in the numerator. Thus, Theorem C.7.4 is not necessarily an improvement due to the extra factor $s/\widetilde{\Delta}(\mathcal{M})$, although the tradeoff is minimal in light of the dominant factor of $\log \binom{d-s}{s} \asymp s \log d$ in the sample complexity of both methods. In Appendix E.6, we illustrate this point on a concrete example.

C.7.3 Beyond Gaussian design

To avoid technical complications, we have assumed Gaussian design and noise in (4.1). Under Gaussianity, the residual variance, which is the main object to deal with in the proofs, is conditionally subject to a χ^2 distribution, e.g. (C.13), for which we can apply concentration bounds as in Lemma D.1.2. Extended to the non-Gaussian setting, one can still derive similar results by resorting to concentration inequalities of sample (co)variance of (uncorrelated) random variables, e.g. sub-Gaussian with Bernstein type bounds. The main modification to the setup and proof will be as follows. Consider i.i.d. sub-Gaussian random vectors X and sub-Gaussian noise variable ϵ, and X is independent of ϵ (cf. (4.1)). Our main results rely on Lemma C.12.1, which is further proved by Lemma C.13.1, C.13.3 and C.13.4. For Lemma C.13.1, the same arguments apply for sub-

Gaussian covariance matrix estimation. For Lemma C.13.3 and C.13.4, we use Hanson-Wright inequality [Rudelson and Vershynin, 2013] for norm of projected sub-Gaussian random vectors (e.g. $\|\Pi_T^{\perp}\epsilon\|^2$) instead of Lemma D.1.2 for concentration of χ^2 distribution.

C.8 Proof of Lemma 4.3.1

Proof of Lemma 4.3.1. The proof is given by the following chain of inequalities: by definition, For any $T \in \mathcal{T}_{d,s} \setminus \{S_*\}$, let $S' = S_* \setminus T$, $T' = T \setminus S_*$, $W = S_* \cap T$, $r := |S_* \setminus T|$, $u = (u_{S'}, u_{T'}) = (u_1, u_2)$, we have

$$\min_{u \in \Theta_{S_* \triangle T}} \frac{u^\top \Sigma_{S_* \triangle T \,|\, S_* \cap T} u}{\min_{|R|=r} \|u_R\|^2} \geq \min_{u \in \mathbb{R}^{2r}} \frac{u^\top \Sigma_{S_* \triangle T \,|\, S_* \cap T} u}{\min_{|R|=r} \|u_R\|^2}.$$

Then

$$\begin{aligned}
&\frac{u^\top \Sigma_{S' \cup T' \,|\, } u}{\min_{|R|=r} \|u_R\|^2} \\
=&\frac{(u_2 - \Sigma^{-1}_{T'T' \,|\, W} \Sigma_{T'S' \,|\, W} u_1)^\top \Sigma_{S' \,|\, W} (u_2 - \Sigma^{-1}_{T'T' \,|\, W} \Sigma_{T'S' \,|\, W} u_1) + u_1^\top \Sigma_{S' \,|\, T} u_1}{\min_{|R|=r} \|u_R\|^2} \\
\geq&\frac{u_1^\top \Sigma_{S' \,|\, T} u_1}{\min_{|R|=r} \|u_R\|^2} \geq \frac{\lambda_{\min}(\Sigma_{S' \,|\, T}) \|u_1\|^2}{\min_{|R|=r} \|u_R\|^2} \geq \lambda_{\min}(\Sigma_{S' \,|\, T}).
\end{aligned}$$

Taking minimum on both sides yields the statement.

To see $\lambda_{\mathrm{K}}(\Sigma) > \lambda_{\mathrm{B}}(\Sigma)$ for any $\Sigma \in \Omega_\Delta$, by construction of Ω_Δ, we only need to show

$\lambda_{\mathrm{K}}(\Sigma) \geq \min_{T \in \mathcal{T}_{d,s} \setminus \{S_*\}} \lambda_{\min}(\Sigma_{T \setminus S_* \,|\, S_*})$. Again, for any $T \in \mathcal{T}_{d,s} \setminus \{S_*\}$,

$$
\begin{aligned}
&\frac{u^\top \Sigma_{S' \cup T' \,|\, u}}{\min_{|R|=r} \|u_R\|^2} \\
=&\frac{(u_1 - \Sigma^{-1}_{S'S' \,|\, W} \Sigma_{S'T' \,|\, W} u_2)^\top \Sigma_{S' \,|\, W} (u_1 - \Sigma^{-1}_{S'S' \,|\, W} \Sigma_{S'T' \,|\, W} u_2) + u_2^\top \Sigma_{T' \,|\, S_*} u_2}{\min_{|R|=r} \|u_R\|^2} \\
\geq&\frac{u_2^\top \Sigma_{T' \,|\, S_*} u_2}{\min_{|R|=r} \|u_R\|^2} \geq \frac{\lambda_{\min}(\Sigma_{T' \,|\, S_*}) \|u_2\|^2}{\min_{|R|=r} \|u_R\|^2} \geq \lambda_{\min}(\Sigma_{T' \,|\, S_*}) \,.
\end{aligned}
$$

Taking minimum over all T completes the proof. □

C.9 Proof of Theorem 4.3.2

Proof of Theorem 4.3.2. We will invoke the analysis in Appendix 4.3.2. Especially, we show λ_{B} and λ_{K} provide lower bound for the signal Δ_1 and Δ_2. For any $T \in \mathcal{T}_{d,s} \setminus \{S_*\}$, by Lemma C.7.2,

$$
\begin{aligned}
&\Delta_1(S_*, T) \vee \Delta_2(S_*, T) \\
\asymp& \frac{1}{\sigma^2} \min_{\alpha_{T \setminus S_*} \in \Theta_{T \setminus S_*}} \mathrm{var}\left[X^\top_{S_* \setminus T} \beta_{S_* \setminus T} - X^\top_{T \setminus S_*} \alpha_{T \setminus S_*} \,|\, S_* \cap T \right] \\
=& \frac{1}{\sigma^2} \min_{\alpha_{T \setminus S_*} \in \Theta_{T \setminus S_*}} \left[\beta_{S_* \setminus T}, -\alpha_{T \setminus S_*} \right]^\top \Sigma_{S_* \triangle T \,|\, S_* \cap T} \left[\beta_{S_* \setminus T}, -\alpha_{T \setminus S_*} \right] \\
\geq& |S_* \setminus T| \times \beta^2_{\min} \lambda_{\mathrm{K}}(\Sigma) / \sigma^2 \,.
\end{aligned}
$$

Similarly,

$$
\Delta_1(S_*, T) \geq |S_* \setminus T| \times \beta^2_{\min} \lambda_{\mathrm{B}}(\Sigma) / \sigma^2 \,.
$$

Therefore, the pointwise sample complexity is given by invoking Theorem C.7.1 and (C.7).

□

C.10 Proof of Theorem 4.3.3

Proof of Theorem 4.3.3. We prove by providing sample complexity upper and lower bounds.

Upper bounds: By construction, for any Σ coming from Ω_{K} and Ω_{B}, we have $\lambda_{\mathrm{K}}(\Sigma) \geq c_0\sigma_{\min}^2$ and $\lambda_{\mathrm{B}}(\Sigma) \geq c_0\sigma_{\min}^2$, respectively. By Lemma 4.3.1 and Theorem 4.3.2, we conclude the sample complexity of KL-BSS on Ω_{K} and Ω_{B}, and BSS on Ω_{B} are

$$\frac{\log d}{\beta_{\min}^2\sigma_{\min}^2/\sigma^2} \vee \log \binom{d-s}{s}.$$

Lower bounds: The lower bounds of Ω_{K} or Ω_{B} are based on Theorem C.10.1 and Corollary C.10.2 presented below, whose proofs are given later on.

Theorem C.10.1. *Given n i.i.d. samples from $P_{\beta,\Sigma,\sigma^2}$ with $(\beta,\Sigma,\sigma^2) \in \mathcal{M} := \mathcal{M}(\Theta,\Omega,\sigma^2)$ for $\Omega = \Omega_{\mathrm{B}}$ or Ω_{K}. If the sample size is bounded as*

$$n \leq \frac{1-2\delta}{2} \times \frac{\log(d-s)}{\beta_{\min}^2\sigma_{\min}^2/\sigma^2},$$

then for any estimator $\widehat{S}$ for $S_* = \mathrm{supp}(\beta)$,

$$\inf_{\widehat{S}} \sup_{(\beta,\Sigma,\sigma^2)\in\mathcal{M}} \mathbb{P}_{\beta,\Sigma,\sigma^2}(\widehat{S} \neq S_*) \geq \delta - \frac{\log 2}{\log(d-1)}.$$

Corollary C.10.2. *Given n i.i.d. samples from $P_{\beta,\Sigma,\sigma^2}$ with $(\beta,\Sigma,\sigma^2) \in \mathcal{M} := \mathcal{M}(\Theta,\Omega,\sigma^2)$ for $\Omega = \Omega_{\mathrm{B}}$ or Ω_{K}. If the sample size is bounded as*

$$n \leq 2(1-\delta) \times \frac{\log\binom{d-1}{s} - 1}{\log(1+s\beta_{\min}^2\sigma_{\min}^2/\sigma^2)},$$

then for any estimator $\widehat{S}$ for $S_* = \mathrm{supp}(\beta)$,

$$\inf_{\widehat{S}} \sup_{(\beta,\Sigma,\sigma^2)\in\mathcal{M}} \mathbb{P}_{\beta,\Sigma,\sigma^2}(\widehat{S} \neq S_*) \geq \delta\,.$$

The upper and lower bounds above together conclude the optimality. □

Remark 14. Both Theorem C.10.1 and Corollary C.10.2 extend beyond SEM to general designs as follows: Replace Ω_{K} with $\Omega'_{\mathrm{K}} = \{\Sigma \in \mathbb{S}^d_{++} : \lambda_{\mathrm{K}}(\Sigma) \geq \omega\}$ and Ω_{B} with $\Omega'_{\mathrm{B}} = \{\Sigma \in \mathbb{S}^d_{++} : \lambda_{\mathrm{B}}(\Sigma) \geq \omega\}$. Then the terms involving $\sigma^2_{\min}$ in the lower bounds can be replaced with ω. For example, the lower bound in Theorem C.10.1 becomes

$$n \asymp \frac{\log(d-s)}{\beta^2_{\min}\omega/\sigma^2}\,.$$

Similarly, matching upper bounds can be derived as well by Theorem 4.3.2. This confirms that the sample complexity of both methods genuinely depends on the eigenvalues λ_{K} and λ_{B} (i.e. through ω) as opposed to $\sigma^2_{\min}$.

Proof of Theorem C.10.1. Consider the design covariance generated by an empty graph where $X_k = \epsilon_k \sim \mathcal{N}(0, \sigma^2_{\min})$ for all $k \in [d]$. The SEM generated in this way satisfies $\lambda_{\mathrm{K}}(\Sigma) = \lambda_{\mathrm{B}}(\Sigma) = \sigma^2_{\min}$. We fix Σ, the empty graph G and coefficient vector $\beta = \beta_{\min}\mathbf{1}_d$, construct the ensemble solely by varying support.

$$\mathcal{S} := \left\{ S : S = \{1, 2, \cdots, s-1\} \cup \{X_\ell\}, \ell \in \{s, s+1, \ldots, d\} \right\}.$$

Therefore, $|\mathcal{S}| = \binom{d-(s-1)}{1} = d - s + 1 \geq d - s$, and for any two elements S, T, we have

form

$$S = \{1, 2, \cdots, s\} \cup \{j\}$$
$$T = \{1, 2, \cdots, s\} \cup \{k\},$$

with $j \neq k$ and $j, k \in \{s, s+1, \ldots, d\}$. Denote the models determined by S and T to be P_S and P_T, we now calculate the KL divergence between them:

$$\begin{aligned}
\mathbf{KL}(P_S \| P_T) &= \mathbb{E}_{P_S} \log \frac{P_S}{P_T} \\
&= \mathbb{E}_X (X_S^\top \beta_S - X_T^\top \beta_T)^2 / 2\sigma^2 \\
&= \mathbb{E}_X (X_j - X_k)^2 \beta_{\min}^2 / 2\sigma^2 \\
&= \mathbb{E}_X (\epsilon_j - \epsilon_k)^2 \beta_{\min}^2 / 2\sigma^2 \\
&= \sigma_{\min}^2 \beta_{\min}^2 / \sigma^2 .
\end{aligned}$$

Finally, we apply Fano's inequality Corollary D.2.3 with KL divergence upper bound $\beta_{\min}^2 \sigma_{\min}^2 / \sigma^2$ and ensemble cardinality lower bound $d - s$, which completes the proof. □

Proof of Corollary C.10.2. Following the proof of Theorem 1 in Wang et al. [2010b], we construct a mixture of all possible supports in $\mathcal{T}_{d,s}$. We adopt the graph and SEM in the proof of Theorem C.10.1, which satisfies $\lambda_{\mathrm{K}}(\Sigma) = \lambda_{\mathrm{B}}(\Sigma) = \sigma_{\min}^2$. To align the notation with Wang et al. [2010b], denote the data matrix $\widetilde{X} = (X_1, X_2, \ldots, X_d) \in \mathbb{R}^{n \times (d-1)}$, $\widetilde{Y} = Y \in \mathbb{R}^n$, let $\mu(\widetilde{X}) = \mathbb{E}[\widetilde{Y} \mid \widetilde{X}] \in \mathbb{R}^n$ and

$$\Lambda(\widetilde{X}) = \mathbb{E}[\widetilde{Y}\widetilde{Y}^\top \mid \widetilde{X}] - \mu(\widetilde{X})\mu(\widetilde{X})^\top \in \mathbb{R}^{n \times n}$$

be the conditional mean and variance of $\widetilde{Y}$ given $\widetilde{X}$. If suffices to recognize that Lemma 1 in Wang et al. [2010b] still holds with covariance matrix I_d replaced by this construction,

i.e.

$$\mathbb{E}_{\widetilde{X}}[\Lambda(\widetilde{X})] = \left(\sigma^2 + s\beta_{\min}^2\sigma_{\min}^2\left(1 - \frac{s}{d}\right)\right) I_n \,.$$

Then combining Lemma 1 with equation (17) in Wang et al. [2010b] leads to the lower bound. □

C.11 Proof of Example 1

Example 1 uses the eigenvalue condition (4.23), which is sufficient but not necessary. Before proving sufficiency, we remark that the condition (4.23) can further be relaxed to

$$\lambda_{\min}(D) = \lambda_{\min}(\Sigma_{S_*^c \,|\, S_*}) > \min_{T \subseteq S_*^c, |T| = s} \lambda_{\min}(\Sigma_{S_* \,|\, T}), \tag{C.10}$$

which can be relaxed even further to

$$\min_{T \in \mathcal{T}_{d,s} \setminus \{S_*\}} \lambda_{\min}(\Sigma_{T \,|\, S_*}) > \min_{T \in \mathcal{T}_{d,s} \setminus \{S_*\}} \lambda_{\min}(\Sigma_{S_* \,|\, T}), \tag{C.11}$$

which recovers the original definition of Ω_Δ in (4.20). Thus, we see that (4.23) is just a special case of (4.20) in light of the relations

$$\min_{T \in \mathcal{T}_{d,s} \setminus \{S_*\}} \lambda_{\min}(\Sigma_{T \,|\, S_*}) \geq \lambda_{\min}(\Sigma_{S_*^c \,|\, S_*}) \quad \text{and} \quad \lambda_{\min}(\Sigma_{S_*}) \geq \min_{T \in \mathcal{T}_{d,s} \setminus \{S_*\}} \lambda_{\min}(\Sigma_{S_* \,|\, T}).$$

This is because the right hand side is essentially $\lambda_{\min}(\Sigma_{S_* \,|\, T})$ that we want to upper bound in Ω_Δ.

We now prove the following generalized version of Example 1. It follows from Lemma C.11.1 below that the example covariance class defined by (4.23) is contained in Ω_Δ, therefore, KL-BSS improves BSS in the sense that $\lambda_{\mathrm{K}}(\Sigma) > \lambda_{\mathrm{B}}(\Sigma)$.

Lemma C.11.1. *For any* Σ *from*

$$\left\{ \begin{pmatrix} C & A^\top \\ A & B + AC^{-1}A^\top \end{pmatrix} \middle| \, \lambda_{\min}(B) > \min_{T \subseteq S_*^c, |T| = s} \lambda_{\min}\Big(C - A_T^\top (B_{TT} + A_T C^{-1} A_T^\top)^{-1} A_T \Big) \right\},$$

we have $\min_{T \in \mathcal{T}_{d,s} \setminus \{S_*\}} \lambda_{\min}(\Sigma_{S_* \setminus T \,|\, T}) < \min_{T \in \mathcal{T}_{d,s} \setminus \{S_*\}} \lambda_{\min}(\Sigma_{T \setminus S_* \,|\, S_*})$.

Proof of Lemma C.11.1. For any

$$\Sigma = \begin{pmatrix} C & A^\top \\ A & B + AC^{-1}A^\top \end{pmatrix},$$

conditioning on X_{S_*}, we have $\mathrm{cov}(X_{S_*^c} \mid X_{S_*}) = B$. For any $T \in \mathcal{T}_{d,s} \setminus \{S_*\}$,

$$\lambda_{\min}(\Sigma_{T \setminus S_* \,|\, S_*}) \geq \lambda_{\min}(\Sigma_{S_*^c \,|\, S_*}) = \lambda_{\min}(B)\,.$$

We only need to upper bound $\min_{T \in \mathcal{T}_{d,s} \setminus \{S_*\}} \lambda_{\min}(\Sigma_{S_* \setminus T \,|\, T})$ by $\lambda_{\min}(B)$. To this end, take the $T \subseteq S_*^c$ with $|T| = s$ achieving the minimum, then $T \cap S_* = \emptyset$, and

$$\lambda_{\min}(\Sigma_{S_* \,|\, T}) = \lambda_{\min}\Big(C - A_T^\top (B_{TT} + A_T C^{-1} A_T^\top)^{-1} A_T \Big) < \lambda_{\min}(B)\,.$$

This completes the proof.

To see how this generalizes the example in the main paper, since $A \neq 0$, choose T such

that $A_T \neq 0$. Therefore,

$$\begin{aligned}
&\lambda_{\min}\Big(C - A_T^\top (B_{TT} + A_T C^{-1} A_T^\top)^{-1} A_T\Big) \\
&= \min_{\|v\|=1} \Big[v^\top C v - v^\top A_T^\top (B_{TT} + A_T C^{-1} A_T^\top)^{-1} A_T v\Big] \\
&< \min_{\|v\|=1} v^\top C v = \lambda_{\min}(C) \leq \lambda_{\min}(B)
\end{aligned}$$

Thus, Lemma C.11.1 also generalizes to the case where $A = 0$, which essentially requiring $\lambda_{\min}(C) < \lambda_{\min}(B)$ such that:

$$\min_{T \subseteq S_*^c, |T|=s} \lambda_{\min}(\Sigma_{S_* \,|\, T}) = \lambda_{\min}(C) < \lambda_{\min}(B) \leq \min_T \lambda_{\min}(\Sigma_{T \setminus S_* \,|\, S_*}) \,. \qquad \square$$

C.12 Proof of Theorem C.7.1

Before we prove the theorem, we firstly introduce the main device, which is an error probability bound for Algorithm 3 to successfully distinguish the true support S_* from any other candidate T:

Lemma C.12.1. *For any $(\beta, \Sigma, \sigma^2) \in \mathcal{M}(\Theta, \Omega, \sigma^2)$, let $S_* = \mathrm{supp}(\beta)$ and $|S_*| = s$. Given n i.i.d. samples from $P_{\beta,\Sigma,\sigma^2}$, apply Algorithm 3 to estimate support from S_* and T with $|S_* \setminus T| = r$ and output $\widehat{S}$. Let $\Delta_1 := \Delta_1(S_*, T)$ and $\Delta_2 := \Delta_2(S_*, T)$. If sample size satisfies $n \gtrsim s + \frac{r}{\Delta_1 \vee \Delta_2}$, then we have for some positive constant C_0,*

$$\mathbb{P}_{\beta,\Sigma,\sigma^2}(\widehat{S} = S_*) \geq 1 - 9\exp\Big(-C_0(n-s)\min\Big(\Delta_1 \vee \Delta_2, 1\Big) + r\Big)\,.$$

The proof is given in Appendix C.13. Then we are ready to prove the theorem.

Proof of Theorem C.7.1. We apply Lemma C.12.1, whose conditions are satisfied by the

stated sample complexity,

$$\begin{aligned}
\mathbb{P}(\widehat{S} \neq S_*) &\leq \mathbb{P}\left[\bigcup_{r=1}^{s} \bigcup_{\substack{T \in \mathcal{T}_{d,s} \setminus \{S_*\} \\ |S_* \setminus T| = r}} \left\{ \mathcal{L}(T; (S_*, T)) - \mathcal{L}(S_*; (S_*, T)) < 0 \right\} \right] \\
&\leq \sum_{r=1}^{s} \sum_{\substack{T \in \mathcal{T}_{d,s} \setminus \{S_*\} \\ |S_* \setminus T| = r}} \mathbb{P}\left[\mathcal{L}(T; (S_*, T)) - \mathcal{L}(S_*; (S_*, T)) < 0 \right] \\
&\leq \sum_{r=1}^{s} \sum_{\substack{T \in \mathcal{T}_{d,s} \setminus \{S_*\} \\ |S_* \setminus T| = r}} 9 \exp\left(- C_0 (n-s) \min\left(r\Delta(\mathcal{M}), 1 \right) + r \right) \\
&\leq \max_r s \times \binom{s}{r} \binom{d-s}{r} \times 9 \exp\left(- C_0 (n-s) \min\left(r\Delta(\mathcal{M}), 1 \right) + r \right).
\end{aligned}$$

We apply Lemma C.12.1 and definition of $\Delta(\mathcal{M})$ (C.4) for the third inequality, which relies on the scaling factor $|S_* \setminus T| = r$. Since $s \leq d/2$, $\binom{s}{r} \leq \binom{d-s}{r}$ and

$$\log 9s \leq \max_r \log 9 \binom{s}{r} \leq \max_r 2 \log \binom{s}{r},$$

when s is large enough. Therefore, the error probability

$$\begin{aligned}
&\mathbb{P}(\widehat{S} \neq S_*) \\
&\leq \max_r \exp\left(\log 9s + \log \binom{s}{r} + \log \binom{d-s}{r} + r - C_0 (n-s) \min\left(r\Delta(\mathcal{M}), 1 \right) \right) \\
&\leq \max_r \exp\left(5 \log \binom{d-s}{r} - C_0 (n-s) \min\left(r\Delta(\mathcal{M}), 1 \right) \right).
\end{aligned}$$

Setting the RHS to be smaller than δ for all r, we have desired sample complexity. □

C.13 Proof of Lemma C.12.1

Before proving the lemma, we first introduce a technical lemma on sample covariance matrix estimation. The proof follows from standard arguments, and so we only sketch its proof here:

Lemma C.13.1. *If $U \in \mathbb{R}^{n\times r}$ with each entry $U_{ij} \overset{iid}{\sim} \mathcal{N}(0,1)$, $\Pi \in \mathbb{R}^{n\times n}$ is an idempotent matrix and $\mathrm{Tr}(\Pi) = n-k$ with $k < n$, then for $t > 0$,*

$$\mathbb{P}\bigg(\|\frac{U^\top \Pi U}{n-k} - I_r\|_{\mathrm{op}} \geq t \bigg) \leq \exp\bigg(-C(n-k)\min(t, t^2) + r \bigg),$$

which implies $1 - t \leq \lambda_{\min}(\frac{U^\top \Pi U}{n-k}) \leq \lambda_{\max}(\frac{U^\top \Pi U}{n-k}) \leq 1 + t$ with high probability.

Proof sketch. We need only to recognize

$$\begin{aligned}
\|\frac{U^\top \Pi U}{n-k} - I_r\|_{\mathrm{op}} &= \max_{v\in\mathbb{R}^r, \|v\|=1} |v^\top (\frac{U^\top \Pi U}{n-k} - I_r) v| \\
&= \max_{v\in\mathbb{R}^r, \|v\|=1} |\frac{(vU)^\top \Pi (Uv)}{n-k} - 1| \\
&= \max_{v\in\mathbb{R}^r, \|v\|=1} |\frac{\chi^2_{n-k}}{n-k} - 1| \,.
\end{aligned}$$

Then the proof follows the standard analysis for sample covariance matrix using covering and packing number arguments on the $\max_{v\in\mathbb{R}^r, \|v\|=1}$. □

We also need the following lemma on the norm of projected Gaussian noise:

Lemma C.13.2. *If $U \in \mathbb{R}^{n\times r}$ with each entry $U_{ij} \overset{iid}{\sim} \mathcal{N}(0,1)$, $\Pi \in \mathbb{R}^{n\times n}$ is an idempotent matrix and $\mathrm{Tr}(\Pi) = n-k$ with $k \leq n$, $\xi \in \mathbb{R}^n$ with each entry $\xi \sim \mathcal{N}(0, \sigma^2)$, $U \perp\!\!\!\perp \xi$, then with $t \in (0,1)$ and the constant C in Lemma C.13.1, assuming $n \geq k + \frac{8r(1+t)}{\delta'}$,*

$$\mathbb{P}\bigg(\frac{\|U^\top \Pi \xi\|^2}{\sigma^2 (n-k)^2} \geq \delta' \bigg) \leq \exp\bigg(-(n-k)\frac{\delta'}{4(1+t)} + \frac{r}{4} \bigg) + \exp(-C(n-k)t^2 + r) \,.$$

Proof. Given U, $U^\top \Pi \xi \sim \mathcal{N}(0, \sigma^2 U^\top \Pi U)$, we can rewrite with $\nu \sim \mathcal{N}(0, I_r)$,

$$\begin{aligned}\frac{\|U^\top \Pi \xi\|^2}{\sigma^2 (n-k)^2} &= \frac{1}{n-k} \nu^\top \frac{U^\top \Pi U}{n-k} \nu \\ &\leq \|\frac{U^\top \Pi U}{n-k}\|_{\mathrm{op}} \frac{\|\nu\|^2}{n-k} \leq \frac{(1+t)\chi_r^2}{n-k}\,.\end{aligned}$$

For the second inequality, we invoke Lemma C.13.1 and corresponding error probability. Given that, we have

$$\begin{aligned}\mathbb{P}\left(\frac{\|U^\top \Pi \xi\|^2}{\sigma^2 (n-k)^2} \geq \delta'\right) &\leq \mathbb{P}\left(\chi_r^2 \geq \frac{(n-k)\delta'}{1+t}\right) \\ &= \mathbb{P}\left(\frac{\chi_r^2}{r} - 1 \geq \frac{(n-k)\delta'}{r(1+t)} - 1\right) \\ &\leq \exp\left(-(n-k)\frac{\delta'}{4(1+t)} + \frac{r}{4}\right),\end{aligned}$$

providing $n \geq k + \frac{8r(1+t)}{\delta'}$. □

Now we are ready to prove Lemma C.12.1.

Proof of Lemma C.12.1. We start with some notations. Write $S_* = S' \cup W$ and $T = T' \cup W$ with $W = S_* \cap T$. Denote $\alpha_\beta := \alpha_\beta(S_*, T)$. Let $\epsilon_0 \sim \mathcal{N}(0, \sigma^2 \Delta_1)$ be the part of Y that cannot be explained by T, i.e. $X_{S'} = \Sigma_{S'T}\Sigma_{TT}^{-1} X_T + \epsilon_{S' \mid T}$, $\epsilon_0 = \beta_{S'}^\top \epsilon_{S' \mid T}$. Let $\epsilon' := \epsilon_0 + \epsilon \sim \mathcal{N}(0, \sigma^2(1+\Delta_1))$. Furthermore, write $\widetilde{\epsilon} = \Pi_W^\perp \epsilon$, $\widetilde{\epsilon}' = \Pi_W^\perp \epsilon'$. Therefore, we can write

$$\begin{aligned}\widetilde{Y} := \Pi_W^\perp Y &= \Pi_W^\perp X_{S'}^\top \beta_{S'} + \Pi_W^\perp \epsilon = \widetilde{X}_{S'}^\top \beta_{S'} + \widetilde{\epsilon} \\ &= \Pi_W^\perp X_{T'}^\top \alpha_\beta + \Pi_W^\perp \epsilon' = \widetilde{X}_{T'}^\top \alpha_\beta + \widetilde{\epsilon}'\,.\end{aligned}$$

Denote the OLS vector $\widehat{\gamma}$ for S' and T' to be $\widehat{\beta}$ and $\widehat{\alpha}$, then we have

$$\widehat{\beta} = \beta_{S'} + (\widetilde{X}_{S'}^\top \widetilde{X}_{S'})^{-1} \widetilde{X}_{S'}^\top \widetilde{\epsilon}$$
$$\widehat{\alpha} = \alpha_\beta + (\widetilde{X}_{T'}^\top \widetilde{X}_{T'})^{-1} \widetilde{X}_{T'}^\top \widetilde{\epsilon}' \,.$$

Let

$$\widehat{\beta}^* = \underset{\widetilde{\beta} \in \Theta_{S'}}{\arg\min} (\widehat{\beta} - \widetilde{\beta})^\top \frac{\widetilde{X}_{S'}^\top \widetilde{X}_{S'}}{n - (s-r)} (\widehat{\beta} - \widetilde{\beta})$$
$$\widehat{\alpha}^* = \underset{\alpha \in \Theta_{T'}}{\arg\min} (\widehat{\alpha} - \alpha)^\top \frac{\widetilde{X}_{T'}^\top \widetilde{X}_{T'}}{n - (s-r)} (\widehat{\alpha} - \alpha)$$
$$\alpha^* = \underset{\alpha \in \Theta_{T'}}{\arg\min} (\alpha_\beta - \alpha)^\top \Sigma_{T' \,|\, W} (\alpha_\beta - \alpha) \,,$$

then we have the scores

$$\mathcal{L}(S_*; (S_*, T)) = \frac{\|\Pi_{S_*}^\perp \epsilon\|^2}{n - s} + \widehat{\Delta}_2(S_*)$$
$$\mathcal{L}(T; (S_*, T)) = \frac{\|\Pi_T^\perp \epsilon'\|^2}{n - s} + \widehat{\Delta}_2(T) \,,$$

where we denote

$$\widehat{\Delta}_2(S_*) = (\beta_{S'} - \widehat{\beta}^*)^\top \frac{\widetilde{X}_{S'}^\top \widetilde{X}_{S'}}{n - (s-r)} (\beta_{S'} - \widehat{\beta}^*) + \frac{\|\Pi_{\widetilde{S}'} \widetilde{\epsilon}\|^2}{n - (s-r)} + 2\langle \beta_{S'} - \widehat{\beta}^*, \frac{\widetilde{X}_{S'}^\top \widetilde{\epsilon}}{n - (s-r)} \rangle$$
$$\widehat{\Delta}_2(T) = (\alpha_\beta - \widehat{\alpha}^*)^\top \frac{\widetilde{X}_{T'}^\top \widetilde{X}_{T'}}{n - (s-r)} (\alpha_\beta - \widehat{\alpha}^*) + \frac{\|\Pi_{\widetilde{T}'} \widetilde{\epsilon}'\|^2}{n - (s-r)} + 2\langle \alpha_\beta - \widehat{\alpha}^*, \frac{\widetilde{X}_{T'}^\top \widetilde{\epsilon}'}{n - (s-r)} \rangle \,, \tag{C.12}$$

with $\Pi_{\widetilde{R}} = \widetilde{X}_R (\widetilde{X}_R^\top \widetilde{X}_R)^{-1} \widetilde{X}_R^\top$ for $R = S'$ or T'. Note we have $\widetilde{X}_R = \Pi_W^\perp X_R = \Pi_W^\perp \epsilon_{R \,|\, W} = \Pi_W^\perp U_{R \,|\, W} \Sigma_{R \,|\, W}^{1/2}$ where $U_{R \,|\, W} \in \mathbb{R}^{n \times r}$ and $U_{(R \,|\, W), ij} \overset{iid}{\sim} \mathcal{N}(0,1)$. Thus given W, $\mathrm{Tr}(\Pi_W^\perp) = n - (s-r)$, invoking Lemma C.13.1, we have for $t \in (0,1)$, with probability greater than

$1 - 2\exp(-C(n-(s-r))t^2 + r)$,

$$\left\| \frac{\Sigma_{R\,|\,W}^{-1/2} \widetilde{X}_R^\top \widetilde{X}_R \Sigma_{R\,|\,W}^{-1/2}}{n-(s-r)} - I_r \right\|_{\mathrm{op}} \leq t \quad \forall R = S', T' \,.$$

Then the proof is based on two additional lemmas:

Lemma C.13.3. *Providing* $n \geq s + \frac{64r}{\Delta_1 \vee \Delta_2}$,

$$\mathbb{P}\bigg(\frac{\|\Pi_T^\perp \epsilon'\|^2 - \|\Pi_{S_*}^\perp \epsilon\|^2}{\sigma^2(n-s)} \geq \frac{2}{3}\Delta_1 - \frac{1}{4}\Delta_1 \vee \Delta_2 \bigg) \geq 1 - 5\exp\bigg(-(n-s)\frac{\min(\Delta_1 \vee \Delta_2, 1)}{64^2} \bigg) .$$

Lemma C.13.4. *Providing* $n \geq (s-r) + \frac{2C'r}{\min(\Delta_1 \vee \Delta_2, 1)}$, *for some constant* C',

$$\begin{aligned} &\mathbb{P}\bigg(\frac{\widehat{\Delta}_2(T) - \widehat{\Delta}_2(S_*)}{\sigma^2} \geq \frac{2}{3}\Delta_2 - \frac{1}{4}\Delta_1 \vee \Delta_2 \bigg) \\ \geq &1 - 4\exp\bigg(-C'\Big(n - (s-r)\Big)\min(\Delta_1 \vee \Delta_2, 1) + r \bigg) . \end{aligned}$$

Combining Lemma C.13.3 and C.13.4, it suffices to have $n \gtrsim s + \frac{r}{\Delta_1 \vee \Delta_2}$ to ensure the conditions are satisfied, i.e. $n \geq s + \frac{\max(64, 2C')r}{\min(\Delta_1 \vee \Delta_2, 1)}$. Let $C_0 = \min(C', 1/64^2)$, with probability at least

$$\begin{aligned} &1 - 5\exp\bigg(-(n-s)\frac{\min(\Delta_1 \vee \Delta_2, 1)}{1024} \bigg) \\ &\quad - 4\exp\bigg(-C'\Big(n - (s-r)\Big)\min(\Delta_1 \vee \Delta_2, 1) + r \bigg) \\ \geq &1 - 9\exp\bigg(-C_0(n-s))\min(\Delta_1 \vee \Delta_2, 1) + r \bigg) , \end{aligned}$$

we have

$$\begin{aligned}\mathcal{L}(T;(S_*,T)) - \mathcal{L}(S_*;(S_*,T)) &\geq \frac{2}{3}(\Delta_1+\Delta_2) - \frac{1}{2}\Delta_1 \vee \Delta_2 \\ &\geq \frac{1}{6}\Delta_1 \vee \Delta_2 > 0\,,\end{aligned}$$

which implies successful recovery $\widehat{S} = S_*$, and completes the proof. □

We proceed to show the proofs for Lemma C.13.3 and C.13.4.

Proof of Lemma C.13.3. Note that

$$\begin{aligned}\frac{\|\Pi_T^{\perp}\epsilon'\|^2 - \|\Pi_{S_*}^{\perp}\epsilon\|^2}{\sigma^2(n-s)} &= \frac{(\|\Pi_T^{\perp}\epsilon\|^2 - \|\Pi_{S_*}^{\perp}\epsilon\|^2) + \|\Pi_T^{\perp}\epsilon_0\|^2 + 2\langle\Pi_T^{\perp}\epsilon_0, \Pi_T^{\perp}\epsilon\rangle}{\sigma^2(n-s)} \\ &=: A_1 + A_2 + A_3\,.\end{aligned} \tag{C.13}$$

We will choose $C_{11}, C_{12}, C_{13} \in (0,1)$, then we have for A_1,

$$\begin{aligned}&\mathbb{P}(A_1 \leq -C_{11}\Delta_1 \vee \Delta_2) \\ =&\mathbb{P}\left(\frac{\|\Pi_T^{\perp}\epsilon\|^2 - \|\Pi_{T\cap S_*}^{\perp}\epsilon\|^2}{\sigma^2(n-s)} - \frac{\|\Pi_{S_*}^{\perp}\epsilon\|^2 - \|\Pi_{T\cap S_*}^{\perp}\epsilon\|^2}{\sigma^2(n-s)} \leq -C_{11}\Delta_1 \vee \Delta_2\right) \\ \leq&2\mathbb{P}\left(|\frac{\chi_r^2}{r} - 1| \geq \frac{C_{11}(n-s)}{2r}\Delta_1 \vee \Delta_2\right) \\ \leq&2\exp(-(n-s)\frac{C_{11}}{8}\Delta_1 \vee \Delta_2)\,.\end{aligned}$$

Given $n \geq s + \frac{8}{C_{11}} \times \frac{r}{\Delta_1 \vee \Delta_2}$. For A_2, since $C_{12} \in (0,1)$,

$$\begin{aligned}\mathbb{P}(A_2 \leq C_{12}\Delta_1) &= \mathbb{P}(\frac{\chi_{n-s}^2}{n-s} - 1 \leq C_{12} - 1) \\ &\leq \exp(-(n-s) \times \frac{(1-C_{12})^2}{16})\,.\end{aligned}$$

For A_3, let $U_\epsilon = \epsilon/\sigma$, $U_{\epsilon_0} = \epsilon_0/\sqrt{\Delta_1\sigma^2}$,

$$\begin{aligned}
\mathbb{P}(A_3 \le -C_{13}\Delta_1 \vee \Delta_2) &= \mathbb{P}\bigg(\frac{\|\Pi_T^\perp \frac{U_\epsilon + U_{\epsilon_0}}{\sqrt{2}}\|^2 - \|\Pi_T^\perp \frac{U_\epsilon - U_{\epsilon_0}}{\sqrt{2}}\|^2}{n-s} \le -C_{13}\frac{\Delta_1 \vee \Delta_2}{\sqrt{\Delta_1}}\bigg) \\
&\le 2\mathbb{P}(|\frac{\chi^2_{n-s}}{n-s} - 1| \ge \frac{C_{13}}{2}\frac{\Delta_1 \vee \Delta_2}{\sqrt{\Delta_1}}) \\
&\le 2\mathbb{P}(|\frac{\chi^2_{n-s}}{n-s} - 1| \ge \frac{C_{13}}{2}\sqrt{\Delta_1 \vee \Delta_2}) \\
&\le 2\exp(-(n-s)\min(\sqrt{\Delta_1 \vee \Delta_2}, \Delta_1 \vee \Delta_2) \times \frac{C_{13}^2}{64}) .
\end{aligned}$$

Let $C_{11} = \frac{1}{8}, C_{12} = \frac{2}{3}, C_{13} = \frac{1}{8}$, we conclude

$$\begin{aligned}
&\mathbb{P}\bigg(\frac{\|\Pi_T^\perp \epsilon'\|^2 - \|\Pi_{S_*}^\perp \epsilon\|^2}{\sigma^2(n-s)} \le \frac{2}{3}\Delta_1 - \frac{1}{4}\Delta_1 \vee \Delta_2\bigg) \\
&\le \mathbb{P}(A_1 \le -C_{11}\Delta_1 \vee \Delta_2) + \mathbb{P}(A_2 \le C_{12}\Delta_1) \\
&+ \mathbb{P}(A_3 \le -C_{13}\Delta_1 \vee \Delta_2) \\
&\le 2\exp(-(n-s)\frac{C_{11}}{8}\Delta_1 \vee \Delta_2) \\
&+ \exp(-(n-s) \times \frac{(1-C_{12})^2}{64}) \\
&+ 2\exp(-(n-s)\min(\sqrt{\Delta_1 \vee \Delta_2}, \Delta_1 \vee \Delta_2) \times \frac{C_{13}^2}{16}) \\
&\le 5\exp(-(n-s)\frac{\min(\Delta_1 \vee \Delta_2, 1)}{64^2}) .
\end{aligned}$$

□

Proof of Lemma C.13.4. Since $\widehat{\beta}^*$ is the minimizer, we have

$$\begin{aligned}
\widehat{\Delta}_2(S_*) &= (\widehat{\beta} - \widehat{\beta}^*)^\top \frac{\widetilde{X}_{S'}^\top \widetilde{X}_{S'}}{n-(s-r)}(\widehat{\beta} - \widehat{\beta}^*) \\
&\le (\widehat{\beta} - \beta_{S'})^\top \frac{\widetilde{X}_{S'}^\top \widetilde{X}_{S'}}{n-(s-r)}(\widehat{\beta} - \beta_{S'}) = \frac{\|\Pi_{\widetilde{S}'}\epsilon\|^2}{n-(s-r)} .
\end{aligned}$$

On the other hand, for some $t \in (0,1)$ which will be specified later, with probability greater than $1 - 2\exp(-C(n-(s-r))t^2 + r)$,

$$\begin{aligned}
\widehat{\Delta}_2(T) &= (\alpha_\beta - \widehat{\alpha}^*)^\top \frac{\widetilde{X}_{T'}^\top \widetilde{X}_{T'}}{n-(s-r)}(\alpha_\beta - \widehat{\alpha}^*) + \frac{\|\Pi_{\widetilde{T}'}\widetilde{\epsilon}'\|^2}{n-(s-r)} + 2\langle \alpha_\beta - \widehat{\alpha}^*, \frac{\widetilde{X}_{T'}^\top \widetilde{\epsilon}'}{n-(s-r)}\rangle \\
&\geq (1-t)(\alpha_\beta - \widehat{\alpha}^*)^\top \Sigma_{T'\,|\,W}(\alpha_\beta - \widehat{\alpha}^*) + \frac{\|\Pi_{\widetilde{T}'}\widetilde{\epsilon}'\|^2}{n-(s-r)} \\
&\quad - 2\|\Sigma_{T'\,|\,W}^{1/2}(\alpha_\beta - \widehat{\alpha}^*)\|\,\|U_{T'\,|\,W}^\top \widetilde{\epsilon}'/(n-(s-r))\| \\
&\geq (1-t)\Delta_2\sigma^2 + \frac{\|\Pi_{\widetilde{T}'}\widetilde{\epsilon}'\|^2}{n-(s-r)} - 2\|\Sigma_{T'\,|\,W}^{1/2}(\alpha_\beta - \widehat{\alpha}^*)\|\,\|U_{T'\,|\,W}^\top \widetilde{\epsilon}'/(n-(s-r))\|\,,
\end{aligned}$$

where the last inequality is because $\Delta_2 = \min_{\alpha \in \Theta_{T'}}(\alpha_\beta - \alpha)^\top \Sigma_{T'\,|\,W}(\alpha_\beta - \alpha)$. Then

$$\begin{aligned}
\frac{\widehat{\Delta}_2(T) - \widehat{\Delta}_2(S_*)}{\sigma^2} &\geq (1-t)\Delta_2 + \left(\frac{\|\Pi_{\widetilde{T}'}\widetilde{\epsilon}'\|^2}{\sigma^2(n-(s-r))} - \frac{\|\Pi_{\widetilde{S}'}\widetilde{\epsilon}\|^2}{\sigma^2(n-(s-r))}\right) \\
&\qquad - 2\frac{\|\Sigma_{T'\,|\,W}^{1/2}(\alpha_\beta - \widehat{\alpha}^*)\|\,\|U_{T'\,|\,W}^\top \widetilde{\epsilon}'\|}{\sigma^2(n-(s-r))} \\
&:= (1-t)\Delta_2 + B_1 - B_2\,.
\end{aligned}$$

We will choose $C_{21}, C_{22} \in (0,1)$. For B_1, since $\mathrm{Tr}(\Pi_{\widetilde{T}'}) = \mathrm{Tr}(\Pi_{\widetilde{S}'}) = r$, let $K, K' \sim \chi_r^2$, conditioned on S_* and T,

$$\begin{aligned}
\mathbb{P}(B_1 \leq -C_{21}\Delta_1 \vee \Delta_2) &= \mathbb{P}\left(\frac{(1+\Delta_1)K - K'}{n-(s-r)} \leq -C_{21}\Delta_1 \vee \Delta_2\right) \\
&\leq \mathbb{P}\left(\frac{K-K'}{n-(s-r)} \leq -C_{21}\Delta_1 \vee \Delta_2\right) \\
&\leq 2\mathbb{P}\left(|\frac{\chi_r^2}{r} - 1| \leq \frac{C_{21}(n-(s-r))}{2r}\Delta_1 \vee \Delta_2\right) \\
&\leq 2\exp(-(n-(s-r))\frac{C_{21}}{8}\Delta_1 \vee \Delta_2)\,,
\end{aligned}$$

given $n \geq (s-r) + \frac{8}{C_{21}} \times \frac{r}{\Delta_1 \vee \Delta_2}$. For B_2, we invoke Lemma C.13.2 for $\|U_{T' \mid W}^\top \widetilde{\epsilon}'\|$:

$$\mathbb{P}\bigg(\frac{\|U_{T' \mid W}^\top \widetilde{\epsilon}'\|^2}{\sigma^2 (n-(s-r))^2} \geq C_{22}\Delta_1 \vee \Delta_2 \bigg) \leq \exp\bigg(-(n-(s-r)) \frac{C_{22}\Delta_1 \vee \Delta_2}{4(1+t)(1+\Delta_1)} + \frac{r}{4} \bigg) + \exp(-C(n-(s-r))t^2 + r)\,,$$

providing $n \geq (s-r) + \frac{8r(1+t)(1+\Delta_1)}{C_{22}\Delta_1 \vee \Delta_2}$. We further discuss $\frac{1+\Delta_1}{\Delta_1 \vee \Delta_2}$ in cases:

- If $\Delta_1 < 1$: $\frac{1+\Delta_1}{\Delta_1 \vee \Delta_2} < \frac{2}{\Delta_1 \vee \Delta_2}$;
- If $\Delta_1 \geq 1$: $\frac{1+\Delta_1}{\Delta_1 \vee \Delta_2} \leq \frac{2\Delta_1}{\Delta_1 \vee \Delta_2} \leq 2$.

Thus we only need $n \geq (s-r) + \frac{16r(1+t)}{C_{22}}(1 \vee \frac{1}{\Delta_1 \vee \Delta_2})$ to ensure

$$\mathbb{P}\bigg(\frac{\|U_{T' \mid W}^\top \widetilde{\epsilon}'\|^2}{\sigma^2 (n-s)^2} \geq C_{22}\Delta_1 \vee \Delta_2 \bigg) \leq \exp\bigg(-(n-(s-r)) \frac{C_{22}\min(\Delta_1 \vee \Delta_2, 1)}{8(1+t)} + \frac{r}{4} \bigg) + \exp(-C(n-(s-r))t^2 + r)\,.$$

Then for $\|\Sigma_{T' \mid W}^{1/2}(\alpha_\beta - \widehat{\alpha}^*)\|$, we start with the fact that $\widehat{\alpha}^*$ is the minimizer of the programming, we have

$$\frac{1}{\sigma^2}(\widehat{\alpha} - \widehat{\alpha}^*)^\top \frac{\widetilde{X}_{T'}^\top \widetilde{X}_{T'}}{n-(s-r)} (\widehat{\alpha} - \widehat{\alpha}^*) \leq \frac{1}{\sigma^2}(\widehat{\alpha} - \alpha^*)^\top \frac{\widetilde{X}_{T'}^\top \widetilde{X}_{T'}}{n-(s-r)} (\widehat{\alpha} - \alpha^*)\,.$$

Expand both sides,

$$\begin{aligned}
&\frac{1}{\sigma^2}(\alpha_\beta - \widehat{\alpha}^*)^\top \frac{\widetilde{X}_{T'}^\top \widetilde{X}_{T'}}{n-(s-r)} (\alpha_\beta - \widehat{\alpha}^*) + \frac{2}{\sigma^2}\langle \Sigma_{T' \mid W}^{1/2}(\alpha_\beta - \widehat{\alpha}^*), U_{T' \mid W}^\top \widetilde{\epsilon}'/(n-(s-r)) \rangle \\
\leq &\frac{1}{\sigma^2}(\alpha_\beta - \alpha^*)^\top \frac{\widetilde{X}_{T'}^\top \widetilde{X}_{T'}}{n-(s-r)} (\alpha_\beta - \alpha^*) + \frac{2}{\sigma^2}\langle \Sigma_{T' \mid W}^{1/2}(\alpha_\beta - \alpha^*), U_{T' \mid W}^\top \widetilde{\epsilon}'/(n-(s-r)) \rangle \\
\leq &(1+t)\Delta_2 + \frac{2}{\sigma^2}\|\Sigma_{T' \mid W}^{1/2}(\alpha_\beta - \alpha^*)\| \|U_{T' \mid W}^\top \widetilde{\epsilon}'/(n-(s-r))\| \\
\leq &(1+t)\Delta_2 + 2\sqrt{\Delta_2}\sqrt{C_{22}\Delta_1 \vee \Delta_2}\,.
\end{aligned}$$

Moreover,

$$\begin{aligned}\frac{1-t}{\sigma^2}\|\Sigma^{1/2}_{T'\,|\,W}(\alpha_\beta-\widehat{\alpha}^*)\|^2 &\le \frac{1}{\sigma^2}(\alpha_\beta-\widehat{\alpha}^*)^\top\frac{\widetilde{X}^\top_{T'}\widetilde{X}_{T'}}{n-(s-r)}(\alpha_\beta-\widehat{\alpha}^*)\\ &\le (1+t)\Delta_2+2\sqrt{\Delta_2}\sqrt{C_{22}\Delta_1\vee\Delta_2}\\ &+\frac{2}{\sigma^2}\|\Sigma^{1/2}_{T'\,|\,W}(\alpha_\beta-\widehat{\alpha}^*)\|\|U^\top_{T'\,|\,W}\widetilde{\epsilon}'/(n-(s-r))\|\,.\end{aligned}$$

Denote our target $x := \|\Sigma^{1/2}_{T'\,|\,W}(\alpha_\beta-\widehat{\alpha}^*)/\sigma\|$, then

$$x^2 \le \frac{1+t}{1-t}\Delta_2+\frac{2}{1-t}\sqrt{\Delta_2}\sqrt{C_{22}\Delta_1\vee\Delta_2}+\frac{2}{1-t}\sqrt{C_{22}\Delta_1\vee\Delta_2}x\,.$$

After rearrangement, since both t and $C_{22}<1$,

$$\begin{aligned}\left(x-\frac{1}{1-t}\sqrt{C_{22}\Delta_1\vee\Delta_2}\right)^2 &\le \frac{1+t}{1-t}\Delta_2+\frac{2}{1-t}C^{1/2}_{22}\Delta_1\vee\Delta_2+\frac{C_{22}}{(1-t)^2}\Delta_1\vee\Delta_2\\ &\le \frac{3\times 2}{(1-t)^2}\Delta_1\vee\Delta_2\\ \implies x &\le (\frac{\sqrt{6}}{1-t}+\frac{1}{1-t}\sqrt{C_{22}})\sqrt{\Delta_1\vee\Delta_2}\\ &\le \frac{4}{1-t}\sqrt{\Delta_1\vee\Delta_2}\,.\end{aligned}$$

Therefore,

$$B_2 \le 2\times\sqrt{C_{22}\Delta_1\vee\Delta_2}\times\frac{4}{1-t}\sqrt{\Delta_1\vee\Delta_2}=\frac{8}{1-t}\sqrt{C_{22}}\Delta_1\vee\Delta_2\,,$$

with probability at least

$$1-\exp\left(-(n-(s-r))\frac{C_{22}\min(\Delta_1\vee\Delta_2,1)}{8(1+t)}+\frac{r}{4}\right)-\exp(-C(n-(s-r))t^2+r)\,.$$

Furthermore, we have

$$\frac{\widehat{\Delta}_2(T) - \widehat{\Delta}_2(S_*)}{\sigma^2} \geq (1-t)\Delta_2 - (C_{21} + \frac{8\sqrt{C_{22}}}{1-t})\Delta_1 \vee \Delta_2\,,$$

with probability at least

$$\begin{aligned}
&1 - 2\exp(-(n-(s-r))\frac{C_{21}}{8}\Delta_1 \vee \Delta_2) \\
&\quad - \exp\bigg(-(n-(s-r))\frac{C_{22}\min(\Delta_1 \vee \Delta_2, 1)}{8(1+t)} + \frac{r}{4}\bigg) \\
&\quad - \exp(-C(n-(s-r))t^2 + r)\,.
\end{aligned}$$

Let $t = \frac{1}{3}, C_{21} = \frac{1}{8}, C_{22} = \frac{1}{96^2}$, we get

$$\begin{aligned}
&\mathbb{P}\bigg(\frac{\widehat{\Delta}_2(T) - \widehat{\Delta}_2(S_*)}{\sigma^2} \geq \frac{2}{3}\Delta_2 - \frac{1}{4}\Delta_1 \vee \Delta_2\bigg) \\
\geq &1 - 4\exp\bigg(-(n-(s-r))\min(\Delta_1 \vee \Delta_2) \times \min(\frac{C}{9}, \frac{1}{96^2 \times 32/3}) + r\bigg) \\
= &1 - 4\exp\bigg(-(n-(s-r))C'\min(\Delta_1 \vee \Delta_2) + r\bigg)\,,
\end{aligned}$$

where $C' = \min(\frac{C}{9}, \frac{1}{96^2 \times 32/3})$. □

C.14 Proof of Lemma C.7.2

Proof of Lemma C.7.2. Let

$$\Delta_2(S, T, \alpha_{T\setminus S}) := \Big(\alpha_\beta(S,T) - \alpha_{T\setminus S}\Big)^\top \Sigma_{T\setminus S \,|\, S\cap T}\Big(\alpha_\beta(S,T) - \alpha_{T\setminus S}\Big)/\sigma^2$$

Therefore,

$$
\begin{aligned}
&\operatorname{var}\left[X_{S\setminus T}^\top \beta_{S\setminus T} - X_{T\setminus S}^\top \alpha_{T\setminus S} \,|\, S\cap T\right] \\
&= \beta_{S\setminus T}^\top \Sigma_{S\setminus T \,|\, S\cap T} \beta_{S\setminus T} + \alpha_{T\setminus S}^\top \Sigma_{T\setminus S \,|\, S\cap T} \alpha_{T\setminus S} \\
&\quad - 2\beta_{S\setminus T}^\top \Sigma_{(S\setminus T)(T\setminus S) \,|\, S\cap T} \alpha_{T\setminus S} \\
&= \beta_{S\setminus T}^\top \Sigma_{S\setminus T \,|\, T} \beta_{S\setminus T} \\
&\quad + \beta_{S\setminus T}^\top \Sigma_{(S\setminus T)(T\setminus S) \,|\, S\cap T} \Sigma_{T\setminus S \,|\, S\cap T}^{-1} \Sigma_{(T\setminus S)(S\setminus T) \,|\, S\cap T} \beta_{S\setminus T} \\
&\quad + \alpha_{T\setminus S}^\top \Sigma_{T\setminus S \,|\, S\cap T} \alpha_{T\setminus S} - 2\beta_{S\setminus T}^\top \Sigma_{(S\setminus T)(T\setminus S) \,|\, S\cap T} \alpha_{T\setminus S} \\
&= \Delta_1(S,T)\sigma^2 + \Delta_2(S,T,\alpha_{T\setminus S})\sigma^2 \,.
\end{aligned}
$$

Note that

$$
\frac{1}{2}\Big(\Delta_1(S,T) + \Delta_2(S,T)\Big) \le \Delta_1(S,T) \vee \Delta_2(S,T) \le \Delta_1(S,T) + \Delta_2(S,T)\,,
$$

with

$$
\begin{aligned}
\Delta_1(S,T) + \Delta_2(S,T) &= \Delta_1(S,T) + \min_{\alpha_{T\setminus S} \in \Theta_{T\setminus S}} \Delta_2(S,T,\alpha_{T\setminus S}) \\
&= \frac{1}{\sigma^2} \min_{\alpha_{T\setminus S} \in \Theta_{T\setminus S}} \operatorname{var}\left[X_{S\setminus T}^\top \beta_{S\setminus T} - X_{T\setminus S}^\top \alpha_{T\setminus S} \,|\, S\cap T\right].
\end{aligned}
$$

□

APPENDIX D

AUXILIARY LEMMAS

D.1 Concentration of χ^2 Random Variable

We will employ the tail probability bounds for χ^2 distribution [Laurent and Massart, 2000].

Lemma D.1.1. *If $Z \sim \chi^2_m$ with degree m, then for any $t \geq 0$,*

$$\mathbb{P}\left[\frac{Z-m}{m} \geq 2(\sqrt{t}+t)\right] \leq \exp(-mt)$$
$$\mathbb{P}\left[\frac{Z-m}{m} \leq -2\sqrt{t}\right] \leq \exp(-mt)\,.$$

Especially, we work with the following concentration bounds.

Lemma D.1.2. *If $Z \sim \chi^2_m$ with degree m, then for any $t \geq 0$,*

$$\mathbb{P}\left[\frac{|Z-m|}{m} \geq 4t\right] \leq \exp(-m\min(t,t^2))\,.$$

Proof. If $t \geq 1$, then $2(\sqrt{t}+t) \leq 4t$, $-4t \leq -2t \leq -2\sqrt{t}$, thus

$$\mathbb{P}\left[\frac{Z-m}{m} \geq 4t\right] \leq \mathbb{P}\left[\frac{Z-m}{m} \geq 2(\sqrt{t}+t)\right] \leq \exp(-mt)$$
$$\mathbb{P}\left[\frac{Z-m}{m} \leq -4t\right] \leq \mathbb{P}\left[\frac{Z-m}{m} \leq -2\sqrt{t}\right] \leq \exp(-mt)\,.$$

If $t \in [0,1)$, let $h = t^2 \in [0,1)$, then $2(\sqrt{h}+h) \leq 4\sqrt{h}$, $-4\sqrt{h} \leq -2\sqrt{h}$, thus

$$\begin{aligned}
&\mathbb{P}\left[\frac{Z-m}{m} \geq 4t\right] = \mathbb{P}\left[\frac{Z-m}{m} \geq 4\sqrt{h}\right] \\
&\leq \mathbb{P}\left[\frac{Z-m}{m} \geq 2(\sqrt{h}+h)\right] \leq \exp(-mh) = \exp(-mt^2) \\
&\mathbb{P}\left[\frac{Z-m}{m} \leq -4t\right] = \mathbb{P}\left[\frac{Z-m}{m}\right. \\
&\left.\leq -4\sqrt{h}\right] \leq \mathbb{P}\left[\frac{Z-m}{m} \leq -2\sqrt{h}\right] \leq \exp(-mh) = \exp(-mt^2)\,.
\end{aligned}$$

□

D.2 Lower Bound Techniques

For lower bound techniques, we mainly apply the Fano's inequality and Tsybokov's method.

Lemma D.2.1 (Yu [1997], Lemma 3)**.** *For a model family $\mathcal{M}$ contains M many distributions indexed by $j = 1, 2, \ldots, M$ such that*

$$\begin{aligned}
\alpha &= \max_{P_j \neq P_k \in \mathcal{M}} \mathbf{KL}(P_j \| P_k) \\
s &= \min_{P_j \neq P_k \in \mathcal{M}} \mathbf{dist}(\theta(P_j), \theta(P_k))\,,
\end{aligned}$$

where θ is a functional of its distribution argument. Then for any estimator $\widehat{\theta}$ for $\theta(P)$,

$$\inf_{\widehat{\theta}} \sup_{P \in \mathcal{M}} \mathbb{E}_P \mathbf{dist}(\theta(P), \widehat{\theta}) \geq \frac{s}{2}\left(1 - \frac{\alpha + \log 2}{\log M}\right).$$

Lemma D.2.2 (Tsybakov [2008], Theorem 2.5)**.** *For a model family $\mathcal{M}$ contains distributions $P_0, P_1, \ldots, P_M$ with $M \geq 2$ and suppose that Θ contains elements $\theta_0, \theta_1, \ldots, \theta_M$ such that:*

1. $\mathbf{dist}(\theta_j, \theta_k) \geq 2s$, $\forall 0 \leq j < k \leq M$;

2. $P_j \ll P_0, \forall j = 1, \ldots, M$, *and*

$$\frac{1}{M}\sum_{j=1}^{M} \mathbf{KL}(P_j, P_0) \leq \alpha \log M$$

with $0 < \alpha < 1/8$ *and* $P_j = P_{\theta_j}$. *Then*

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{P}_\theta\Big(\mathbf{dist}(\theta, \widehat{\theta}) \geq s\Big) \geq \frac{\sqrt{M}}{1+\sqrt{M}}\left(1 - 2\alpha - \frac{2\alpha}{\log M}\right).$$

Set the parameter of interest $\theta(P_j) = j$ to be the model index, and distance between parameters to be $\mathbf{dist}(\cdot, \cdot) = \mathbf{1}\{\cdot \neq \cdot\}$, consider P_j to be a product measure of n i.i.d. samples for any $P_j \in \mathcal{M}$, then Lemma D.2.1 and D.2.2 under model selection context can be stated as follows:

Corollary D.2.3 (Fano's inequality)**.** *For a model family* $\mathcal{M}$ *contains* M *many distributions indexed by* $j = 1, 2, \ldots, M$ *such that* $\alpha = \max_{P_j \neq P_k \in \mathcal{M}} \mathbf{KL}(P_j \| P_k)$. *If the sample size is bounded as*

$$n \leq \frac{(1-2\delta)\log M}{\alpha},$$

then for any estimator $\widehat{\theta}$ *for the model index:*

$$\inf_{\widehat{\theta}} \sup_{j \in [M]} P_j(\widehat{\theta} \neq j) \geq \delta - \frac{\log 2}{\log M}.$$

Corollary D.2.4 (Tsybakov's method)**.** *For a model family* $\mathcal{M}$ *contains* M *many distributions indexed by* $j = 1, 2, \ldots, M$ *such that* $\alpha = \max_{j \in [M]} \mathbf{KL}(P_j \| P_0)$. *If the sample size is bounded as*

$$n \leq \frac{\log M}{16\alpha},$$

then for any estimator $\widehat{\theta}$ for the model index:

$$\inf_{\widehat{\theta}} \sup_{j \in [M]} P_j(\widehat{\theta} \neq j) \geq \frac{1}{16}.$$

The model in the context of structure learning is the underlying graph G (or the CPDAG $\overline{G}$); in the context of variable selection is the true support S_*.

APPENDIX E

EXPERIMENT DETAILS

Here we provide full details of our experiments along with additional experiments to compare KL-BSS and Vanilla KL-BSS (Appendix E.6). Finally, Appendix E.5 summarizes the results from all the simulation setups.

E.1 Simulation Setup

For graph types, we generate:

- *Erdös-Rényi (ER)*. Graphs whose edges are selected from all possible $\binom{d}{2}$ edges independently with specified expected number of edges;
- *Scale-Free network (SF)*. Graphs simulated according to the Barabasi-Albert model;
- *Bipartite graph*. Generated as follows:
  1. Randomly divide $[d]$ into V_1 and V_2;
  2. Let $\widetilde{s} = \min\{s, |V_1|\}$;
  3. For each $j \in V_2$, randomly sample the number of parents $|\mathrm{pa}(j)|$ from $[\widetilde{s}]$;
  4. Randomly sample $|\mathrm{pa}(j)|$ many nodes from V_1 to be $\mathrm{pa}(j)$;
  5. Randomly permute the nodes.
- *Complete graph*. Graphs with all possible $\binom{d}{2}$ edges. Nodes are randomly permuted.

We generated graphs from ER and SF with $\{d, 2d, 4d\}$ edges each, which are denoted as XX-k where XX $\in \{\mathrm{ER}, \mathrm{SF}\}$ denotes graph type and k denotes the average number of edges (i.e. expected total number of edges is kd).

Given the DAG G, the data (X, Y) is then generated by

$$X_k = \sum_{j \in \mathrm{pa}_G(k)} b_{jk} X_j + \epsilon_k$$
$$Y = \sum_{\ell \in S_*} \beta_\ell X_\ell + \epsilon\,,$$

with S_* randomly sampled from $[d]$, $\beta_\ell = \mathrm{Rad}_\ell \times \beta_{\min}$, $b_{jk} \sim \mathrm{Rad}_{jk} \times \mathrm{Unif}(b_{\min}, b_{\max})$ where $\mathrm{Rad}_\ell, \mathrm{Rad}_{jk}$ are independent Rademacher random variables and $b_{\max} = 5, \beta_{\min} = b_{\min} = 0.1$. No effort is made to enforce our assumptions, to avoid path cancellation, etc.

For the noise distributions, we consider different distributions centered at zero: {Gaussian, t, Uniform, Laplace}. Scale the random variable such that $\mathrm{var}(\epsilon_k) = \sigma_k^2$ and $\mathrm{var}(\epsilon) = \sigma^2$. Let $\sigma = 1$, $\sigma_k \sim \mathrm{Unif}(\sigma_{\min}, \sigma_{\max})$ with $\sigma_{\min} = 0.5, \sigma_{\max} = 1$. We also consider "mixed" noise distributions where for each of ϵ_k and ϵ, we randomly choose one distribution from the four above and sample from it.

Finally, we generated random datasets with sample size $n \in \{1000, 2000, \ldots, 8000\}$ for number of nodes $d \in \{8, 9, 10\}$ and sparsity level $s \in \{2, 3, 4\}$. For $d = 50$, we consider $s \in \{10, 15, 20\}$, and set $b_{\max} = 2$ instead of 5 to avoid numerical issues. Especially, we set $b_{\max} = 1$ for Complete graph. We implement both KL-BSS and BSS using the MIP detailed in Section 4.3.3 with $M = 10$. We use `Gurobi` with tolerance parameter set to be `MIPGap=1e-9`. For the Lasso, we use `sklearn` with `n_alphas=500`.

For unknown sparsity, we consider three additive penalties τ given by:

- *BIC*: $\tau = \frac{\log n}{n}$;
- *EBIC*: $\tau = \frac{\log d}{n}$;
- *Delta*: $\tau = \frac{\sigma_{\min}^2 \beta_{\min}^2}{4}$.

The last one (*Delta*) is the theoretical choice given in Gao et al. [2025a] when the model satisfies the optimality condition, which is realized as $\sigma_{\min}^2 \beta_{\min}^2 / 4 = 0.5^2 \times 0.1^2 / 4$ in

our setup. For unknown $\beta_{\min}$, we consider candidate choices for $\widetilde{\beta}_{\min}$: $\{\beta_{\min}^{\ell}\}_{\ell=1}^{L} = 10^{-2.4,-2.2,-2,\ldots,0.2,0.4,0.6}$ with $K = 5$-fold CV.

We run the experiments for the same data generating process in the low dimensions ($d \in \{8, 9, 10\}$) with $\bar{s} = 4$; and high dimensions ($d = 50$) with $\bar{s} = 25$.

For evaluation, we consider four metrics based on the estimated support $\widehat{S}$:

- *Recovery probability*: $\mathbb{1}\{\widehat{S} = S_*\}$;
- *Hamming distance*: $|S_* \setminus \widehat{S}| + |\widehat{S} \setminus S_*|$;
- *False discovery rate*: $\frac{|\widehat{S} \setminus S_*|}{|\widehat{S}|}$;
- *True positive rate*: $\frac{|\widehat{S} \cap S_*|}{|S_*|}$.

The results are reported by average over $N = 200$ replications.

E.2 Real Data Application

E.2.1 Selection of Genes

We start by removing all the genes with variances smaller than 0.01; then select the top 25 genes based on their marginal variances and group the remaining genes according to their variances in ascending order into $d - 25$ bins. Then for each replication, we randomly sample one gene from each bin to form the X (of dimension d). We consider $d = 50, 60, 70, 80, 90$ with $s = 10$. Let $S_* = (2, 4, 6, \ldots, 18, 20)$. We randomly shuffle the rows of X, center every gene, and let

$$Y = X\beta + \epsilon$$

where $\beta_j = \beta_{\min} \times \mathrm{Rad}_j$, Rad_j are i.i.d. Rademacher random variables, $\epsilon_i \sim \mathcal{N}(0, 1)$, and $\beta_{\min} = 0.1$. Apply KL-BSS and BSS on this data and evaluate the performance for

$n = 200, 200, \ldots, 800$. The results are reported as average of $N = 200$ replications.

E.2.2 Prediction Performance

We pick the gene with the largest variance as Y, then the remaining genes as candidate X's to explain Y. For each replication, we randomly choose $d = 50$ genes, and randomly split the dataset in training set $\mathcal{D}_0$ and test set $\mathcal{D}_1$. Apply KL-BSS and BSS with $s = 10$ on the training set to estimate the support $\widehat{S}$ and coefficients $\widehat{\beta}_{\widehat{S}}$. Finally, evaluate the estimate using prediction error on the test set:

$$\frac{1}{n_1} \sum_{i \in \mathcal{D}_1} (Y_i - X_{i\widehat{S}}^\top \widehat{\beta}_{\widehat{S}})^2 .$$

For both BSS and KL-BSS, we use the MIP implementation for this exercise. In particular, we apply the method in Section 4.3.3 for the choice of $\beta_{\min}$ for KL-BSS. We run for $N = 100$ replications.

E.3 Structure Learning

We perform experiments to compare the performance of BSS and KL-BSS on structure learning. We take the same experiment setup described in Appendix E.1 with ER/SF-2 graphs to generate X. We set $\sigma_k \equiv 2$ for all k. The learning produce is as follows: Initialize $\widehat{G}$ as an empty graph. Given G and data X, we either take one valid topological ordering of G, or apply EqVar algorithm [Chen et al., 2019] for ordering estimation, denoted as π (a permutation of $[d]$). For each $k = 2, 3 \ldots, d$, we apply BSS/KL-BSS with unknown sparsity and $\bar{s} = \deg(G) + 1$ and BIC penalty to conduct variable selection for X_{π_k} from $X_{\pi_{[1:k-1]}}$, which is used as estimate for parents of π_k in $\widehat{G}$. Finally, we evaluate the performances by structural Hamming distance (SHD) between G and $\widehat{G}$.

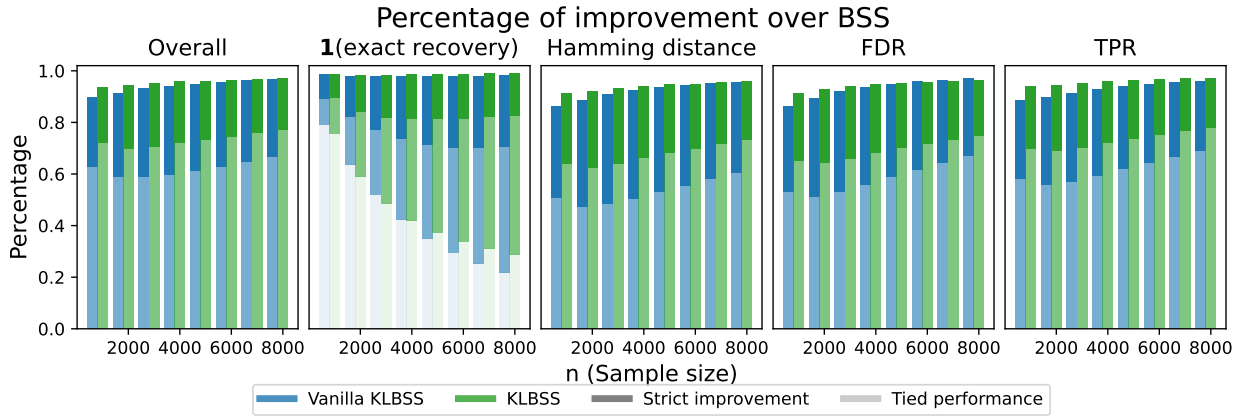


Figure E.1: Detailed comparison of KL-BSS vs. BSS. Overall evaluation by percentage of improvement over BSS on a per-dataset basis (i.e. not averaged for all replications in each setting). The solid bars indicate percentage of strict improvement. The transparent bars above the solid bars indicate the percentage of tied performance. For exact variable selection metric (the second panel from left), if representing the outcomes of KL-BSS and BSS as a tuple, we count $(1,0)$ as strict improvement, $(1,1)$ as tied performance, and the most transparent bars indicate percentage of $(0,0)$.

E.4 Additional Metrics

For comparison, at the end of this appendix we have included results for other metrics for the setups in Figure 4.2 as described in Appendix E.1.

- Figure E.3: Hamming distance results for ER-4, SF-4 and Complete graph types, known and unknown sparsity, $(d, s, \bar{s}) = (10, 3, 4)$ and $(50, 10, 25)$;
- Figure E.4: FDR results for ER-4, SF-4 and Complete graph types, known and unknown sparsity, $(d, s, \bar{s}) = (10, 3, 4)$ and $(50, 10, 25)$;
- Figure E.5: TPR results for ER-4, SF-4 and Complete graph types, known and unknown sparsity, $(d, s, \bar{s}) = (10, 3, 4)$ and $(50, 10, 25)$.

E.5 Overall Comparison

To visualize the improvement over BSS and summarize results from our comprehensive experiments in a succinct way, we use the percentage of runs where KL-BSS outputs a strictly better result (across four different metrics) over BSS across all replications (simulated datasets) in all experiment setups for evaluation. For each sample size n and each replication, we compute the metric (successful recovery indicator, Hamming distance, FDR and TPR) and calculate the percentages of KL-BSS (and Vanilla KL-BSS) giving tied and better metric against BSS averaged over all datasets, whose total number is 200 (number of replications per setup) $\times$ 8 (number of graph types) $\times$ 5 (number of noise distributions) $\times$ 12 (number of pairs of (d, s), 9 for KL-BSS since not implemented for $d = 50$) $\times$ 4 (known and unknown sparsity with BIC/EBIC/Delta penalty) $= 384,000$. The result is shown in Figure E.1, where we compute the overall percentage and also separately for each metric. The percentages of strict improvement and tied performance are plotted via solid and transparent bars, respectively. In particular, for exact variable selection metric, if the outcomes are denoted as a tuple $(\mathbb{1}\{\widehat{S}^{\text{KL-BSS}} = S_*\}, \mathbb{1}\{\widehat{S}^{\text{BSS}} = S_*\})$, then we count $(1, 0)$ as strict improvement, $(1, 1)$ as tied performance, and the most transparent bars (bottom) indicate the percentage of $(0, 0)$. We can see KL-BSS strictly improves BSS on around 20%-30% of the simulated datasets, and gives equivalent performance as BSS for nearly all of the rest datasets.

E.6 Vanilla KL-BSS and KL-BSS

In this appendix, we use random SF graphs to showcase a setting where KL-BSS has advantage over Vanilla KL-BSS. Vanilla KL-BSS requires a factor of s in the sample complexity (Theorem C.7.4), which comes from the covariance matrix estimation. In the cases where (certain submatrices of) $\Sigma = \text{cov}(X)$ are hard to estimate, KL-BSS will give better

performance. Here we consider to increase the upper limit of the linear coefficients in generating X, and remove the Rademacher random multiplier before them (Appendix E.1), which means the linear coefficients are always positive and makes X's highly asymmetric in their (co)variances, and the maximum eigenvalue of Σ is large.

Concretely, consider the following experiment setup: $d = 7, s = 3$ and an SF-2 graph. The only difference is $b_{jk} \sim \text{Unif}(b_{\min}, b_{\max})$ are all positive with $b_{\min} = 0.1, b_{\max} = 15$, and $\sigma_{\max} = 2$. The results are shown in Figure E.2, from which we can see the slightly better performance of KL-BSS compared to Vanilla KL-BSS. This is an example of a hard instance that encapsulates the *minimax* (i.e. worst-case) behaviour in neighbourhood selection, but is not indicative of the *average* (similarly, pointwise) behaviour, where Vanilla KL-BSS is better.

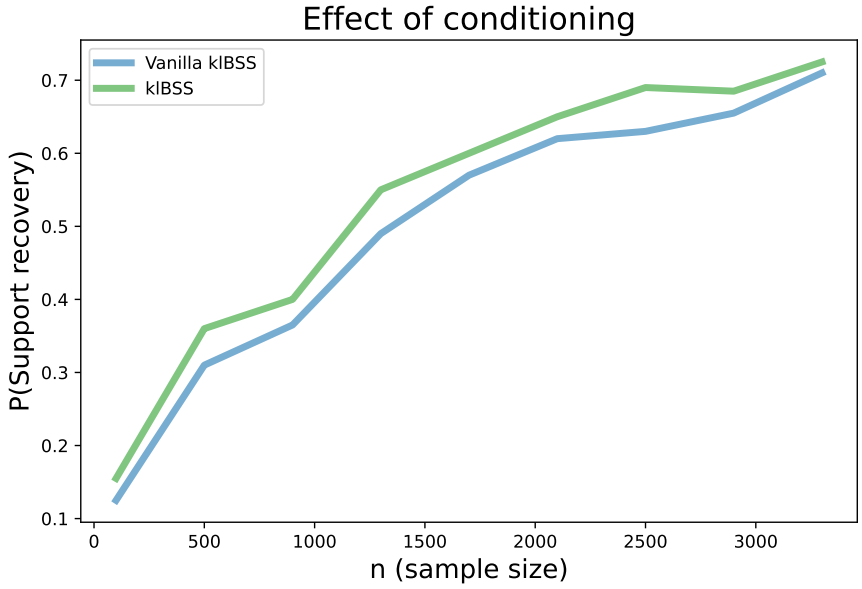


Figure E.2: Comparison between KL-BSS and Vanilla KL-BSS on covariance matrix with large maximum eigenvalue. KL-BSS gives better performance when the covariance (sub)matrix is relatively harder to estimate.

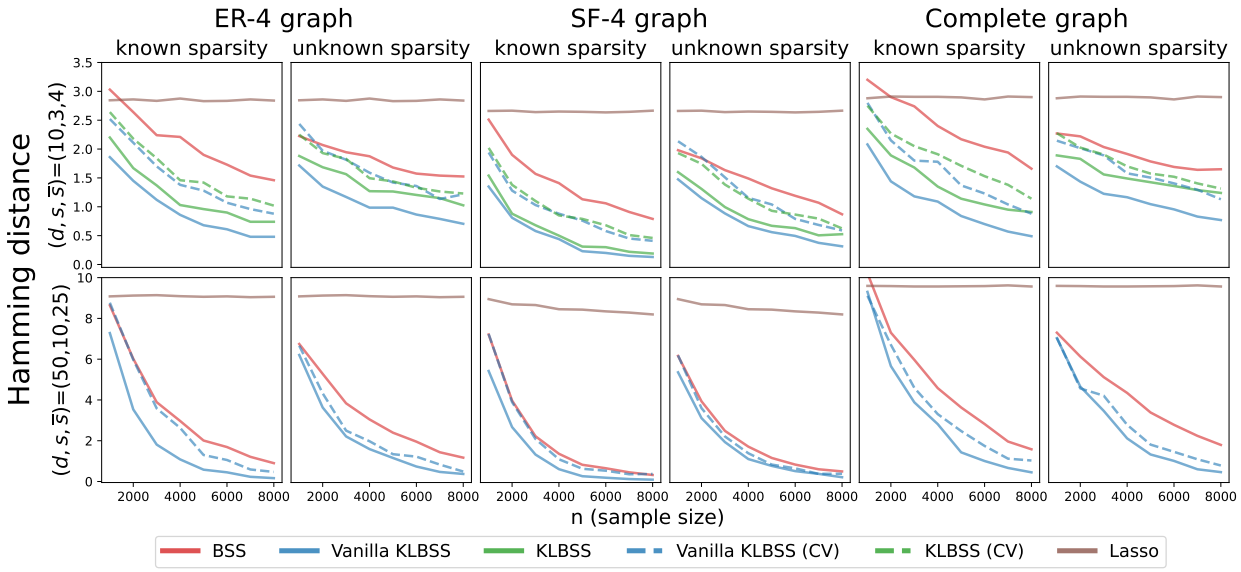


Figure E.3: Experiment results of Hamming distance for the same setups with Figure 4.2.

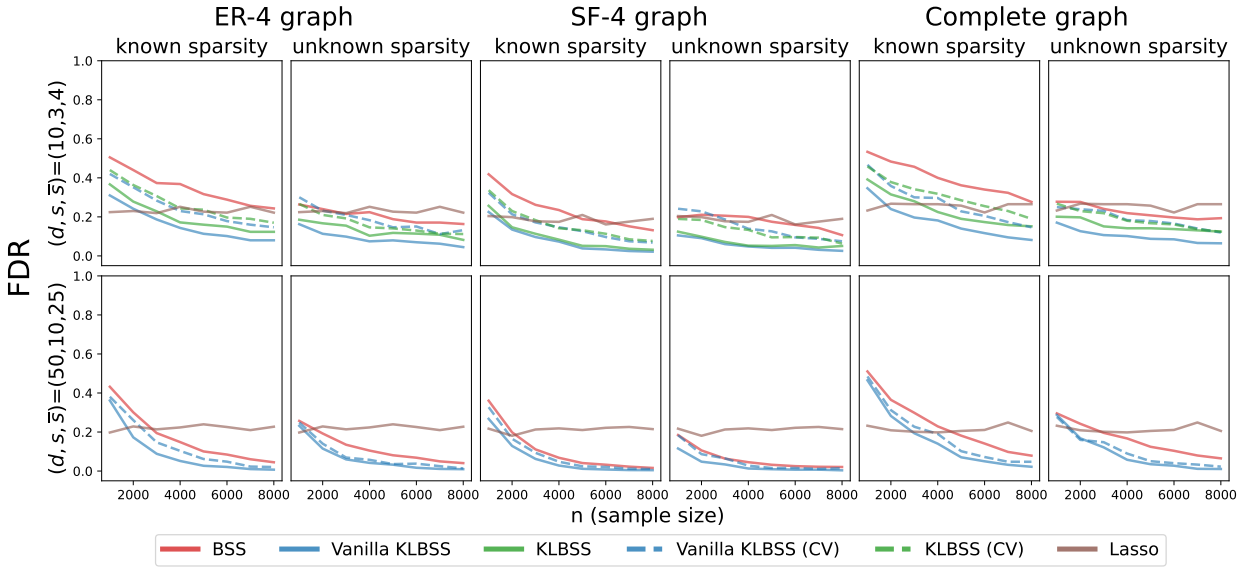


Figure E.4: Experiment results of FDR for the same setups with Figure 4.2.

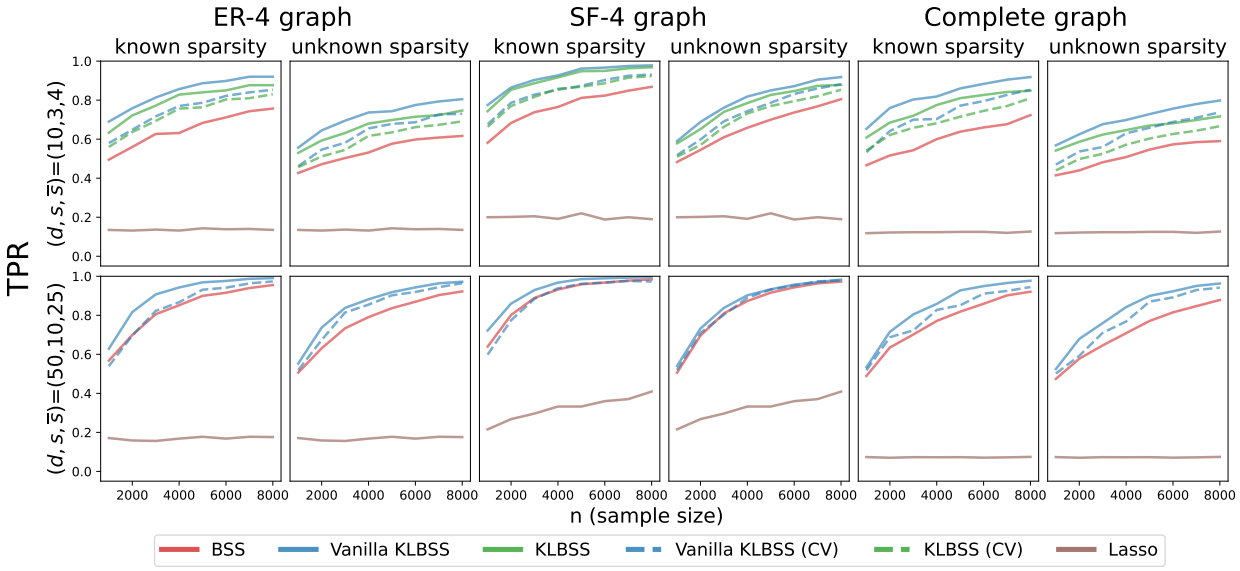


Figure E.5: Experiment results of TPR for the same setups with Figure 4.2.

BIBLIOGRAPHY

Emmanuel Abbe. Community detection and stochastic block models: recent developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018.

Shuchin Aeron, Venkatesh Saligrama, and Manqi Zhao. Information theoretic bounds for compressed sensing. *IEEE Transactions on Information Theory*, 56(10):5111–5130, 2010.

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

Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 2003.

Mehmet Akçakaya and Vahid Tarokh. Shannon-theoretic limits on noisy compressive sampling. *IEEE Transactions on Information Theory*, 56(1):492–504, 2009.

Cem Aksoylar, George K Atia, and Venkatesh Saligrama. Sparse signal processing with linear and nonlinear observations: A unified shannon-theoretic approach. *IEEE Transactions on Information Theory*, 63(2):749–776, 2016.

Constantin F Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D Koutsoukos. Local causal and markov blanket induction for causal discovery and feature selection for classification part i: algorithms and empirical evaluation. *Journal of Machine Learning Research*, 11(1), 2010.

Gökmen Altay and Frank Emmert-Streib. Inferring the conservative causal core of gene regulatory networks. *BMC systems biology*, 4(1):1–13, 2010.

Theodore W Anderson. *An introduction to multivariate statistical analysis*, volume 2. Wiley New York, 1958.

Steen A Andersson, David Madigan, and Michael D Perlman. A characterization of markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25(2):505–541, 1997.

Sanjeev Arora and Boaz Barak. *Computational complexity: a modern approach*. Cambridge University Press, 2009.

Mona Azadkia, Armeen Taeb, and Peter Bühlmann. A fast non-parametric approach for local causal structure learning. *arXiv preprint arXiv:2111.14969*, 2021.

Afonso S Bandeira, Ahmed El Alaoui, Samuel Hopkins, Tselil Schramm, Alexander S Wein, and Ilias Zadik. The franz-parisi criterion and computational trade-offs in high dimensional statistics. *Advances in Neural Information Processing Systems*, 35:33831–33844, 2022.

Yannick Baraud. Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, pages 577–606, 2002.

Pierre C Bellec, Guillaume Lecué, and Alexandre B Tsybakov. Slope meets lasso: improved oracle bounds and optimality. *The Annals of Statistics*, 46(6B):3603–3642, 2018.

Thomas B Berrett and Richard J Samworth. Nonparametric independence testing via mutual information. *Biometrika*, 106(3):547–566, 2019.

Quentin Berthet and Philippe Rigollet. Optimal detection of sparse principal components in high dimension. *The Annals of Statistics*, 41(4):1780, 2013.

Dimitris Bertsimas and Bart Van Parys. Sparse high-dimensional regression. *The Annals of Statistics*, 48(1):300–323, 2020.

Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. *The annals of statistics*, 44(2):813–852, 2016.

Guy Bresler. Efficiently learning ising models on arbitrary graphs. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 771–782, 2015.

Rares-Darius Buhai, Jingqiu Ding, and Stefan Tiegel. Computational-statistical gaps for improper learning in sparse linear regression. In Shipra Agrawal and Aaron Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 752–771. PMLR, 30 Jun–03 Jul 2024. URL `https://proceedings.mlr.press/v247/buhai24a.html`.

Peter Bühlmann, Markus Kalisch, and Marloes H Maathuis. Variable selection in high-dimensional linear models: partially faithful distributions and the pc-simple algorithm. *Biometrika*, 97(2):261–278, 2010.

Peter Bühlmann, Jonas Peters, and Jan Ernest. Cam: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.

Cristina Butucea, Mohamed Ndaoud, Natalia A Stepanova, and Alexandre B Tsybakov. Variable selection with hamming loss. *The Annals of Statistics*, 46(5):1837–1875, 2018.

T Tony Cai and Zongming Ma. Optimal hypothesis testing for high dimensional covariance matrices. *Bernoulli*, 19(5B):2359–2388, 2013.

T Tony Cai and Lie Wang. Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Transactions on Information theory*, 57(7):4680–4688, 2011.

T Tony Cai, Zongming Ma, and Yihong Wu. Sparse pca: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(237):3074–3110, 2013.

Tony Cai, Weidong Liu, and Xi Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.*, 35(1):2313–2351, 2007.

Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.

Clément L Canonne. *Topics and techniques in distribution testing*. now Publishers, 2022.

Clément L Canonne, Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Testing conditional independence of discrete distributions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 735–748, 2018.

Siu-On Chan, Ilias Diakonikolas, Paul Valiant, and Gregory Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 1193–1203. SIAM, 2014.

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

David Maxwell Chickering. Learning bayesian networks is np-complete. *Learning from data: Artificial intelligence and statistics V*, pages 121–130, 1996.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Max Chickering. Statistically efficient greedy equivalence search. In *Conference on Uncertainty in Artificial Intelligence*, pages 241–249. PMLR, 2020.

Max Chickering, David Heckerman, and Chris Meek. Large-sample learning of bayesian networks is np-hard. *Journal of Machine Learning Research*, 5:1287–1330, 2004.

C Chow and T Wagner. Consistency of an estimate of tree-dependent probability distributions (corresp.). *IEEE Transactions on Information Theory*, 19(3):369–371, 1973.

CKCN Chow and Cong Liu. Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14(3):462–467, 1968.

Laëtitia Comminges and Arnak S Dalalyan. Tight conditions for consistency of variable selection in the context of high dimensionality. *The Annals of Statistics*, 40(5):2667–2696, 2012.

John N Darroch, Steffen L Lauritzen, and Terry P Speed. Markov fields and log-linear interaction models for contingency tables. *The Annals of Statistics*, pages 522–539, 1980.

Sanjoy Dasgupta. Learning polytrees. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 134–141, 1999.

Constantinos Daskalakis, Vardis Kandiros, and Rui Yao. Learning gaussian dag models without condition number bounds. In *Forty-second International Conference on Machine Learning*, 2025.

A Philip Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 41(1):1–15, 1979.

Luc Devroye. The equivalence of weak, strong and complete convergence in l1 for kernel density estimates. *The Annals of Statistics*, pages 896–904, 1983.

Ilias Diakonikolas and Daniel M Kane. A new approach for testing properties of discrete distributions. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 685–694. IEEE, 2016.

Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 73–84. IEEE, 2017.

NR Draper. *Applied regression analysis*. McGraw-Hill. Inc, 1998.

Mathias Drton. Algebraic problems in structural equation modeling. In *The 50th Anniversary of Gröbner Bases*, volume 77, pages 35–87. Mathematical Society of Japan, 2018.

Mathias Drton and Marloes H Maathuis. Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4:365–393, 2017.

Michael Alin Efroymson. Multiple regression analysis. *Mathematical methods for digital computers*, pages 191–203, 1960.

Felix Elwert. Graphical causal models. In *Handbook of causal analysis for social research*, pages 245–273. Springer, 2013.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(5): 849–911, 2008.

Samuele Fiorini. gene expression cancer RNA-Seq. UCI Machine Learning Repository, 2016. DOI: https://doi.org/10.24432/C5R88H.

Ronald A Fisher. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521, 1915.

Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936a.

Ronald Aylmer Fisher. Design of experiments. *British Medical Journal*, 1(3923):554, 1936b.

Ronald Aylmer Fisher. Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*, pages 66–70. Springer, 1970.

Alyson K Fletcher, Sundeep Rangan, and Vivek K Goyal. Necessary and sufficient conditions for sparsity pattern recovery. *IEEE Transactions on Information Theory*, 55(12): 5758–5772, 2009.

Dean Foster, Howard Karloff, and Justin Thaler. Variable selection is hard. In *Conference on Learning Theory*, pages 696–709. PMLR, 2015.

Dean P Foster and Edward I George. The risk inflation criterion for multiple regression. *The Annals of Statistics*, 22(4):1947–1975, 1994.

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

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

Nir Friedman, Iftach Nachman, and Dana Pe'er. Learning bayesian network structure from massive datasets: The" sparse candidate" algorithm. *arXiv preprint arXiv:1301.6696*, 2013.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. *Advances in neural information processing systems*, 20, 2007.

Ming Gao and Bryon Aragam. Efficient bayesian network structure learning via local markov boundary search. *Advances in Neural Information Processing Systems*, 34:4301–4313, 2021.

Ming Gao and Bryon Aragam. Optimality and computational barriers in variable selection under dependence. *Submitted*, 2025. URL `https://arxiv.org/abs/2510.03990`.

Ming Gao, Yi Ding, and Bryon Aragam. A polynomial-time algorithm for learning nonparametric causal graphs. *Advances in Neural Information Processing Systems*, 33:11599–11611, 2020.

Ming Gao, Wai Ming Tai, and Bryon Aragam. Optimal estimation of gaussian dag models. In *International Conference on Artificial Intelligence and Statistics*, pages 8738–8757. PMLR, 2022.

Ming Gao, Wai Ming Tai, and Bryon Aragam. KL-BSS: Rethinking optimality for neighbourhood selection in structural equation models. *arXiv preprint arXiv:2306.02244*, 2025a.

Ming Gao, Yuhao Wang, and Bryon Aragam. Optimal structure learning and conditional independence testing. *Submitted*, 2025b. URL `https://arxiv.org/abs/2507.05689`.

Tian Gao and Qiang Ji. Efficient markov blanket discovery and its application. *IEEE transactions on Cybernetics*, 47(5):1169–1179, 2016.

Christopher R Genovese, Jiashun Jin, Larry Wasserman, and Zhigang Yao. A comparison of the lasso and marginal regression. *The Journal of Machine Learning Research*, 13(1): 2107–2143, 2012.

Asish Ghoshal and Jean Honorio. Information-theoretic limits of bayesian network structure learning. In *Artificial Intelligence and Statistics*, pages 767–775. PMLR, 2017a.

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

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

Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. *Advances in neural information processing systems*, 20, 2007.

Yongtao Guan and Matthew Stephens. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics*, 5(3):1780, 2011.

F Richard Guo, Anton Rask Lundborg, and Qingyuan Zhao. Confounder selection: Objectives and approaches. *arXiv preprint arXiv:2208.13871*, 2022a.

Yilin Guo, Haolei Weng, and Arian Maleki. Signal-to-noise ratio aware minimaxity and higher-order asymptotics, 2022b.

Sumeet Gupta and Hee W Kim. Linking structural equation modeling to bayesian networks: Decision support for customer retention in virtual communities. *European Journal of Operational Research*, 190(3):818–833, 2008.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC, 2019.

Trevor Hastie, Robert Tibshirani, and Ryan Tibshirani. Best subset, forward stepwise or lasso? analysis and recommendations based on extensive comparisons. *Statistical Science*, 35(4):579–592, 2020.

Hussein Hazimeh and Rahul Mazumder. Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. *Operations Research*, 68(5):1517–1537, 2020.

David Heckerman, Dan Geiger, and David M Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20:197–243, 1995.

Georg Heinze, Christine Wallisch, and Daniela Dunkler. Variable selection–a review and recommendations for the practicing statistician. *Biometrical journal*, 60(3):431–449, 2018.

Ronald R Hocking and RN Leslie. Selection of the best subset in regression analysis. *Technometrics*, 9(4):531–540, 1967.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

C Terrance Ireland and Solomon Kullback. Contingency tables with given marginals. *Biometrika*, 55(1):179–188, 1968.

Martin Emil Jakobsen, Rajen D Shah, Peter Bühlmann, and Jonas Peters. Structure learning for directed trees. *The Journal of Machine Learning Research*, 23(1):7076–7172, 2022.

Fateme Jamshidi, Luca Ganassali, and Negar Kiyavash. On the sample complexity of conditional independence testing with von mises estimator with application to causal discovery. In *Forty-first International Conference on Machine Learning*, 2024. URL `https://openreview.net/forum?id=oSOZ31ISBV`.

Pengsheng Ji and Jiashun Jin. Ups delivers optimal phase diagram in high-dimensional variable selection. *The Annals of Statistics*, pages 73–103, 2012.

Markus Kalisch and Peter Bühlman. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(3), 2007.

Jonathan A Kelner, Frederic Koehler, Raghu Meka, and Dhruv Rohatgi. On the power of preconditioning in sparse linear regression. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 550–561. IEEE, 2022.

Ilmun Kim, Matey Neykov, Sivaraman Balakrishnan, and Larry Wasserman. Local permutation tests for conditional independence. *The Annals of Statistics*, 50(6):3388–3414, 2022.

Ilmun Kim, Matey Neykov, Sivaraman Balakrishnan, and Larry Wasserman. Conditional independence testing for discrete distributions: Beyond x 2-and g-tests. *Electronic Journal of Statistics*, 18(2):4767–4794, 2024.

JinHyung Kim and Judea Pearl. A computational model for causal and diagnostic reasoning in inference systems. In *International Joint Conference on Artificial Intelligence*, pages 0–0. Citeseer, 1983.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Sadanori Konishi and Genshiro Kitagawa. Generalised information criteria in model selection. *Biometrika*, 83(4):875–890, 1996.

Dmitriy Kunisky, Alexander S Wein, and Afonso S Bandeira. Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio. In *ISAAC Congress (International Society for Analysis, its Applications and Computation)*, pages 1–50. Springer, 2019.

Wai-Yin Lam, Bryan Andrews, and Joseph Ramsey. Greedy relaxations of the sparsest permutation algorithm. In *Uncertainty in Artificial Intelligence*, pages 1052–1062. PMLR, 2022.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

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

Chun Li and Xiaodan Fan. On nonparametric conditional independence tests for continuous variables. *Wiley Interdisciplinary Reviews: Computational Statistics*, 12(3):e1489, 2020.

Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10 (10), 2009.

Han Liu, Min Xu, Haijie Gu, Anupam Gupta, John Lafferty, and Larry Wasserman. Forest density estimation. *The Journal of Machine Learning Research*, 12:907–951, 2011.

Po-Ling Loh. Neighborhood selection methods. In *Handbook of Graphical Models*, pages 289–308. CRC Press, 2018.

Po-Ling Loh and Peter Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *The Journal of Machine Learning Research*, 15(1):3065–3105, 2014.

Po-Ling Loh and Martin J Wainwright. Support recovery without incoherence: A case for nonconvex regularization. *The Annals of Statistics*, 45(6):2455–2482, 2017.

Zongming Ma and Yihong Wu. Computational barriers in minimax submatrix detection. *The Annals of Statistics*, pages 1089–1116, 2015.

Marloes Maathuis, Mathias Drton, Steffen Lauritzen, and Martin Wainwright. *Handbook of graphical models*. CRC Press, 2018.

Colin L Mallows. Some comments on cp. *Technometrics*, 42(1):87–94, 2000.

Dimitris Margaritis and Sebastian Thrun. Bayesian network induction via local neighborhoods. *Advances in neural information processing systems*, 12, 1999.

Florian Markowetz and Rainer Spang. Inferring cellular networks–a review. *BMC bioinformatics*, 8(6):1–17, 2007.

Alexander Marx, Arthur Gretton, and Joris M Mooij. A weaker faithfulness assumption based on triple interactions. In *Uncertainty in Artificial Intelligence*, pages 451–460. PMLR, 2021.

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

Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1), 2009.

Alan J Miller. Selection of subsets of regression variables. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 147(3):389–410, 1984.

Sidhant Misra, Marc Vuffray, and Andrey Y Lokhov. Information theoretic optimal learning of gaussian graphical models. In *Conference on Learning Theory*, pages 2888–2909. PMLR, 2020.

Ankur Moitra and Alexander S Wein. Precise error rates for computationally efficient testing. *arXiv preprint arXiv:2311.00289*, 2023.

Rajeev Motwani. *Randomized Algorithms*. Cambridge University Press, 1995.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Preetam Nandy, Alain Hauser, Marloes H Maathuis, et al. High-dimensional consistency in score-based and hybrid structure learning. *The Annals of Statistics*, 46(6A):3151–3183, 2018.

Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.

Mohamed Ndaoud and Alexandre B Tsybakov. Optimal variable selection and adaptive noisy compressed sensing. *IEEE Transactions on Information Theory*, 66(4):2517–2532, 2020.

Matey Neykov, Sivaraman Balakrishnan, and Larry Wasserman. Minimax optimal conditional independence testing. *The Annals of Statistics*, 49(4):2151–2177, 2021.

Ryuei Nishii. Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics*, pages 758–765, 1984.

Ryuie Nishii. Maximum likelihood principle and model selection when the true model is unspecified. *Journal of Multivariate analysis*, 27(2):392–403, 1988.

Samet Oymak, Amin Jalali, Maryam Fazel, Yonina C Eldar, and Babak Hassibi. Simultaneously structured models with application to sparse and low-rank matrices. *IEEE Transactions on Information Theory*, 61(5):2886–2908, 2015.

Gunwoong Park. Learning generalized hypergeometric distribution (ghd) dag models. *arXiv preprint arXiv:1805.02848*, 2018.

Gunwoong Park. Identifiability of additive noise models using conditional variances. *The Journal of Machine Learning Research*, 21(1):2896–2929, 2020.

Gunwoong Park and Sion Park. High-dimensional poisson structural equation model learning via ℓ_1-regularized regression. *Journal of Machine Learning Research*, 20(95):1–41, 2019.

Gunwoong Park and Garvesh Raskutti. Learning quadratic variance function (QVF) dag models via overdispersion scoring (ODS). *The Journal of Machine Learning Research*, 18 (1):8300–8342, 2017.

Judea Pearl. Causal inference. *Causality: objectives and assessment*, pages 39–58, 2010.

Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversity-Press*, 19(2):3, 2000.

Karl Pearson. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

Jose M Pena, Roland Nilsson, Johan Björkegren, and Jesper Tegnér. Towards scalable and data efficient learning of markov boundaries. *International Journal of Approximate Reasoning*, 45(2):211–232, 2007.

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

Jonas Peters and Peter Bühlmann. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Identifiability of causal graphs using functional models. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 589–598, 2011.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1): 2009–2053, 2014.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Kamiar Rahnama Rad. Nearly sharp sufficient conditions on exact sparsity pattern recovery. *IEEE Transactions on Information Theory*, 57(7):4672–4679, 2011.

Goutham Rajendran, Bohdan Kivva, Ming Gao, and Bryon Aragam. Structure learning in polynomial time: Greedy algorithms, bregman information, and exponential families. *Advances in Neural Information Processing Systems*, 34:18660–18672, 2021.

Joseph D Ramsey. A scalable conditional independence test for nonlinear, non-gaussian data. *arXiv preprint arXiv:1401.5031*, 2014.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE transactions on information theory*, 57 (10):6976–6994, 2011.

G Rebane. The recovery of causal poly-trees from statistical data. *Uncertainty in Artificial Intelligence'87*, pages 222–228, 1987.

Galen Reeves and Michael Gastpar. Sampling bounds for sparse support recovery in the presence of noise. In *2008 IEEE International Symposium on Information Theory*, pages 2187–2191. IEEE, 2008.

Galen Reeves, Jiaming Xu, and Ilias Zadik. The all-or-nothing phenomenon in sparse linear regression. In *Conference on Learning Theory*, pages 2652–2663. PMLR, 2019.

Dominik Rothenhäusler, Jan Ernest, Peter Bühlmann, et al. Causal inference in partially linear structural equation models. *The Annals of Statistics*, 46(6A):2904–2938, 2018.

Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18(none):1 – 9, 2013. doi:10.1214/ECP.v18-2865.

Jakob Runge. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *International Conference on Artificial Intelligence and Statistics*, pages 938–947. Pmlr, 2018.

Narayana P Santhanam and Martin J Wainwright. Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Transactions on Information Theory*, 58(7):4117–4134, 2012.

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

Rajen D Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514, 2020.

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

Jun Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, 7:221–242, 1997a.

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

Xiaotong Shen, Wei Pan, and Yunzhang Zhu. Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, 107(497):223–232, 2012.

Xiaotong Shen, Wei Pan, Yunzhang Zhu, and Hui Zhou. On constrained and regularized high-dimensional regression. *Annals of the Institute of Statistical Mathematics*, 65(5):807–832, 2013.

Ritei Shibata. An optimal selection of regression variables. *Biometrika*, 68(1):45–54, 1981.

Shohei Shimizu. Lingam: Non-gaussian methods for estimating causal structures. *Behaviormetrika*, 41:65–98, 2014.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

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

Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72, 1991.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

Chandler Squires and Caroline Uhler. Causal structure learning: A combinatorial perspective. *Foundations of Computational Mathematics*, pages 1–35, 2022.

Nathan Srebro. Maximum likelihood bounded tree-width markov networks. *Artificial intelligence*, 143(1):123–138, 2003.

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

Vincent YF Tan, Animashree Anandkumar, and Alan S Willsky. Learning gaussian tree models: Analysis of error exponents and extremal structures. *IEEE Transactions on Signal Processing*, 58(5):2701–2714, 2010.

Vincent YF Tan, Animashree Anandkumar, and Alan S Willsky. Learning high-dimensional markov forest distributions: Analysis of error rates. *Journal of Machine Learning Research*, 12:1617–1653, 2011.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Joel A Tropp and Anna C Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on information theory*, 53(12):4655–4666, 2007.

Ioannis Tsamardinos, Constantin F Aliferis, and Alexander Statnikov. Time and sample efficient discovery of markov blankets and direct causal relations. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 673–678, 2003a.

Ioannis Tsamardinos, Constantin F Aliferis, Alexander R Statnikov, and Er Statnikov. Algorithms for large scale markov blanket discovery. In *FLAIRS conference*, volume 2, pages 376–380. St. Augustine, FL, 2003b.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519.

Caroline Uhler, Garvesh Raskutti, Peter Bühlmann, and Bin Yu. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, pages 436–463, 2013.

Sara van de Geer, Peter Bühlmann, and Shuheng Zhou. The adaptive and the thresholded lasso for potentially misspecified models (and a lower bound for the lasso). *Electronic Journal of Statistics*, 5:688–749, 2011.

Mishfad Shaikh Veedu, Deepjyoti Deka, and Murti Salapaka. Information theoretically optimal sample complexity of learning dynamical directed acyclic graphs. In *International Conference on Artificial Intelligence and Statistics*, pages 4636–4644. PMLR, 2024.

Nicolas Verzelen. Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic Journal of Statistics*, 6:38–90, 2012.

Marc Vuffray, Sidhant Misra, Andrey Lokhov, and Michael Chertkov. Interaction screening: Efficient and sample-optimal learning of ising models. *Advances in neural information processing systems*, 29, 2016.

Martin J Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE transactions on information theory*, 55(12):5728–5741, 2009a.

Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202, 2009b.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

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

Wei Wang, Martin J Wainwright, and Kannan Ramchandran. Information-theoretic bounds on model selection for gaussian markov random fields. In *2010 IEEE International Symposium on Information Theory*, pages 1373–1377. IEEE, 2010a.

Wei Wang, Martin J Wainwright, and Kannan Ramchandran. Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices. *IEEE Transactions on Information Theory*, 56(6):2967–2979, 2010b.

Y Samuel Wang and Mathias Drton. High-dimensional causal discovery under non-gaussianity. *Biometrika*, 107(1):41–59, 2020.

Yuhao Wang, Ming Gao, Wai Ming Tai, Bryon Aragam, and Arnab Bhattacharyya. Optimal estimation of gaussian (poly) trees. In *International Conference on Artificial Intelligence and Statistics*, pages 3619–3627. PMLR, 2024.

Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

Larry Wasserman and Kathryn Roeder. High dimensional variable selection. *Annals of statistics*, 37(5A):2178, 2009.

Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, 2016.

Bin Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam: research papers in probability and statistics*, pages 423–435. Springer, 1997.

Bin Zhang, Chris Gaiteri, Liviu-Gabriel Bodea, Zhi Wang, Joshua McElwee, Alexei A Podtelezhnikov, Chunsheng Zhang, Tao Xie, Linh Tran, Radu Dobrin, et al. Integrated systems approach identifies genetic nodes and networks in late-onset alzheimer's disease. *Cell*, 153(3):707–720, 2013.

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

Jiawei Zhang, Yuhong Yang, and Jie Ding. Information criteria for model selection. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(5):e1607, 2023.

K Zhang, J Peters, D Janzing, and B Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pages 804–813. AUAI Press, 2011.

Kun Zhang and Aapo 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.

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

Tong Zhang. Sparse recovery with orthogonal matching pursuit under rip. *IEEE transactions on information theory*, 57(9):6215–6221, 2011.

Yuchen Zhang, Martin J Wainwright, and Michael I Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Conference on Learning Theory*, pages 921–948. PMLR, 2014.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

Junxian Zhu, Canhong Wen, Jin Zhu, Heping Zhang, and Xueqin Wang. A polynomial algorithm for best-subset selection problem. *Proceedings of the National Academy of Sciences*, 117(52):33117–33123, 2020.

Or Zuk, Shiri Margel, and Eytan Domany. On the number of samples needed to learn the correct structure of a bayesian network. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 560–567, 2006.