Title
Plug-in Estimation Approaches to Causal Inference and Discovery

Author
Ruiz, Gabriel

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Plug-in Estimation Approaches to Causal Inference and Discovery

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Statistics

by

Gabriel Ruiz

2022

ABSTRACT OF THE DISSERTATION

Plug-in Estimation Approaches to
Causal Inference and Discovery

by

Gabriel Ruiz
Doctor of Philosophy in Statistics
University of California, Los Angeles, 2022
Professor Oscar Hernan Madrid Padilla, Co-Chair
Professor Qing Zhou, Co-Chair

This dissertation covers techniques for the estimation of parameters related to making causal inferences and discoveries. Both for its generality and its simplicity, the focus is in the plug-in estimation of these parameters, whereby the statistical estimator(s) of a parameter(s) is plugged in to obtain an estimator for another, possibly more difficult to estimate, parameter. In particular, the following is addressed.

In Chapter 2, we focus on causal discovery, the learning of causality in a data mining scenario. Causal discovery has been of strong scientific and theoretical interest as a starting point to identify "what causes what?" Contingent on assumptions and a proper learning algorithm, it is sometimes possible to identify and accurately estimate a causal directed acyclic graph (DAG), as opposed to a Markov equivalence class of graphs that gives ambiguity of causal directions. The focus of this chapter is in highlighting the identifiability and estimation of DAGs with general error distributions through a general sequential sorting procedure that orders variables one at a time, starting at root nodes, followed by children of the root nodes,

and so on until completion. We demonstrate a novel application of this general approach to estimate the topological ordering of a DAG. At each step of the procedure, only simple likelihood ratio scores are calculated on regression residuals to decide the next node to append to the current partial ordering. The computational complexity of our algorithm on a p-node problem is $\mathcal{O}(pd)$, where d is the maximum neighborhood size. Under mild assumptions, the population version of our procedure provably identifies a true ordering of the underlying DAG. We provide extensive numerical evidence to demonstrate that this sequential procedure scales to possibly thousands of nodes and works well for high-dimensional data. We accompany these numerical experiments with an application to a single-cell gene expression dataset.

The focus of the Chapter 3 is the Linear Non-Gaussian Acyclic Model (LiNGAM). Compared to what has been done, we present a novel estimation approach which involves specifying a parametric objective function and arguing when our sequential optimization approach will be statistically consistent, including if the dimension of underlying graph diverges, and when we can provide finite sample guarantees on its accuracy. This involves (1) defining well our target parameter: an ordering of the Directed acyclic graph (DAG)'s vertices such that parents always precede children; and (2) translating deviation bounds on the parameters for the corresponding structural equation model (SEM) into a statement about our topological order estimate's deviation from a true topological ordering. We also incorporate the use of a priori known neighborhood sets to our theoretical results.

In Chapter 4, we assume that the underlying causal structure is known, for example, due to the successful application of a causal discovery algorithm similar to those in the previous two chapters. This grants us the identifiability of parameters on the distribution of so-called potential outcomes, the key random variables we would like to make causal claims about. The premise of this chapter, in a vein similar to predictive inference with quantile regression, is that observations may lie far away from their conditional expectation. In the context of causal inference, due to the missing-ness of one outcome, it is difficult to check whether an individual's treatment effect lies close to its prediction given by the estimated Average Treatment Effect

(ATE) or Conditional Average Treatment Effect (CATE). With the aim of augmenting the inference with these estimands in practice, we further study an existing distribution-free framework for the plug-in estimation of bounds on the probability an individual benefits from treatment (PIBT), a generally inestimable quantity that would concisely summarize an intervention's efficacy if it could be known. Given the innate uncertainty in the target population-level bounds on PIBT, we seek to better understand the margin of error for the estimation of these target parameters in order to help discern whether estimated bounds on treatment efficacy are tight (or wide) due to random chance or not. In particular, we present non-asymptotic guarantees to the estimation of bounds on marginal PIBT for a randomized experiment setting. We also derive new non-asymptotic results for the case where we would like to understand heterogeneity in PIBT across strata of pre-treatment covariates, with one of our main results in this setting making strategic use of regression residuals. These results, especially those in the randomized experiment case, can be used to help with formal statistical power analyses and frequentist confidence statements for settings where we are interested in inferring PIBT through the target bounds under minimal parametric assumptions. Our results extend to both real-valued and binary-valued outcomes, and these results can also instead be applied to reason about whether an individual is likely to be harmed by an intervention.

The dissertation of Gabriel Ruiz is approved.

Arash Ali Amini

Jingyi Jessica Li

Qing Zhou, Committee Co-Chair

Oscar Hernan Madrid Padilla, Committee Co-Chair

University of California, Los Angeles

2022

To my mother and father
who—among so many other things—
taught me the value and enjoyment
in a day's hard work.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am really grateful to my advisors, Professor Qing Zhou and Professor Oscar-Hernan Madrid Padilla, for their continuous support on my research projects. They both have always been available, patient, and give illuminating advice on the projects and ideas I present to them. It is a privilege to complete this degree under their direction. I would like to thank my committee members, Jingyi-Jessica Li and Arash-Ali Amini, for their valuable advice on my dissertation, along with their excellent instruction in coursework on non-asymptotic and asymptotic statistics which were key to this dissertation. Moreover, the work in this dissertation is thanks to the financial support of the National Science Foundation (NSF) Graduate Research Fellowship Program (DGE-1650604) and the National GEM Consortium Ph.D Engineering and Science Fellowship.

I would also like to extend my gratitude to Professor Nicolas Christou, Professor Honquan Xu, Glenda Jones, Laurie Leyden, Chie Ryu, Verghese Nallengara, Enrique Reyes, my classmates Dr. Zhanhao Peng, Dr. Qiaoling Ye, Dr. Hangjian Li, Dr. Dale Kim, Dr. Jireh Huang, Stephen Smith, Yuhao Yin, Tianyi Sun and many others for their friendship and help throughout my time in the Department of Statistics. I am also grateful to my internship supervisors at Adobe Inc., Drs. Eugene Chen, Yi-Hong Kuo, and Micahel Burkhart, for allowing me to apply my data science knowledge. For their mentorship and encouragement before my graduate studies, I am appreciative of Professor Subir Ghosh, Professor Jill Smith, Dr. Dmitri Zaykin, the MARC U program and the Highlander Statistics Society at UC Riverside, and Mr. Robert Santiago who first introduced me to statistics at Rubidoux High School.

Finally, for making this degree possible with their unconditional love and support, I am grateful for my wife Lupita, my parents Cuca and Martin, and my siblings Victor and Alex.

VITA

2019, 2021	Data Science Intern, Digital Experience, Adobe Inc., San Jose, California. 12 weeks in the Summer.
2018–2021	Teaching Assistant, Department of Statistics, UCLA, Los Angeles, California. Led student discussion sections in 9 academic quarters for Statistics 10, Statistics 100A, Statistics 100C, and Statistics 200C.
2020	Ph.D. Candidate (Statistics), University of California, Los Angeles.
2017	B.S. (Statistics) and Minor (Mathematics), University of California, Riverside.

PUBLICATIONS

Ruiz, G. & Padilla, O.H.M. (2022). "Non-asymptotic confidence bands on the probability an individual benefits from treatment (PIBT)." arXiv:2205.09094. `https://arxiv.org/abs/2205.09094`.

Ruiz, G., Padilla, O.H.M. & Zhou, Q. (2022). "Sequentially learning the topological ordering of causal directed acyclic graphs with likelihood ratio scores." arXiv:2202.01748. `https://arxiv.org/abs/2202.01748`.

Padilla, O.H.M., Chen, Y, & Ruiz, G. (2022). "A causal fused lasso for interpretable heterogeneous treatment effects estimation." arXiv:2110.00901. `https://arxiv.org/abs/2110.00901`.

Burkhart, M. & Ruiz, G. (2022). "Neuroevolutionary Feature Representations for Causal Inference." Computational Science – ICCS 2022. https://doi.org/10.17863/CAM.84824

Vsevolozhskaya, O., Ruiz, G. & Zaykin, D (2017). "Bayesian prediction intervals for assessing P-value variability in prospective replication studies." Translational Psychiatry 7, 1271. https://doi.org/10.1038/s41398-017-0024-3

Vsevolozhskaya, O.A., Kuo, C.-L., Ruiz, G., Diatchenko, L., & Zaykin, D.V. (2017). "The more you test, the more you find: The smallest P-values become increasingly enriched with real findings as more tests are conducted." Genetic Epidemiology.; 41: 726– 743. https://doi.org/10.1002/gepi.22064

CHAPTER 1

Introduction

With observational data alone, causal inference using an accurate directed acyclic graph (DAG) has been shown to provide results that are up to par with the quintessential randomized controlled experiment (Pearl 2009, Malinsky et al. 2019). However, it is difficult to imagine that this approach, with its reliance on strong domain knowledge about the system of variables at hand, can be applied to cases with large numbers of variables and little background on how they are all related, such as in bioinformatics and fields of science where "big data" was previously unavailable and we are now trying to get a grasp of it. On the other hand, causal discovery—learning the DAG structure for Bayesian networks from scratch—has its own pitfalls, such as a super-exponentially increasing number of networks in the search space as the number of nodes grows and the fact that it is possible for multiple directed acyclic graphs and the Bayesian networks they encode to map to the same joint distribution–a phenomenon called Markov equivalence (Frydenberg 1990, Verma and Pearl 1990, Peters et al. 2017, Verma and Pearl 2022). Nonetheless, a large effort has been devoted to the structure learning of DAGs as a preliminary data mining step in the scientific pipeline. Section 1.2 discusses more on the motivation for DAG structures and structure learning.

Moreover, what do we do once we have high confidence about a fixed graphical model or some properties of it, such as knowing appropriate confounding variables between two variables W and Y of interest (see Assumption 4.1.5 and Assumption 4.1.7 in Chapter 4)? For such a scenario, Section 1.3 below discusses causal modeling for an estimand known as the Conditional Average Treatment Effect (CATE), which allows us to study heterogeneity in

the causal effect of W on Y due to differing values of a confounding variable (effect modifier) X. We also introduce a related parameter of interest: the probability an individual benefits from treatment (PIBT), which is discussed in depth in Chapter 4. On the one hand, a CATE model is a predictive model we can never check the goodness of fit for due to the fundamental problem of causal inference: we only observe one potential outcome for each individual in a sample, while the other potential outcomes that could result from unrealized values of the treatment variable remain hidden. On the other hand, we cannot directly estimate the probability an individual benefits from treatment, due to the same fundamental reason. This means we have to estimate bounds on PIBT in practice that could be conservative. Section 1.3 discusses CATE and the inestimable parameter of interest.

We next discuss the outline and contribution of this dissertation. This includes a short summary of each of the chapters.

1.1 Outline and Contribution of this dissertation

1. In Chapter 2, we develop a novel structure learning method that allows us to estimate the topological ordering of a large number of variables in an unknown causal graph. A key component of our method is the use of a general sequential procedure that appends a node to a partial ordering one at a time until completion, greatly reducing the search complexity across the large number of possible orderings. We theoretically establish the identifiability of such a topological ordering with our sequential procedure under well explained regularity conditions. To the best of our knowledge, ours is the most recent method that pushes the boundary of learning large causal graphs with such theoretical guarantees. Through our use of regression residuals in an easy to follow application of the plug-in estimation principle, an additional takeaway of our work is the suggestion it gives to scale up similar sequential algorithms that estimate more flexible classes of structural equation models.

2. Moreover, given the importance of applying theoretically justified methods in practice, Chapter 3 provides finite sample and asymptotic guarantees for the learning of the linear structural equation model discussed in Chapter 2. The results here can help provide a reasonable expectation for a causal discovery method, such as that introduced in Chapter 2, in terms of accuracy and its relation to dimension, connectedness of nodes, relative signal to noise, and sample size.

3. In Chapter 4, we develop novel statistical estimation theory that allows us to reason in a frequentist sense about the probability an individual benefits from treatment (PIBT)–an inestimable parameter. A key component of our approach is the use of distribution-free bounds on the parameter of interest along with non-asymptotic concentration inequalities. In doing so, we theoretically establish an understanding of how much sample size is needed to reason, within some target margin of error and frequentist confidence level, about the parameter of interest through the distribution-free bounds of choice. To the best of our knowledge, ours is the first approach that provides such a non-asymptotic statistical inference on PIBT. Through our use of regression residuals in another easy to follow application of the plug-in estimation principle, we also provide results to reason about the parameter of interest in strata of confounding covariates. Moreover, Section 4.5 of this chapter, in an application to a large randomized experiment dataset, demonstrates how our proposed methodology can help augment existing approaches for average and heterogeneous causal effect modeling.

The remainder of this chapter gives background and motivation for these three middle chapters. Finally, Chapter 5 concludes.

1.2 Motivation for Directed Graphical Models, Structure Learning

In the ideal case, our estimate of a DAG, completed partially direct acyclic graph (CPDAG) as in Figure 1.1 (Spirtes and Glymour 1991), or related structure would be accurate so that

Figure 1.1: The original DAG (left) and its corresponding CPDAG (right), obtained by keeping the orientation of edges corresponding to the v-structures $Z \to X \leftarrow U$ and $Z \to Y \leftarrow U$, and removing the orientation from all other edges (Spirtes and Glymour 1991). Note the ambiguity about the causal direction $X \to Y$ vs. $X \leftarrow Y$ in the CPDAG.

its use to augment our causal reasoning is justified. That is, the structure we estimate would guide our use of Pearl (2009)'s back-door adjustment, front-door (mediator) adjustment, instrumental variable analysis, and combinations thereof. This is the goal! As a basic example of how a graphical model can augment causal inference, consider the linear structural equation model in Figure 1.2 where a natural causal estimand of interest for the effect an intervention on variable X will have on the response variable Y is the total causal effect (see chapter 7 of Hernán and Robins (2020) or chapter 3 of Pearl et al. (2016) for more background). Due to linearity, the total causal effect can be summarized by the coefficient of X in the expectation of Y on X when the experimenter intervenes on X to be x (denoted by the do-operator). This coefficient is equivalently the expected difference of two counterfactual (potential) outcomes $Y(1)$ and $Y(0)$ under $do(X = 1)$ and $do(X = 0)$, respectively:

$$\gamma_{x\to y} \doteq \frac{\partial}{\partial x}\mathbb{E}[Y|do(X = x)] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \alpha_2\alpha_3.$$

Let $\beta_A(B \sim A+S)$ denote the coefficient of variable A in the population-level linear regression of variable B on variable A and a set of variables S. For the example in Figure 1.2, using the observational data distribution, the standard back-door adjustment argument shows that $\gamma_{x\to y}$ is equal to $\beta_X(Y \sim X + U)$ since U is the sole common cause of both X and Y. Should U be unobserved on the other hand, the standard instrumental variable analysis argument shows that $\gamma_{x\to y}$ is the same as $\beta_Z(Y \sim Z)/\beta_Z(X \sim Z)$ since Z and Y are causally related through X, while X's parents Z and U are marginally independent. And if the instrumental

variable is also unobserved or is weak (e.g. if $\alpha_1 \approx 0$ after all), the front-door (mediator) adjustment approach tells us that also $\gamma_{x \to y}$ is the same as $\beta_M(Y \sim M + X) \times \beta_X(M \sim X)$ since X closes the backdoor path between M and Y. (See the Appendix for more details.) Because the causal DAG in this example is known, we may estimate the causal estimand by the sample analogues of such regressions using non-interventional data. The key is obtaining this graph to begin with.

1.2.1 Combining Bayesian network structure learning and causal inference

Alas, it is likely unavoidable to make an estimation error in the graph estimate due to data variability, not too mention model misspecification. So a better goal is to demonstrate that causal discovery models are useful as working models in data mining settings to generate causal leads a scientist will follow up on with more experiments and/or compelling domain knowledge. To this end, procedures have been developed to take the equivalence class of DAGs outputted by a CPDAG-learning algorithm to further estimate the corresponding set of total causal effects for each DAG using Pearl et al. (2016)'s do-calculus. These works include Maathuis et al. (2009), Nandy et al. (2017), Malinsky and Spirtes (2017), and Henckel et al. (2019). Notably, Maathuis et al. (2009), the precursor to its generalizations in Nandy et al. (2017) and Malinsky and Spirtes (2017), was applied to gene expression data from an observational distribution, and many of the causal leads were validated using readily available experiment data for the plant organism of interest (Stekhoven et al. 2012). This result was despite the use of linear structural equations, albeit after a strategic log (Box-Cox) transformation of the count matrix. Maathuis et al. (2010) further discuss such applications of causal discovery in practice.

Taking for granted that the assumptions for a CPDAG-learning algorithm are met, a question arises on how to resolve ambiguities in the causal effect *set* that is given by a procedure such as Nandy et al. (2017) or Maathuis et al. (2009). Consider the example in Figure 1.1. For this example, one would forcibly estimate a causal effect of X on Y to be null

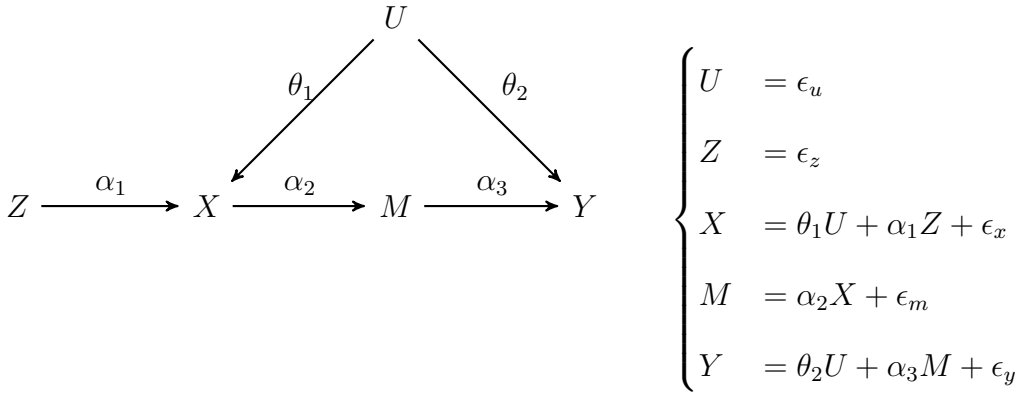

Figure 1.2: An example of a DAG and the linear structural equation model it encodes, where the exogenous ϵ terms are mutually independent and zero-centered.

using the DAG in the equivalence class which orients the edges betwen $\{X, Y\}$ as $Y \to X$. For this and the general case where there is ambiguity about the orientation of edges on a path between two nodes in a CPDAG, Nandy et al. (2017) propose to remedy the ambiguity by taking the absolute value of the total causal effect query for each DAG in the equivalence class, then summarizing the set of (absolute values of) total causal effects by taking the minimum[1]. This is done under a linear structural equation model assumption to go along with each DAG, in which case the estimand is $\frac{\partial}{\partial x}\mathbb{E}[Y|do(X = x)]$. For the example in Figure 1.1, the lower bound for the absolute value of the total causal effect of X on Y is zero, yet the possibility for a practically significant causal effect exists if the DAG on the left with edge orientation $X \to Y$ is the true causal DAG.

Therefore, the uniquely identifiable Linear Non-Gaussian Acyclic Model (LiNGAM) of Shimizu et al. (2006) can be of interest to the causal discovery practitioner who wishes for less ambiguity of causal effect estimates. Very importantly, the practitioner must be willing to grant this wish by adding stronger assumptions. If the model assumptions are not met, an estimated LiNGAM must be used cautiously as one small part of a larger quest to understand

[1]Resolving this ambiguity for the general case of one or more simultaneous intervention nodes, i.e. under $do\left(X_{\mathcal{I}} = x_{\mathcal{I}}\right)$ where $\mathcal{I} \subseteq [p]$ arbitrary, would work in a similar way: their procedure would simply keep track of total causal effect coefficient for each simultaneous intervention node across each DAG in the MEC.

causality for the system of variables at hand. When modeling assumptions are not met, a LiNGAM estimate would simply be the linear structural equation model of best fit–in our case in Chapter 2, according to the specified non-Gaussian noise-term densities and corresponding estimation procedure.

1.3 A Brief Review of Causal Modeling

Consider now that we are in a setting in which we are confident about how variables are causally related, perhaps after we have successfully applied data mining methods discussed in the previous section and Chapters 2 and 3, have had plenty of domain expert input, and also conducted various validation studies. In such a setting, it can be interesting for practitioners to understand what confounding characteristics of individuals in a population are predictive of benefiting from treatment. To this end, we now review a popular estimand known as Conditional Average Treatment Effect (CATE) along with popular estimation approaches for it. We then motivate Chapter 4 which discusses further how to reason statistically about the probability an individual benefits from treatment, an arguably more informative quantity than CATE if only it could be known.

1.3.1 The Conditional Average Treatment Effect (CATE) and its Estimation

Denote below $Y_i(w)$ as the potential outcome for individual $i = 1, \ldots, n$ in our sample when they are in binary treatment group $w = 0, 1$ (Rubin 1974, Imbens and Rubin 2015). Denote also the individual treatment effect as

$$\Delta_i := Y_i(1) - Y_i(0).$$

Here, Δ_i is a summary of the effect treatment has on individual i specifically. It could be well predicted by some covariates X_i that confound the naturally occurring treatment, W_i, and Y_i. This premise leads to the definition of CATE.

Definition 1.3.1 (Conditional Average Treatment Effect (CATE))**.**
To understand average treatment effect heterogeneity across strata of our confounder, X_i, denote:

$$\tau_x \stackrel{\Delta}{=} \mathbb{E}\left[Y_i(1) - Y_i(0) | X_i = x\right].$$

Due to linearity of expectation and Chapter 4's Assumption 4.1.7 (Strong Conditional Ignorability)[2], we have that:

$$\tau_x = \mathbb{E}\left[Y_i | W_i = 1, X_i = x\right] - \mathbb{E}\left[Y_j | W_j = 0, X_j = x\right].$$

This equivalent formulation of τ_x as the difference of two regression curves, $\mu_{1T}(x) = \mathbb{E}\left[Y_i | W_i = 1, X_i = x\right]$ and $\mu_{0T}(x) = \mathbb{E}\left[Y_j | W_j = 0, X_j = x\right]$, suggests why τ_x is identifiable: the conditional expectations can be with respect to differing individuals ($i \neq j$) such that $X_i = X_j = x$.

After learning $\mu_{1T}(x)$ and $\mu_{0T}(x)$ with the treatment group sample and control group sample, respectively, we can use $\hat{\mu}_{1T}(x) - \hat{\mu}_{0T}(x)$ as our estimate of τ_x–known as the T-Learner approach, where "T" stands for "two." One may also estimate a single regression curve, $\mu_S(x, w) = \mathbb{E}\left[Y_i | W_i = w, X_i = x\right]$ by pooling treatment and control group outcomes together. The estimate of τ_x in this approach is given by taking the difference $\hat{\mu}_S(x, 1) - \hat{\mu}_S(x, 0)$. This procedure is known as the S-Learner, where "S" stands for "single." The regression curves of both the S and T-learner approaches can be learned with any appropriate method, called the "base-learner," including linear regression, kernel regression, tree-based procedures, or artificial neural networks (Künzel et al. 2019).

For provably better statistical efficiency compared to the T and S-learner frameworks under regularity conditions, Künzel et al. (2019) introduce the X-learner, which stands for "meta-learner." This approach strategically builds off of a T-learner by using it to impute Δ_i as $W_i[Y_i - \hat{\mu}_{0T}(X_i)] + (1 - W_i)[\hat{\mu}_{1T}(X_i) - Y_i]$, then using this imputation as the regression label

[2]This assumption should be argued well in practice.

to learn τ_x. In order to obtain desirable efficiency results in their own right, Nie and Wager (2020) learns τ_x by mimicking an oracle loss function involving the main effects regression curve $m(x) = \mathbb{E}[Y_i|X_i = x]$, and the propensity score, $e(x) = \Pr(W_i = 1|X_i = x)$. Nie and Wager (2020) do so by plugging in the corresponding estimates $\hat{m}(x)$ and $\hat{e}(x)$ to the loss function which is with respect to τ_x, with special care given to sample splitting in order to mitigate biases. This approach is named the R-learner, where the "R" here stands for its namesake: the author of the key reference (Robinson 1988). The causal random forests of Wager and Athey (2018), later made more robust by the generalized random forests in Athey et al. (2019), fall under the R-learner framework to estimate CATE.

Moreover, Burkhart and Ruiz (2022) discuss a heuristic approach that can help provide greater accuracy to each of the above mentioned CATE learning frameworks. The idea is to train a feed-forward neural network (Goodfellow et al. 2016) in which X_i is the feature we input to the model in order to predict Y_i. After this is done, we use the last hidden layer as the basis for a feature representation of X_i, which we can call $\Psi(X_i)$. One would then use the learned feature representation $\Psi(X_i)$ in place of X_i in a CATE-learning algorithm. Given that $\Psi(X_i)$ already incorporates the non-linear relation between X_i and Y_i, the base-learner for this approach with $\Psi(X_i)$ can be a linear model if one wishes.

1.3.2 Does CATE really imply what it suggests about benefiting from treatment?

Asking whether CATE really implies what it suggests about benefiting from treatment, in part, boils down to how far points on the conditional distribution $\Delta_i|X_i = x$ tend to be from τ_x. That is, it is partly a matter of whether the distribution $\Delta_i|X_i = x$ is thin tailed. Does this distribution have outliers which pull the mean toward them? Are there multiple modes?

In regression lingo, this question might be answered in part by the proportion of Δ_i's variance explained by X_i through τ_{X_i}:

$$R^2_{ITE} = \frac{\mathbf{V}[\Delta_i] - \mathbb{E}[\mathbf{V}[\Delta_i|X_i]]}{\mathbf{V}[\Delta_i]} = 1 - \frac{\mathbb{E}[\mathbf{V}[\Delta_i|X_i]]}{\mathbf{V}[\Delta_i]},$$

which is between 0 and 1 due to the law of total variance. If R^2_{ITE} is large, say 90% or larger, then one can feel more confident in making a statement about treatment efficacy for individuals with large values of CATE as it would appear that most values of ITE are near the curve τ_x.

If we know the ITE for each individual in our sample used to learn CATE, R^2_{ITE} can be estimated by:

$$\widehat{R^2}_{ITE} = 1 - \frac{\sum_{i=1}^{n} \left(\Delta_i - \hat{\tau}_{X_i}\right)^2}{\sum_{i=1}^{n} \left(\Delta_i - \frac{1}{n}\sum_{j=1}^{n} \Delta_j\right)^2},$$

which makes use of the residual sum of squares and total sum of squares. Alas, computing the residual $\Delta_i - \hat{\tau}_{X_i}$ as we typically would in supervised regression is impossible.

Given that $\widehat{R^2}_{ITE}$ cannot be calculated to assess the goodness of fit of $\hat{\tau}_{X_i}$, we propose to study estimation strategies for bounds on the proportion of individuals in stratum x that benefit from treatment. We define this proportion, without loss of generality[3], as:

$$\theta_x := \Pr\left(\Delta_i > \delta | X_i = x\right).$$

θ_x, if it could be known, seems more practically informative than $\widehat{R^2}_{ITE}$. Here, δ is a fixed threshold of interest, such as $\delta = 0$. The quantity θ_x is importantly different in general from

$$\eta_x := \Pr\left(Y_i(1) - Y_j(0) > \delta | X_i = X_j = x; i \neq j\right),$$

as discussed below and in Chapter 4 (Hand 1992, Fay et al. 2018, Greenland et al. 2020).

Let us now study the relation between θ_x and τ_x to gain some intuition for why the study of individuals benefiting from treatment in Chapter 4 is interesting. Consider the toy generative model on Δ_i given by:

$$\begin{aligned} \Delta_i &= \tau_{X_i} + \nu(X_i)\left[\epsilon_{i1} - \epsilon_{i0}\right] \\ \begin{bmatrix}\epsilon_{i0} \\ \epsilon_{i1}\end{bmatrix} &\sim \mathcal{N}_{2\times 1}\left(\begin{bmatrix}0 \\ 0\end{bmatrix}, \begin{bmatrix}1 & \rho \\ \rho & 1\end{bmatrix}\right). \end{aligned} \tag{1.1}$$

[3]See the Introduction and Section 4.4 of Chapter 4 to understand why there is not a loss of generality.

Notice that conditional on $X_i = x$, we have that the ITE residual satisfies:

$$\Delta_i - \tau_{X_i} = \nu(x)[\epsilon_{i1} - \epsilon_{i0}] \quad \sim \mathcal{N}\left(0, 2\nu(x)^2(1-\rho)\right).$$

Here, it can be appreciated that $\Delta_i - \tau_{X_i}$ satisfies homoscedasticity (constant variation across possible values of X_i) so long as $\nu(\cdot)$ is a constant function. Otherwise, the variance of $\Delta_i - \tau_{X_i}$ depends on $X_i = x$.

Further,

$$\theta_x = Pr(\Delta_i - \tau_{X_i} > \delta - \tau_x | X_i = x) = 1 - \Phi\left(\frac{\delta - \tau_x}{|\nu(x)|\sqrt{2(1-\rho)}}\right) \tag{1.2}$$

and similarly,

$$\eta_x = 1 - \Phi\left(\frac{\delta - \tau_x}{|\nu(x)|\sqrt{2}}\right). \tag{1.3}$$

Here Φ is the CDF for a standard normal distribution.

From Equations (1.2) and (1.3), one can begin to appreciate the complexity of reasoning about the relation between τ_x and θ_x that comes when $\Delta_i - \tau_{X_i}$ violates homoscedasticity. Figure 1.3 gives this insight for specific choices of $\nu(x)$, ρ, σ, and threshold δ. For the relation between θ_x and τ_x, we see in the homoscedastic case where $\nu(x) = 1$ that θ_x increases monotonically with τ_x, with the rate of increase being dictated by ρ. For the heteroskedastic cases (non-constant variation of residuals across possible values of X_i), especially when $\nu(x) = 1 + \sin(\tau_x)$, we see that we may counter-intuitively have that $\theta_x < \theta_{x'}$ even if $\tau_x > \tau_{x'}$.

For the general case that homoscedastic violations are allowed, stronger tools than those presented in Chapter 4's Theorem 4.3.4 for the homoscedastic case are needed in order to estimate bounds on θ_x and to estimate η_x. Because there appears to be a gap in the literature for the homoscedastic case, we deem its discussion in this chapter relevant nonetheless. We consider violations of homoscedasticity outside of the scope of this paper.

With respect to the relation between τ_x and θ_x, Figure 4.5.1 in Chapter 4 gives some interesting insight. Upon a partitioning of individuals in the sample using their similar CATE

prediction, we can estimate PIBT in each hand-crafted strata of the partition to better understand the implication of a CATE estimate.

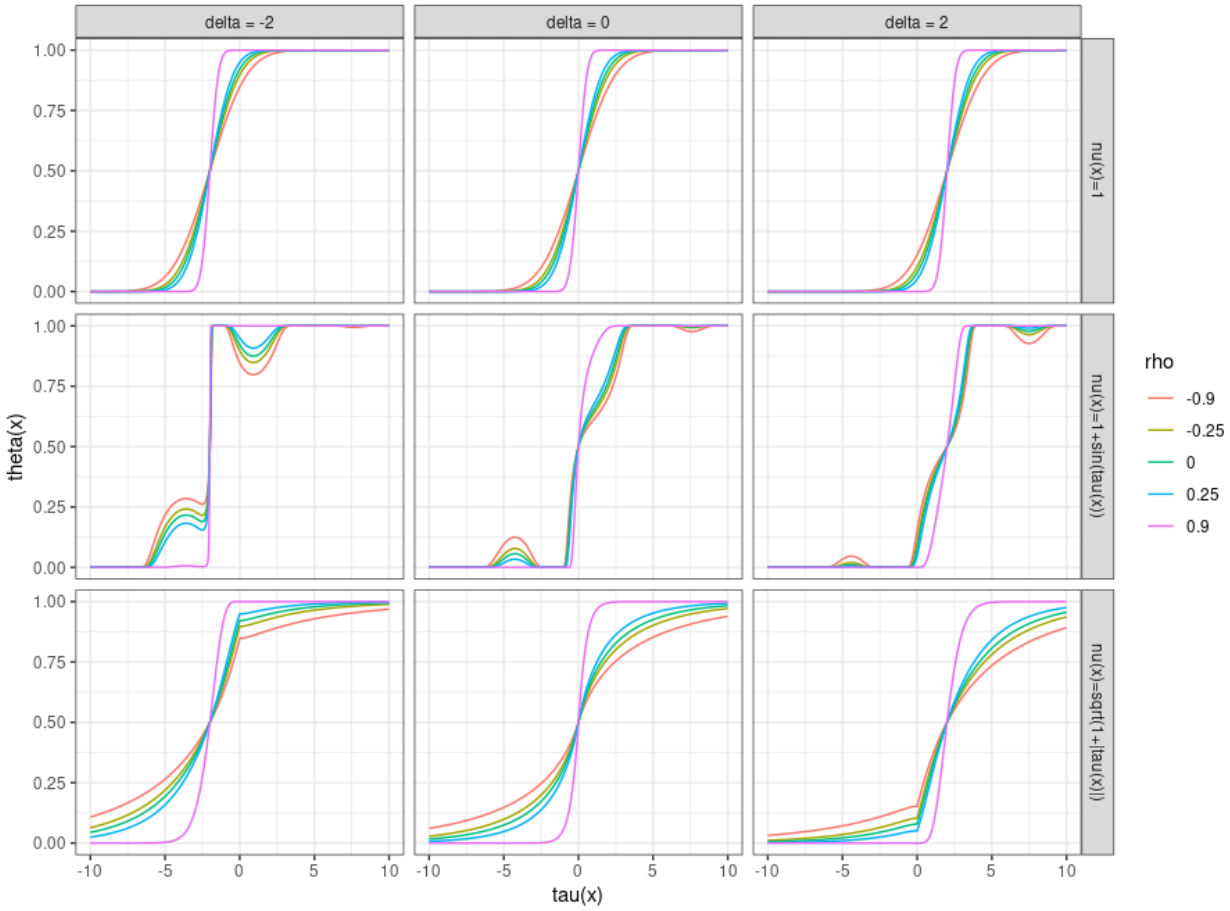

Figure 1.3: For the toy example given in Equation (1.1), the corresponding relation between τ_x and θ_x. The rows on the grid of plots correspond to $\nu(x) = 1$ (homoscedastic case), $\nu(x) = 1 + \sin(\tau_x)$, and $\nu(x) = \sqrt{1 + |\tau_x|}$. The columns on the grid of plots corresponds to examples choices of our threshold δ. The color corresponds to examples for ρ, the Pearson correlation between ϵ_0 and ϵ_1. The x-axis of each plot on the grid corresponds to values of τ_x. The y-axis on the top grid corresponds to θ_x (note that $\theta_x = \eta_x$ when $\rho = 0$ in this example).

APPENDIX

1.A Deriving Causal Estimands from the Example in Figure 1.2

Recall the estimand of interst is $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \alpha_2\alpha_3$. This is the case because an experimenter setting the value of X to be 1 or 0 does not change the value of U, but it does change the value of M to $M(1) = \alpha_2 + \epsilon_m$ and $M(0) = \epsilon_m$, respectively. In turn, this means that $Y(1) = \alpha_3(\alpha_2 + \epsilon_m) + \theta_2 U + \epsilon_y$, while $Y(0) = \alpha_3\epsilon_m + \theta_2 U + \epsilon_y$. Taking the difference in expectations gives $\alpha_2\alpha_3$ as our estimand.

Using the observational data distribution, the standard back-door adjustment argument shows that the causal effect is $\beta_X(Y \sim X + U)$, since

$$\mathbb{E}\left[Y|X,U\right] = \mathbb{E}\left[\theta_2 U + \alpha_3 M|X,U\right] = \theta_2 U + \alpha_3 \mathbb{E}\left[M|X\right] = \theta_2 U + \alpha_3 \mathbb{E}\left[\alpha_2 X|X\right] = \theta_2 U + \alpha_2\alpha_3 X.$$

The second equality holds because M and U are d-separated by X.

When we use Z as an instrumental variable, the standard argument shows that the causal estimand of interest is the same as $\beta_Z(Y \sim Z)/\beta_Z(X \sim Z)$, since $\mathbb{E}[X|Z] = \alpha_1 Z$ while

$$\mathbb{E}\left[Y|Z\right] = \mathbb{E}\left[\theta_2 U + \alpha_3 M|Z\right] = \theta_2 \mathbb{E}\left[U\right] + \alpha_3 \mathbb{E}\left[M|Z\right] = \theta_2 \mathbb{E}\left[U\right] + \alpha_3 \mathbb{E}\left[\alpha_2 X|Z\right] = \theta_2 \mathbb{E}\left[U\right] + \alpha_3\alpha_2\alpha_1 Z.$$

In the second equality we used that U and Z are d-separated by the empty set.

For the front-door (mediator) adjustment approach, the causal estimand of interest is the same as $\beta_M(Y \sim M + X) \times \beta_X(M \sim X)$, since $\mathbb{E}[M|X] = \alpha_2 X$ while

$$\mathbb{E}\left[Y|M,X\right] = \mathbb{E}\left[\theta_2 U + \alpha_3 M|M,X\right] = \alpha_3 M + \theta_2 \mathbb{E}\left[U|M,X\right] = \alpha_3 M + \theta_2 \mathbb{E}\left[U|X\right].$$

The last equality holds because U is d-separated from M by X.

CHAPTER 2

Sequentially Learning the Topological Ordering of Causal Directed Acyclic Graphs with Likelihood Ratio Scores

2.1 Introduction

The present chapter provides discussion of a general learning algorithm, which we show can be quite scaleable, along with novel identification theory for a specific application. It is based on our work in the pre-print Ruiz et al. (2022). The main task of the approach we advocate for is to sequentially estimate a topological ordering of the DAG, a permutation of node labels such that every parent must precede its children. To help with scalability in practice, we also make use of a priori known neighborhood sets, such as a Markov blanket of a node. In order to demonstrate the theoretical promise of this procedure, we discuss existing identification results that make use of it. We also provide new theory for a linear structural equation model (SEM) first studied in Shimizu et al. (2006). The novelty of our application of the sequential sorting procedure to this SEM compared to the state-of-the-art for it is the scalability of our procedure to a large number of nodes in the underlying graphical model.

Representative methods for causal discovery under the assumption of no unobserved confounding (causal sufficiency) include the Peter-Clark (PC) algorithm (Spirtes and Glymour 1991) and Greedy Equivalence Search (GES) (Chickering 2002). The PC algorithm is a constraint-based method due to its use of conditional independence queries, while GES is

considered a score-based method for the objective function it seeks to optimize across the space of graphical models. Without additional structural assumptions, the best these methods can generally do in the limit of sample size ($n \to \infty$) is to obtain a Markov equivalence class (MEC) of DAGs, visualized typically by a single Completed Paritially Directed Acyclic Graph (CPDAG) as in Figure 1.1. Each DAG in the MEC, obtainable by orienting undirected edges in the CPDAG without introducing a cyclic path nor a "v-structure," encodes the same set of *d-separation* relations that imply marginal and conditional independence relations between triplets of variable subsets in their underlying joint distribution (Spirtes and Glymour 1991).

When additional assumptions are justified, such as strict non-linearity of structural equations, or non-Gaussianity of noise terms in a linear structural equation model, a unique DAG can be identified (Bühlmann et al. 2014, Shimizu et al. 2006). When we are not willing to make the assumption of causal sufficiency, the Fast Causal Inference (FCI) algorithm provides an alternative at the cost of a potentially less precise, though possibly more accurate, graphical model compared to a DAG or CPDAG (Spirtes et al. 2000). Beyond what we highlight here for the case of iid samples from a distribution that our DAG of interest satisfies the Markov property with respect to, Glymour et al. (2019) and Peters et al. (2017) provide reviews on the trade-offs of different algorithms and what can and cannot be done when there is additional structure, such as the case that the system of variables varies in time. In the context of Earth system sciences, Runge et al. (2019) review causal discovery methods. Structure learning has also been explored for its possibility to explain the black-box nature of state-of-the-art deep learning architectures (Sani et al. 2020). Moreover, Zheng et al. (2018) and its extension to Zheng et al. (2020) provide an approach to optimize a non-convex score function in DAG space by using a smooth characterization of an adjacency matrix's acyclicity constraint.

2.1.1 Review of relevant work

Specific to our task of learning a topological ordering for an underlying acyclic graphical model, we now review some relevant work. Let us first formally define a so-called topological ordering of a DAG, the target parameter we seek to estimate. Let $[m] = \{1, \ldots, m\}$ for integer $m \geq 1$.

Definition 2.1.1. A topological ordering for a DAG $\mathcal{G}$ is given by a permutation $\pi : [p] \to [p]$ such that every parent node precedes its child in the ordering:

$$j \in PA_k \implies \pi^{-1}(j) < \pi^{-1}(k).$$

Importantly, we note that the discrete search space across $p!$ permutation functions in search of one that satisfies Definition 2.1.1 can be quite cumbersome (Raskutti and Uhler 2018, Solus et al. 2021). Several heuristic score-based methods have been developed to cope with the search space, however, it remains the case that score-based approaches for ordering search are NP hard in general (Chickering 1996, Ye et al. 2021a). Along these lines, recent work by (Ye et al. 2021a) provides one approach under the case of a linear Bayesian network with Gaussian noise. The similar non-parametric approach of Solus et al. (2021) and Wang et al. (2017) requires a consistent conditional independence testing procedure to decide the presence or absence of an edge in the DAG corresponding to a given $\tilde{\pi}$ in the search space. The empirical results of these approaches are all promising. However, these methods do not contain an application to more than 100 nodes. Moreover, these works do not provide guarantees on whether the search can terminate at a point well before querying all permutations or DAGs in order to achieve optimal (statistically consistent) results.

Complementary to these advances, we here study a simple approach for which the search has a pre-determined number of steps: $\mathcal{O}(pd)$ in the number of least squares residual updates, where $d \leq p$ is the maximum neighborhood size of a node. The objective of our work is to estimate one such permutation π from observed data $\mathbf{X} \in \mathbb{R}^{n \times p}$ of sample size n. We denote this estimate as $\hat{\pi}$. To do so, we will apply Algorithm 1 for $t = 1, 2, \ldots, p$ until all nodes are

sorted. At step t, given an input partial ordering $\mathcal{A}_t = (\hat{\pi}(1), \ldots, \hat{\pi}(t-1))$, the algorithm identifies a node $\hat{\pi}(t) \notin \mathcal{A}_t$ to append to the estimated ordering by maximizing the score $\mathcal{S}(k, \mathcal{A}_t; \mathbf{X})$.

Algorithm 1: Continue a Topological Ordering

Data: The partial ordering $\mathcal{A}_t$ and data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$

Result: The continued partial ordering $\mathcal{A}_{t+1}$

for $k \notin \mathcal{A}_t$ **do**
  $s_k \leftarrow \mathcal{S}(k, \mathcal{A}_t; \mathbf{X})$

end

$\hat{\pi}(t) \leftarrow \arg\max_{k \notin \mathcal{A}_t} s_k$

$\mathcal{A}_{t+1} \leftarrow \mathcal{A}_t \cup \{\hat{\pi}(t)\}$

There exist structure learning methods that use the general approach in Algorithm 1 to sequentially construct a topological ordering. These approaches motivate our present work and include the following. Peters et al. (2014) apply Algorithm 1 under an assumption of strictly nonlinear structural equations with additive noise. Meanwhile Ghoshal and Honorio (2017), Park (2020), Park and Kim (2020), and Chen et al. (2019) apply this sequential sorting procedure under a bounded conditional variance assumption: $a \leq \mathbf{V}[X_j | X_{PA_j}] \leq b$ for each $j \in [p]$ and some unknown positive constants $a \leq b$ restricted by the signal a parent sends its child node. Park (2020) can be considered the most general of the three similar approaches as it contains an extended discussion on the case of a node's possibly non-linear relation with its parents. Gao et al. (2020) further explore the scaleability for the sequential application of Algorithm 1 to estimate non-linear structural equation models under this bounded conditional variance assumption. With respect to linear SEMs, applications of Algorithm 1 include Shimizu et al. (2011), Hyvärinen and Smith (2013), and Wang and Drton (2019), while Zeng et al. (2020) construct the topological ordering in reverse starting with child-less nodes. We believe there exists potential to scale up the estimation of each of these

models. We focus on the linear SEM here.

2.1.2 Chapter Contribution and Outline

Our application of Algorithm 1 to a linear SEM with non-Gaussian noise under causal sufficiency, a model known as the linear non-Gaussian acyclic model (LiNGAM). In terms of theoretical guarantees for the estimation of LiNGAM, Shimizu et al. (2006) and Shimizu et al. (2011) provide identifiabiliy results for the respective LiNGAM learning procedures–that is, with knowledge of the true distribution defined by the LiNGAM and an oracle for conditional independence queries in the case of latter. Meanwhile Wang and Drton (2019) provide formal statistical consistency results for their LiNGAM-learning procedure.

Although the above methods have very nice theoretical guarantees, their practical application is limited as they do not presently scale well to large graphs, say with thousands of nodes, as confirmed in the numerical results in this chapter. Therefore, we develop a fast sequential learning method that can estimate large graphs in practice. At each step of this method, a node is selected to append to a partial ordering, so that after p steps, where p is the number of nodes in the underlying graph, a full ordering of all the nodes will be produced. Compared to the existing works on LiNGAM, the main contributions of our work are:

1. Based on a specified error distribution, we define a novel likelihood ratio score which is used at each step in our sequential algorithm. The evaluation of the likelihood ratio only involves linear regression and residual calculation. There are no tuning parameters.

2. We prove that at the population-level, this sequential algorithm will identify a true ordering of the underlying DAG under proper assumptions on the LiNGAM.

3. Our sequential method is computationally tractable with computational complexity $O(p^2)$ for the number of updates used in the entire algorithm. If prior knowledge on the Markov blankets of the nodes is provided, the computational complexity can be further reduced to $O(pd)$, where d is the maximum size of the Markov blankets. This is

in sharp contrast to traditional score-based approaches for ordering search, which are NP hard in general (Chickering 1996, Ye et al. 2021a).

The rest of the chapter is organized as follows. In the rest of this section, we formally introduce the linear SEM of interest. Next, in Section 2.2 we will introduce our approach: §2.2.1 discusses the conditions for this approach to work; §2.2.2 provides a formal identifiability result; and §3.2 provides the finite sample version of the algorithm. Section 2.3 presents simulation results for our procedure for small and large-sized Bayesian networks, along with an application to single-cell gene expression data. Finally, we conclude with a summary of our findings and discussion of future work.

2.1.3 Where the LiNGAM Falls Within Existing Work

2.1.3.1 Review of LiNGAM

We follow closely here the definition of a LiNGAM given by Shimizu et al. (2006).

Definition 2.1.2. (Linear Non-Gaussian Acyclic Model)
For $p \geq 2$, let $\mathcal{G}$ be a DAG on p nodes and $\mathbf{B} \in \mathbb{R}^{p \times p}$ be the weighted adjacency matrix of $\mathcal{G}$ such that $\mathbf{B}_{jk} \neq 0$ means $j \in PA_k$, the parent set of node k. Let $\epsilon = (\epsilon_1, \ldots, \epsilon_p)$ such that $\epsilon_k \sim g(\cdot; \theta_k)$ independently with $g(\cdot; \theta)$ a density of a non-Gaussian distribution parameterized by $\theta \in \mathbb{R}^q$. We say $X \in \mathbb{R}^p$ follows a LiNGAM with DAG $\mathcal{G}$ if

$$X_k = \sum_{j \in PA_k} \mathbf{B}_{jk} X_j + \epsilon_k, \qquad k = 1, \ldots, p. \tag{2.1}$$

The scalar form of the linear SEM in Equation (2.1) can be rewritten in vector form as $X = \mathbf{B}^T X + \epsilon$. Put $\mathbf{M} = (\mathbb{I}_p - \mathbf{B})^{-T}$, a matrix with ones on its diagonal. Let AN_k denote the ancestor set of node k: $a \in AN_k$ means there exists a direct path starting at node a and ending at node k, $a \to \cdots \to k$. Then we arrive at $X = \mathbf{M}\epsilon$ in vector form and $X_k = \sum_{j \in AN_k \bigcup \{k\}} \mathbf{M}_{kj} \epsilon_j$ in scalar form for all $k \in [p]$. Noting that $\mathbf{M}$ serves as a mixing

matrix for the independent components in ϵ, we may think of the estimation of this linear SEM as an instance of Independent Component Analysis (ICA) (Hyvärinen and Oja 2000). Shimizu et al. (2006) discuss the connection between LiNGAM and ICA.

2.2 Methodology and algorithm

In this section, we introduce both the population-level and finite-sample versions of our sorting procedure. We also show that our choice of summary score $\mathcal{S}(k, \mathcal{A}_t; \mathbf{X})$ in Algorithm 1 will lead to the identification of a topological ordering of the true DAG $\mathcal{G}$ used to define the linear SEM of Definition 2.1.2. We start with a few main assumptions on the linear SEM we will work with.

2.2.1 Assumptions

Our main assumptions are on the distributions of the independent errors ϵ. We consider restricting our class of densities $\{g(\cdot; \theta_k)\}_{1 \leq k \leq p}$ for the noise terms in Definition 2.1.2 to a scale-location family in which the $\theta_k > 0$ are the scale parameters, such as the Laplace family of distributions, the Logistic family of distributions, or a Scaled-t distribution family (same degrees of freedom). This is summarized in Assumption 2.2.1.

Assumption 2.2.1.
Let $U \sim g(\cdot; \theta_0)$ with $\theta_0 > 0$ and $\mathbb{E}[U] = 0$. For each $k = 1, 2, \ldots, p$, the density of the error ϵ_k satisfies

$$g(e; \theta_k) = \frac{\theta_0}{\theta_k} g(\theta_0 e / \theta_k; \theta_0).$$

That is, $\epsilon_k \overset{d}{=} (\theta_k / \theta_0) U$, an equality in distribution.

Our next assumption for the linear SEM of interest is on linear combinations of the noise terms. This condition is related to Lemma 2.B.1 in the appendix, a key result about how to characterize the regression residuals of Equation (2.2) as linear combinations of "independent

components."

Assumption 2.2.2. For any $j = 1, 2, \ldots, p$ and any $a \in \mathbb{R}^p$ with at least two non-zero entries, the linear combination $a^T\epsilon$ does not follow the same distribution as ϵ_j.

Notable disagreements with Assumption 2.2.2 are when the ϵ_j are all Gaussian distributed (not the case for LiNGAM), or when the ϵ_j are all Poisson-distributed. Notable agreements with Assumption 2.2.2 (and Assumption 2.2.1) are the cases where the ϵ_j are all Laplace-distributed, all Logistic-distributed, or all Scaled-t distributed (same degrees of freedom). This can be concluded with the characteristic function for a linear combination of two or more ϵ_j's.

To allow for a quicker sorting procedure in practice, we may make use of an *a priori* known support set for the neighborhood of each node in the DAG. We consider these neighborhood sets to arise based on domain knowledge, previous studies, or a pre-processing step such as with neighborhood lasso regression of Meinshausen and Bühlmann (2006). We highlight this usage in Assumption 2.2.3:

Assumption 2.2.3. For node k, denote its neighborhood estimate as $\widehat{N}_k$. Assume for each $k = 1, 2, \ldots, p$ that:

$$\widehat{N}_k \supseteq MB_k := PA_k \cup CH_k \cup \bigcup_{j \in CH_k} PA_j \backslash \{k\},$$

where MB_k is known as the Markov Blanket of node k: the set of its parents PA_k, its children CH_k, and its co-parents $\bigcup_{j \in CH_k} PA_j \backslash \{k\}$.

Let $\widehat{N}_{kt} := \widehat{N}_k \cap \mathcal{A}_t$, which is the subset of the neighborhood set that has been ordered at step t of our procedure (Algorithm 1). For the cases where $|\widehat{N}_{kt}| \geq 1$, we will make use of least squares residuals for calculating the score $\mathcal{S}(k, \mathcal{A}_t)$. The corresponding sample version

is discussed in §3.2. At the population-level, the residual is

$$R_{kt} := \begin{cases} X_k & \text{if } |\widehat{N}_{kt}| = 0 \\ X_k - \beta_{kt}^T X_{\widehat{N}_{kt}} & \text{otherwise} \end{cases}, \tag{2.2}$$

where β_{kt} is the least-squares regression coefficient vector,

$$\beta_{kt} = \left(\mathbb{E}\left[X_{\widehat{N}_{kt}} X_{\widehat{N}_{kt}}^T\right]\right)^{-1} \mathbb{E}\left[X_{\widehat{N}_{kt}} X_k\right].$$

Remark 2.2.4. When we consider the population-level version of our algorithm in this section (i.e. we have infinite n), we can take $\widehat{N}_k = [p]\backslash\{k\}$ for each k so that Assumption 2.2.3 holds trivially. For the finite sample version of our procedure discussed in Section 3.2, we will make use of Ordinary Least Squares (OLS) linear regressions which require the design matrix to be of full column rank. So if $p \ll n$, we may also take $\widehat{N}_k = [p]\backslash\{k\}$ for each k. In the case that $p \gg n$ or $p \approx n$, the neighborhood sets can reduce the number of covariates in OLS regression if $|\widehat{N}_k| \ll n$ for all k.

2.2.2 Our Choice of a Likelihood Ratio Score

In Algorithm 1, we will select the next node to continue our constructed topological ordering as:

$$\hat{\pi}(t) = \arg\max_{k \notin \mathcal{A}_t} \mathbb{E}_{f_{kt}(r_{kt})}\left[\log \frac{g(R_{kt}; \eta_{kt})}{\phi(R_{kt}; \sigma_{kt})}\right]. \tag{2.3}$$

Here, $\mathbb{E}_{f_{kt}(r_{kt})}[\cdot]$ denotes expectation with respect to R_{kt}'s true density, $f_{kt}(r_{kt})$. Also,

$$\eta_{kt} := \arg\max_{\eta} \mathbb{E}_{f_{kt}(r_{kt})}\left[\log g(R_{kt}; \eta)\right],$$

while $\phi(r_{kt}; \sigma_{kt})$ is the density for $\mathcal{N}\left(0, \sigma_{kt}^2 = \mathbf{V}[R_{kt}]\right)$, i.e. the normal density that matches the mean and variance of R_{kt}. Note that $f_{kt}(r_{kt})$ is in general different from $g(r_{kt}; \eta_{kt})$.

The log-likelihood ratio in (2.3) can be thought of as a score that tells us "how non-Gaussian" the residual R_{kt} is. If the residual is explained by a Gaussian distribution well relative to the non-Gaussian distribution in the assumed family, then we expect the log-likelihood ratio to be smaller. Otherwise, if the Gaussian density is not a good fit relative to

$g(r_{kt}; \eta_{kt})$, then we have stronger evidence to believe that node k is a valid node to continue the ordering. In Theorem 2.2.5, we claim that using (2.3) leads to the identification of a valid topological ordering, our main result.

Theorem 2.2.5. *Let $X \in \mathbb{R}^p$ follow a LiNGAM with DAG $\mathcal{G}$. If Assumptions 2.2.1, 2.2.2 and 2.2.3 hold, then applying Algorithm 1 at all steps $t = 1, 2, \ldots, p$ with the score*

$$\mathcal{S}(k, \mathcal{A}_t) = \mathbb{E}_{f_{kt}(r_{kt})} \left[\log \frac{g(R_{kt}; \eta_{kt})}{\phi(r_{kt}; \sigma_{kt})} \right]$$

will identify a permutation $\hat{\pi} = (\hat{\pi}(1), \ldots, \hat{\pi}(p))$ that is a topological ordering of $\mathcal{G}$.

Theorem 2.2.5 suggests that the maximization at each iteration in which we apply Algorithm 1 can be done easily. This differs from maximizing a score over a whole ordering which may also lead to identification of the true MEC, but is in general NP hard (not tractable). Relatedly, Appendix 2.A gives additional motivation for the choice of $\mathcal{S}(k, \mathcal{A}_t)$ in Theorem 2.2.5 as one that allows us greedily optimize the mean log-likelihood when the full ordering is only partially discovered. The proof of Theorem 2.2.5 in Appendix 2.B.1 is an inductive application of key lemmas found in Appendix 2.B.2.

2.2.3 Finite Sample Sorting Procedure

Assume that we have a data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ such that $\mathbf{X}_{i\cdot}$, the i-th row, is iid across $i = 1, 2, \ldots, n$ from a distribution defined by a LiNGAM satisfying Assumptions 2.2.1 and 2.2.2. Also let Assumption 2.2.3 hold, where the sets $\widehat{N}_k$ are given by domain knowledge, or they are estimated with data independent of $\mathbf{X}$ by an asymptotically consistent procedure.

Denote by $\mathbf{X}_{\cdot S}$ the columns of $\mathbf{X}$ indexed by the set S. When S is a singleton, such as $S = \{k\}$, we will simply write $\mathbf{X}_{\cdot k}$ for the k-th column. Analogous to Section 2.2.2, consider:

$$\hat{\beta}_{kt} = \left(\mathbf{X}_{\cdot \widehat{N}_{kt}}^T \mathbf{X}_{\cdot \widehat{N}_{kt}} \right)^{-1} \mathbf{X}_{\cdot \widehat{N}_{kt}}^T \mathbf{X}_{\cdot k} \in \mathbb{R}^{|\widehat{N}_{kt}| \times 1},$$

which exists so long as $1 \leq |\widehat{N}_{kt}| \leq n$ and $\mathbf{X}_{\cdot \widehat{N}_{kt}}$ is of full column rank almost surely. Further,

we define $\hat{R}_{kt} \in \mathbb{R}^{n\times 1}$ as

$$\hat{R}_{kt} = \begin{cases} \mathbf{X}_{\cdot k} & \text{if } |\widehat{N}_{kt}| = 0 \\ \mathbf{X}_{\cdot k} - \mathbf{X}_{\cdot \widehat{N}_{kt}} \hat{\beta}_{kt} & \text{if } |\widehat{N}_{kt}| \geq 1 \end{cases},$$

the vector of residuals which we will use to estimate the pertinent scale parameter of (2.3), denoted as $\hat{\eta}_{kt}$ and $\hat{\sigma}_{kt}$, respectively. Explicitly, we select the next node to continue an ordering using the empirical analogue of the mean log-likelihood ratio in Equation (2.3):

$$\hat{\pi}(t) = \arg\max_{k \notin \mathcal{A}_t} \frac{1}{n} \sum_{i=1}^{n} \log \frac{g(\hat{R}_{i,kt}; \hat{\eta}_{kt})}{\phi(\hat{R}_{i,kt}; \hat{\sigma}_{kt})}, \tag{2.4}$$

where $\hat{R}_{i,kt}$ is the i-th entry of the vector $\hat{R}_{kt}$, while $\hat{\sigma}^2_{kt} := \frac{1}{n}\|\hat{R}_{kt}\|_2^2$ and $\hat{\eta}_{kt} := \arg\max_\eta \sum_{i=1}^n \log g(\hat{R}_{i,kt}; \eta)$. For example, if η_{kt} is the scale parameter for a Laplace distribution, it can be seen that $\hat{\eta}_{kt} = \frac{1}{n}\|\hat{R}_{kt}\|_1$. In this case, (3.1) is equivalent to

$$\hat{\pi}(t) = \arg\max_{k \notin \mathcal{A}_t} \log \frac{\hat{\sigma}_{kt}}{\hat{\eta}_{kt}} = \arg\max_{k \notin \mathcal{A}_t} \frac{\|\hat{R}_{kt}\|_2}{\|\hat{R}_{kt}\|_1}. \tag{2.5}$$

The Laplace update (2.5) exemplifies how simple the maximization of our likelihood ratio score is. After the regression of each unsorted node X_k, $k \notin \mathcal{A}_t$, onto $\widehat{N}_{kt}$, we only need to compare the ratio between the two norms of the residual vector $\hat{R}_{kt}$ across unsorted nodes to find $\hat{\pi}(t)$. Algorithm 2 in the Supplementary Material shows the pseudo-code for the sorting procedure we use in practice, with a strategic update of regression residuals using partial regression that greatly reduces the computation cost. We have also provided the details on the estimation of the scale parameters for Logistic and Scaled-t distributions in Appendix 2.D.1.

2.3 Empirical Results

2.3.1 Simulations on Small Networks

We now present simulation results for networks that are on the smaller end ($35 \leq p \leq 223$), downloaded from the `bnlearn.com` Bayesian network repository. We compared our sorting

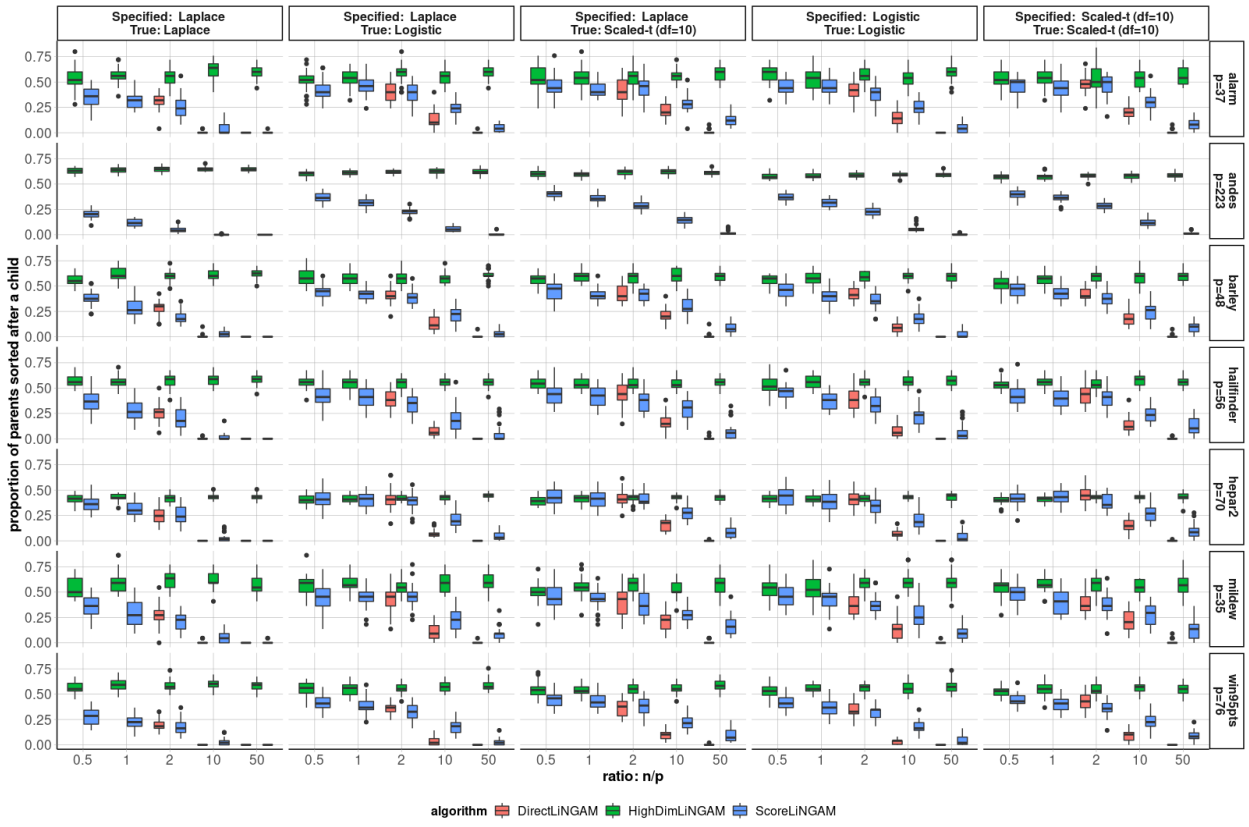

Figure 2.2.1: The simulation results comparing LiNGAM estimation procedures.

procedure to other LiNGAM learning procedures. Due to their readily available code, the algorithms of interest are "DirectLiNGAM" (Shimizu et al. 2011), "HighDimLingam" (Wang and Drton 2019), and "ScoreLiNGAM" (our procedure). For each simulation setting, we conduct 30 replicates.

For each choice of $\mathcal{G}$ underlying a LiNGAM, our synthetic data generation schema was as follows. We generated $\mathbf{B}_{jk} \overset{i.i.d.}{\sim} \text{Uniform}[-0.9, -0.4] \cup [0.4, 0.9]$ for each (j, k) such that $j \in PA_k$, and otherwise set $\mathbf{B}_{jk} = 0$. We generated $\theta_k \overset{i.i.d.}{\sim} \text{Uniform}[0.4, 0.7]$ across $1 \leq k \leq p$, where θ_k is the scale parameter for the error distributions as in Assumption 2.2.1. Finally, we varied sample size as $n = 0.5p, p, 2p, 10p, 50p$. Note that $n = 0.5p$ and $n = p$ represent the high-dimensional setting ($p \geq n$). Next, a data set $\mathbf{X} \in \mathbb{R}^{n \times p}$ of iid samples is drawn

from the distribution given by the LiNGAM parameterized by $(\mathbf{B}, \theta_1, \ldots, \theta_p)$ and having errors $\epsilon_k \sim g(\cdot; \theta_k)$ across $1 \leq k \leq p$. Moreover, we varied the family of the densities g in Assumption 2.2.1 to be the Laplace, the Logistic, or the Scaled-t distribution (10 degrees of freedom) scale-location families. Finally, ScoreLiNGAM and HighDimLiNGAM were run with knowledge of the true Markov blanket for each node, while DirectLiNGAM was not as it does not have this option. Afterward, the data matrix $\mathbf{X}$ was standardized so that each column has sample standard deviation equal to 1 and sample mean equal to 0.

Figure 2.2.1 reports the results in terms of order estimation error (lower is better), which we define as:

$$\frac{1}{p^2} \sum_{j=1}^{p} \sum_{k=1}^{p} \mathbf{1}\{\mathbf{B}_{jk} \neq 0, \hat{\pi}^{-1}(k) < \hat{\pi}^{-1}(j)\}.$$

Our ScoreLiNGAM achieved the highest accuracy for all high-dimensional settings ($n \leq p$). DirectLiNGAM became quite comparable until the sample size increased to $n = 2p$ and did a bit better than ScoreLiNGAM when $n \geq 10p$ (large sample size cases). Note that results are not presented for DirectLiNGAM when $n = 0.5p$ nor $n = p$, because it is not applicable for $n \leq p$. For the Andes network, results for DirectLiNGAM are also not presented as this procedure takes about 118 minutes for a single replicate, which adds up across 90 total replicates. On the other hand, HighDimLiNGAM is generally the least accurate algorithm across all networks and sample sizes. Recall that the data matrix $\mathbf{X}$ is re-scaled. The inaccuracy of HighDimLiNGAM is likely owed to the fact that this procedure is not invariant to a re-scaling of the data, as ScoreLiNGAM and DirectLiNGAM are. We also compared the three methods when the error distributions were mis-specified for ScoreLiNGAM (second and third columns of Figure 2.2.1). The true error distributions were Logistic or Scaled-t, but we still used the Laplace update (2.5) in ScoreLiNGAM. It is seen that its accuracy was comparable to the result when we correctly specified the error distributions (the other three columns), suggesting that our method is robust to model mis-specification.

In terms of speed, Figure 2.3.1 summarizes this for the win95pts network. The advantage of our method is speed, with our method being no less than 100 times faster the next fastest

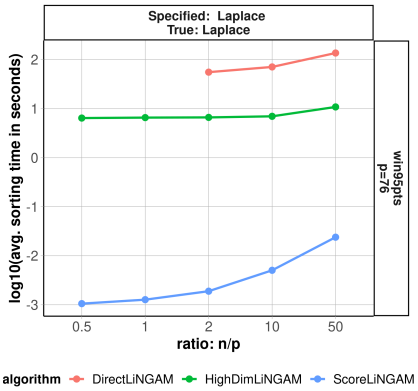

Figure 2.3.1: The $\log_{10}$(avg. sorting time in seconds) scale for the various methods applied to the win95pts network.

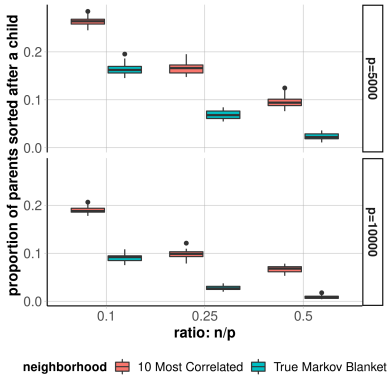

Figure 2.3.2: Sorting errors for ScoreLiNGAM under $p = 5000, 10000$ and $n = 0.1p, 0.25p, 0.5p$. Color indicates how the neighborhood sets are constructed.

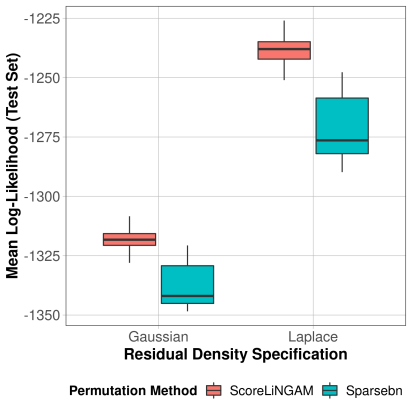

Figure 2.3.3: The mean log-likelihood on 1,000 genes for a subset of cells in the data of (Yao et al. 2021), across 50 repetitions.

method. (Note: HighDimLiNGAM's procedure is parallelized across 7 threads.) Appendix 2.C contains details about the implementation of each procedure, along with the machine used to run these experiments. Moreover, Figure 2.C.1 in Appendix 2.C contains sorting times for all the settings we considered.

2.3.2 Larger Network Results

Next, we simulated large networks with $p = 5000, 10000$ and $n = 0.1p, 0.25p, 0.5p$ to further demonstrate the scalability of ScoreLiGAM. We do not include results in these settings for DirectLiNGAM nor HighDimLiNGAM as they would take too long to run. The network generation is such that 5% of nodes are root nodes (no parents), and all other nodes have between 1 and 2 parents (with equi-probability) which are selected at random from the set of predecessors in a randomly generated permutation. Moreover, $\mathbf{B}_{jk} \overset{i.i.d.}{\sim} \text{Uniform}[-0.9, -0.4] \cup [0.4, 0.9]$ across (j, k) such that $j \in PA_k$, while $\theta_k \overset{i.i.d.}{\sim} \text{Uniform}[0.25, 0.9]$ across $1 \leq k \leq p$

is the scale parameter for the Laplace noise in the synthetic LiNGAM. A new LiNGAM is generated according to this schema for each data replicate.

Figure 2.3.2 presents simulation results for ScoreLiNGAM with two different a priori known neighborhood sets. "True Markov Blanket" means that we set $\widehat{N}_{kt} = MB_k$ for each $1 \leq k \leq p$ and run the sorting procedure with these oracle sets. The results for "10 Most Correlated" use 20% of the data to specify $\widehat{N}_{kt}$ as the 10 most Pearson-correlated variables (in absolute value) to X_k for each $1 \leq k \leq p$, and the other 80% of the data to estimate the topological ordering.

It is encouraging to see in Figure 2.3.2 that the accuracy of our method is high even for such a challenging high-dimensional setting. In fact, the average error rate is quite comparable to that for the smaller networks reported in Figure 2.2.1. As expected, an accurate neighborhood set provides better sorting results. Further, our method can run relatively quickly for large p, but its accuracy naturally is dictated by sample size. Figure 2.C.2 in Appendix 2.C contains the sorting times to go along with Figure 2.3.2.

2.3.3 Application: Single-Cell Gene Expression Data

We apply our method on the data of Yao et al. (2021)[1]. With it, we seek to estimate a linear SEM to model a gene regulatory network, where each X_k in Equation (2.1) is the expression level of a gene. We focus our attention on their dataset for which isolated single cells were processed for RNA sequencing using SMART-Seq v4 (labeled "Mouse Cortex+Hipppocampus (2019/2020)"). Noting the paper's finding that cells' gene expressions cluster according to region and cell type, we subset the data as follows. We focus on glutamatergic cells from the mice brains' primary visual cortex. We also focus on cells for which injection materials are not specified (see Saleeba et al. (2019) for background on neuronal tracers). This takes us from 74,973 cells down to 7,159–the largest subset of all cell class, isocortex location, and

[1]Available at `http://cells.ucsc.edu/?ds=allen-celltypes+mouse-cortex&meta=regionlabel` in compressed TSV format

injection material combinations. A sizable amount of genes had expression measurements of exactly 0, so we subset genes to those which were measured to be non-zero in 50% or more of these cells. This brings us from 45,768 to 10,012 genes.

2.3.3.1 Comparison to another scalable linear SEM estimation procedure

As for large simulated networks in Section 2.3.2, DirectLiNGAM and HighDimLiNGAM were too slow for this application. In order to compare ScoreLiNGAM to another linear structural equation modeling procedure, we applied the package `sparsebn` (Aragam et al. 2019) to our data, which is a score-based method that maximizes a regularized Gaussian likelihood over the DAG space (Aragam and Zhou 2015). To make comparisons across 50 repetitions, we randomly select 1,000 of the original 10,012 genes. For each repetition, we randomly sample 2,000 cells: half of the cells are designated to be in the training set, and the other half in test set; each data matrix is standardized such that columns have sample standard deviation 1 and sample mean 0.

In the training set, 20% of cells are randomly selected to estimate the Pearson correlation matrix. We specify the neighborhoods, $\widehat{N}_k$, for ScoreLiNGAM as the 50 genes $j \in [1000] \backslash \{k\}$ with the largest Pearson correlation (in absolute value) with gene k. The remaining 80% of training data is used to estimate a topological ordering and the linear SEM's coefficients (via ordinary least squares). For Sparsebn, no a priori neighborhood selection is used: parent sets for the linear SEM are learned with 100% of the training data using default options in the `estimate.dag` command, and the selection of the final DAG in the solution path is done by the recommended `select.parameter` command. For Sparsebn, the linear SEM's model parameters are estimated according to the selected DAG. Moreover, the noise densities we fit to the residuals in the training set are either Gaussian or Laplace.

As can be seen in Figure 2.3.3, the Laplace density specification for the additive errors provides a significantly higher mean log-likelihood on the test set compared to a Gaussian density for both methods. This shows that the Laplace distribution, with its thicker tails

than the Gaussian distribution, fits this data better. Furthermore, ScoreLiNGAM showed substantially higher test-data likelihood than Sparsebn under both error distributions for calculating the likelihood.

2.3.3.2 Application of ScoreLiNGAM to All 10,012 Genes

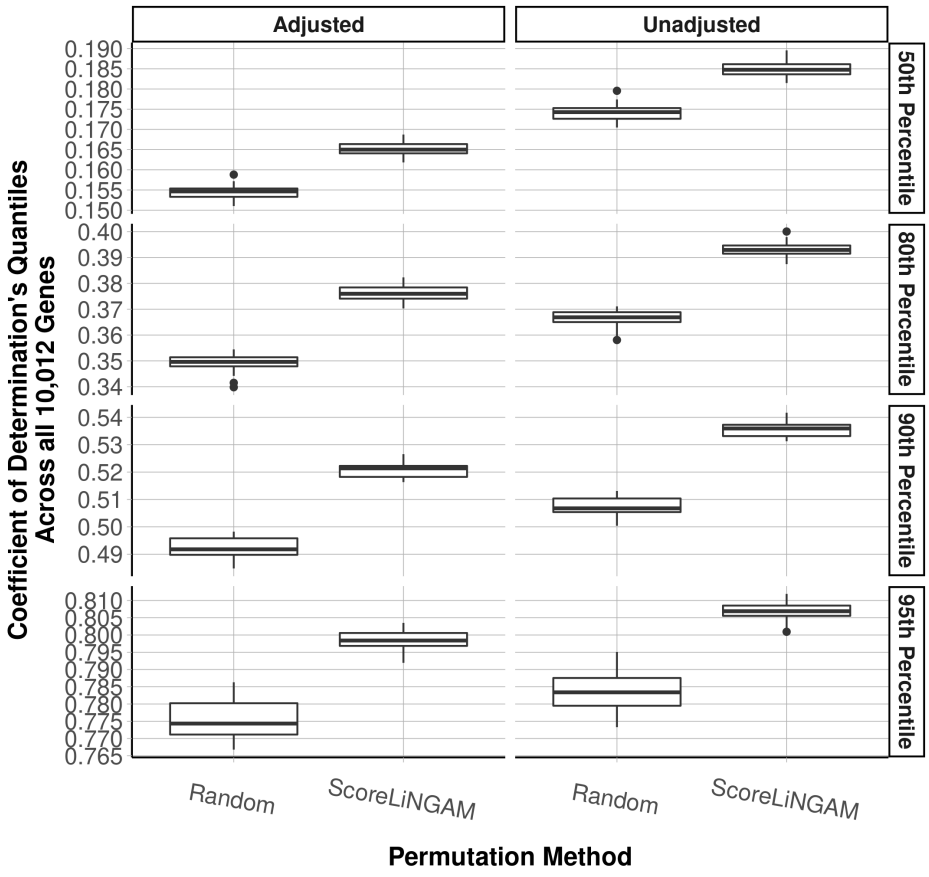

Figure 2.3.4: Across 30 replications, a comparison of the estimated coefficient of determination on Fold 2B for each gene. We summarize the coefficient of determination across genes by taking the median, 80th, 90th, and 95th percentiles.

We now present the application of ScoreLiNGAM to all original $p = 10,012$ genes discussed at the start of this section. The application is as follows:

1. We randomly split the gene expression measurement matrix with 7,159 cells (rows) into two folds having 358 ($\sim 5\%$) and 6801 ($\sim 95\%$) of the cells, respectively.

2. On the first fold, we ran neighborhood linear regressions in which we restrict coefficients to be non-negative via `R`'s `glmnet` package (**?**) with no ridge or lasso regularization. We then selected the sets $\widehat{N}_k$ to correspond to genes $\widehat{N}_k \subseteq [p] \backslash \{k\}$ such that coefficients are non-zero. The non-negative coefficient constraint for each linear regression, known as non-negative least squares, can itself be seen as a form of regularization that gives a sparse solution to the coefficient vectors (**?**). Indeed, this constraint in the neighborhood regressions resulted in neighborhood sets of size 45 to 277 genes (of 10,011 possible genes), with a median of 145 genes per neighborhood set. Our use of non-negative least squares regression is motivated by the use of non-negative linear regression to impute single-cell gene expression measurements (**?**). It is also computationally faster compared to neighborhood Lasso regression.

3. We then randomly split the second fold into two folds, 2A and 2B, having 3401 ($\sim 47.5\%$) and 3400 ($\sim 47.5\%$) of the original cells, respectively.

4. On Fold 2A, we estimated the permutation $\hat{\pi}$ using ScoreLiNGAM.

5. On Fold 2B, the validation fold, we estimated via linear least squares regression the 10,012 by 10,012 weighted adjacency matrix for the DAG corresponding to the topological ordering defined by $\hat{\pi}$ and such that the support of the k-th column is the set of indices $\widehat{N}_k \cap \{\hat{\pi}(j)\}_{0<j<\hat{\pi}^{-1}(k)}$.

6. Using the weighted adjacency matrix estimate from the previous step, we then calculated linear least squares residuals from the predictions given by the estimated parent set for each node k, on Fold 2B.

7. With the residuals of the last step, we calculated the coefficient of determination (R_k^2): for each gene $k \in [10,012]$, R_k^2 is an estimate of the proportion of variation explained by

linearly regressing gene k's measurement on the estimated parent genes. It is 1 minus the ratio of gene k's residual sum of squares and the total sum of squares.

The result of this application is summarized in Figure 2.3.4. In order to make a comparison, we also calculated residuals from the linear SEM induced by a randomly generated topological ordering, denoted in Figure 2.3.4 as "Random." The two linear SEMs, with topological ordering given by ScoreLiNGAM or randomly generated, select parent sets as the intersection of a node's neighborhood set, $\widehat{N}_k$, and the node's predecessors in the corresponding topological ordering if any. We repeat Steps 3-7 a total of 30 times.

Considering that some genes may have more estimated parents than others, and that a coefficient of determination can be artificially large as the number of regressors increases, Figure 2.3.4 also includes the adjusted coefficient of determination which incorporates a penalty for the number of regressors (**?**). The adjusted coefficient of determination is

$$1 - (1 - R_k^2)\frac{n_B - 1}{n_B - |\widehat{PA}_k|},$$

where $|\widehat{PA}_k|$ is the number of estimated parents for node k and $n_B = 3400$ is the sample size in Fold 2B.

As we can see from Figure 2.3.4, ScoreLiNGAM gives higher coefficients of determination (adjusted and unadjusted) on the test datasets (validation fold 2B) compared to a randomly generated permutation across all random replications–as summarized by the median, 80th, 90th, and 95th percentiles taken across the 10,012 coefficients of determination. Taking for granted the linearity assumption, the above higher-end percentiles of the R_k^2 across all genes provide meaningful comparisons because it may very well be the case that a majority of the genes have quite random expression patterns. Based on the 90th percentile for the adjusted coefficient of determination in Figure 2.3.4, it appears that for 10% of all the genes, more than 51% of their expression variation is explainable by its estimated parents in the linear SEM given by ScoreLiNGAM. As shown by the 95th percentiles in Figure 2.3.4, for the top 5% of genes in terms of adjusted R^2, 79% or more of the genes' expression variation is

explainable by its estimated parents in the linear SEM given by ScoreLiNGAM. These R^2 levels are significantly higher than those given by random permutations, with no overlap in the boxplots in Figure 2.3.4.

Across the 30 replications, ScoreLiNGAM's sorting time for all 10,012 genes had a median of 10.28 minutes, confirming its scalability for such large and high-dimensional datasets of $p > 10,000$ and $n > 3,000$.

2.4 Discussion

In this chapter, we demonstrated that sequentially applying Algorithm 1 can give promising structure learning results. We demonstrated this with a novel sequential procedure based on parametric specification that provides an alternative to the state of the art for the identifiability and estimation of a linear DAG model with non-Gaussian errors. We discussed the conditions, Assumptions 2.2.1 and 2.2.2, under which the proposed causal discovery procedure will identify the valid DAG. We also proposed a relatively simple procedure that can make strategic use of an a priori known neighborhood set for each node. Finally, we presented numerical evidence that our procedure scales to large dimensions, which is otherwise not the case for the state-of-the-art for LiNGAM. We accompanied these simulations with a real-data application. Further extensions of the work presented here include formal statistical guarantees along with extensions of the likelihood ratio approach to nonlinear SEMs.

As a practical manner, consider prospective applications to single cell gene expression data (scRNA-seq). Recent work suggests a hierarchical structure between true (hidden) expressions and measured expressions with missing, and possibly zero, counts (Sarkar and Stephens 2021). Should the hierarchical nature be justified, further work on causal discovery for gene co-expression models may need to incorporate the fact that what we really would like is a graphical model, possibly causal if a domain expert agrees, on the true (hidden) expressions. Along these lines, future causal discovery procedures for such data can build

on the procedures of McDavid et al. (2019) and Yu et al. (2020), which themselves build on Gaussian graphical models, using a heavier tail distribution for residuals as we do here.

APPENDIX

2.A Greedy Choice of a Factor to Optimize the Joint Likelihood Function

Let vector $X \sim f(x)$, where $f(x)$ corresponds to the density induced by the generative model in Definition 2.1.2. Consider X's expected log-likelihood as a function of the permutation π:

$$\mathcal{L}(\pi) = \sum_{j=1}^{p} \mathbb{E}_{X \sim f(x)} \left[\log g \left(X_j - [\mathbf{B}^{\pi}_{\cdot j}]^T X; \theta^{\pi}_j \right) \right], \tag{2.6}$$

Here, $\mathbf{B}^{\pi}$ is the acyclic weighted adjacency matrix that arises from a population-level least squares objective such that the $\pi(j)$-th column is given by:

$$\mathbf{B}^{\pi}_{\cdot \pi(j)} = \arg \min_{\substack{\theta \in \mathbb{R}^{p \times 1}: \ \theta_k = 0 \ \forall k \\ \text{s.t } \pi^{-1}(k) \geq j}} \mathbb{E} \left[(X_{\pi(j)} - \theta^T X)^2 \right].$$

That is, the column $\mathbf{B}^{\pi}_{\cdot \pi(j)}$ is comprised of the least squares coefficients when linearly regressing $\pi(j)$ onto its predecessors, if any, in the ordering given by π. Moreover, θ^{π}_j is the corresponding scale parameter according to Assumption 2.2.1. Now let ϕ^{π}_j be the density for the Gaussian distribution having the same first two moments as:

$$R^{\pi}_j := X_j - [\mathbf{B}^{\pi}_{\cdot j}]^T X.$$

Define

$$\tilde{\mathcal{L}}(\pi) := \sum_{j=1}^{p} \mathbb{E}_{X \sim f(x)} \left[\log \phi^{\pi}_j \left(R^{\pi}_j \right) \right] \text{ and } \kappa := \mathbb{E}_{\tilde{X} \sim f(x)} \left[\log \mathcal{N}(\tilde{X}; \mathbb{E}[X], \mathbf{V}[X]) \right].$$

Here, $\mathcal{N}(x; \mathbb{E}[X], \mathrm{Var}[X])$ denotes the density for a p-variate Gaussian distribution with the same first and second order moments as X. Due to the relation between $\mathbf{B}^{\pi}$ and the generalized Cholesky factorization of $\mathbf{V}[X]$, Ye et al. (2021a) shows that we actually have the equality:

$$\tilde{\mathcal{L}}(\pi) = \kappa. \tag{2.7}$$

Thus, maximizing (2.6) with respect to π is the same as maximizing the expected log-likelihood ratio given by:

$$(\mathcal{L} - \tilde{\mathcal{L}})(\pi) = \sum_{j=1}^{p} \mathbb{E}_{X \sim f(x)} \left[\log \frac{g\left(R_j^\pi; \theta_j^\pi\right)}{\phi_j^\pi\left(R_j^\pi\right)} \right] = \mathcal{L}(\pi) - \kappa. \tag{2.8}$$

With all this in mind, we can think of our choice of a node to append to the ordering $\mathcal{A}_t$ at step t as greedily choosing the largest summand,

$$\mathbb{E}_{X \sim f(x)} \left[\log \frac{g\left(R_{\hat{\pi}(t)}^{\hat{\pi}}; \theta_{\hat{\pi}(t)}^{\hat{\pi}}\right)}{\phi_{\hat{\pi}(t)}^{\hat{\pi}}\left(R_{\hat{\pi}(t)}^{\hat{\pi}}\right)} \right]$$

to add to the known log-likelihood ratio at step t:

$$(\mathcal{L} - \tilde{\mathcal{L}})_t(\hat{\pi}) := \begin{cases} 0 & t = 1 \\ \sum_{j=1}^{t-1} \mathbb{E}_{X \sim f(x)} \left[\log \frac{g\left(R_{\hat{\pi}(j)}^{\pi}; \theta_{\hat{\pi}(j)}^{\hat{\pi}}\right)}{\phi_{\hat{\pi}(j)}^{\hat{\pi}}\left(R_{\hat{\pi}(j)}^{\pi}\right)} \right] & 2 \leq t \leq p+1 \end{cases}.$$

That is, our sequential application of Algorithm 1 is attempting to greedily maximize (2.8) one summand at a time.

2.B Proof of Theorem 2.2.5

2.B.0.1 Proof sketch for Theorem 2.2.5

The formal proof of Theorem 2.2.5 in § 2.B.1 below is a relatively straightforward inductive application of the following reasoning after applying Algorithm 1 at any given step t. Key to the proof, we note that (2.3) can also be written equivalently as the difference of two KL-divergence terms:

$$\arg\max_{k \notin \mathcal{A}_t} \{ D_{KL}\left(f_{kt}(r_{kt}) || \phi(r_{kt}; \sigma_{kt})\right) - D_{KL}\left(f_{kt}(r_{kt}) || g_k(r_{kt}; \eta_{kt})\right) \}. \tag{2.9}$$

Lemma 2.B.1 suggests that invalid nodes' residuals, R_{kt}, are a linear combination of two or more entries in the vector ϵ, while for valid nodes ℓ we have $R_{\ell t} = \epsilon_\ell$. Under Assumption

2.2.2, this means that the term $D_{KL}\left(f_{kt}(r_{kt})||g_k(r_{kt};\eta_{kt})\right)$ in (2.9) will be zero only if node k is valid to continue the ordering at step t. The natural follow up question is what the behavior is for the term $D_{KL}\left(f_{kt}(r_{kt})||\phi(r_{kt};\sigma_{kt})\right)$ in (2.9) when k is valid vs. invalid to continue the ordering. Lemma 2.B.6 provides this insight: for valid nodes to continue an ordering, this term's value is no less than the same term's value for invalid nodes.

In light of Lemma 2.B.1, Lemma 2.B.6 makes sense under a Central Limit Theorem-like argument: a sum of two or more random variables is closer to Gaussian than each summand alone. Of particular note, a key result that helps show why Lemma 2.B.6 holds is *Theorem 17.8.1* of (Cover and Thomas 2005), a restatement of the entropy-power inequality. This restatement says that the differential entropy for a sum of any two independent random variables, U and V, is no less than the differential entropy for the sum of two strategically defined Gaussian random variables, each having the same differential entropy as U and V (rather than the same first two moments), respectively.

2.B.1 Formal Proof of Theorem 2.2.5

Proof of Theorem 2.2.5.
Our proof boils down to making the correct decision in Algorithm 1 at step 1, then making the correct choice at step 2 assuming the choice in step 1 was correct, and so on.

For the sake of induction, let us assume that $\mathcal{A}_t$ is correct in the sense that $PA_a \subseteq \mathcal{A}_t$ for all $a \in \mathcal{A}_t$. This is true at the base case $t = 1$ when $\mathcal{A}_t = \emptyset$, since having made no ordering choices also means we have made no mistakes.

Let $k \in S_t$ be an invalid node to continue the ordering in the sense that $PA_k \cap \mathcal{A}_t \neq \emptyset$. And let $\ell \in S_t$ be a valid node to continue the ordering in the sense that $PA(\ell) \subseteq \mathcal{A}_t$.

Lemma 2.B.6 tells us that the least squares residual $R_{\ell t} \sim f_{\ell t}(r_{\ell t})$ is no closer to Gaussian

than $R_{kt} \sim f_{kt}(r_{kt})$ in the sense that:

$$D_{KL}\left(f_{kt}(r_{kt})||\phi(r_{kt};\sigma_{kt})\right) \leq D_{KL}\left(f_{\ell t}(r_{\ell t})||\phi_{\ell t}(r_{\ell t})\right)$$

Furthermore, regularity Assumption 2.2.2 ensures that:

$$D_{KL}\left(f_{kt}(r_{kt})||g_k(r_{kt};\eta_{kt})\right) > 0.$$

On the other hand, so long as we properly specified the error density for node ℓ, we have that:

$$D_{KL}\left(f_{\ell t}(r_{\ell t})||g_{\ell}(r_{\ell t};\eta_{\ell t})\right) = 0.$$

Thus,

$$\begin{aligned}
\mathbb{E}_{f_{kt}(r_{kt})}\left[\log\frac{g_k(R_{kt};\eta_{kt})}{\phi(r_{kt};\sigma_{kt})}\right] &= D_{KL}\left(f_{kt}(r_{kt})||\phi(r_{kt};\sigma_{kt})\right) - D_{KL}\left(f_{kt}(r_{kt})||g_k(r_{kt};\eta_{kt})\right) \\
< \mathbb{E}_{f_{\ell t}(r_{\ell t})}\left[\log\frac{g_{\ell}(R_{\ell t};\eta_{\ell t})}{\phi_{\ell t}(R_{\ell t})}\right] &= D_{KL}\left(f_{\ell t}(r_{\ell t})||\phi_{\ell t}(r_{\ell t})\right).
\end{aligned}$$

Altogether, this implies that

$$\max_{j\in S_t}\mathcal{S}(j,\mathcal{A}_t) > \mathcal{S}(k,\mathcal{A}_t).$$

and

$$\ell = \arg\max_{j\in S_t}\mathcal{S}(j,\mathcal{A}_t),$$

since ℓ and k were arbitrary valid and invalid nodes, respectively.

So at step t, we will always make the correct choice for a node to continue the ordering. □

2.B.2 Proofs of Lemma 2.B.1 and Lemma 2.B.6

In this section, we formally prove Lemma 2.B.1, Lemma 2.B.6.

2.B.2.1 Some Useful Shorthand Notation

Let us define some new strategic sets which contain indices in $[p]$, and review some we have been using already.

- The set
$$\mathcal{A}_t = \begin{cases} \emptyset & t = 1 \\ \{\hat{\pi}(1), \ldots, \hat{\pi}(t-1)\} & t \geq 2 \end{cases}.$$
This is the partial ordering at step $t = 1, 2, \ldots$. In our population-level identification results, we will typically assume it is correct at step t, which means that for all $a \in \mathcal{A}_t$, $PA_a \subset \mathcal{A}_t$.

- $S_t = [p] \backslash \mathcal{A}_t$ is the set of unordered nodes at step t.

- $MB_k = PA_k \cup CH_k \cup_{j \in CH_k} PA_j$ is the Markov Blanket of node k.

- $\widehat{N}_k$ is the Markov Blanket superset such that $\widehat{N}_k \supseteq MB_k$. In finite data, we will typically estimate $\widehat{N}_k$ by a procedure such as neighborhood lasso regression, so this containment may not hold. For the sake of this section, because we are deriving quantities at the population-level, we assume that $\widehat{N}_k$ is known and contains the true Markov blanket. Note that trivially, we may consider $\widehat{N}_k = [p] \backslash \{k\}$, and the results of this section would still hold.

- $\widehat{N}_{kt} = \mathcal{A}_t \cap \widehat{N}_k$ is the intersection of the Markov blanket superset with the partial ordering.

- $L_{kt} = \bigcup_{j \in \widehat{N}_{kt}} \{j\} \cup AN_j$, which are either nodes of $\widehat{N}_{kt}$ or ancestors of nodes in $\widehat{N}_{kt}$. When $\mathcal{A}_t$ is correct, it is necessarily the case that $L_{kt} \subseteq \mathcal{A}_t$ for each $k \notin \mathcal{A}_t$.

- L_{kt}^C, the complement of set L_{kt} which either contains nodes in $\mathcal{A}_t$ which are not in L_{kt}, i.e. the nodes of $\mathcal{A}_t \backslash L_{kt}$, or which are unordered, i.e. we have that $S_t \subseteq L_{kt}^C$.

Note that for each node $k \in S_t$ we can write:

$$X_k = \mathbf{M}_{k\cdot}\epsilon = \mathbf{M}_{kL_{kt}}\epsilon_{L_{kt}} + \mathbf{M}_{kL_{kt}^C}\epsilon_{L_{kt}^C}, \tag{2.10}$$

where the second equality holds since $L_{kt} \cup L_{kt}^C = [p]$. We can similarly write

$$X_{\widehat{N}_{kt}} = \mathbf{M}_{\widehat{N}_{kt}\cdot}\epsilon = \mathbf{M}_{\widehat{N}_{kt}L_{kt}}\epsilon_{L_{kt}}. \tag{2.11}$$

We omit a term with $\epsilon_{L_{kt}^C}$ since by definition of L_{kt}, the sub-mixing matrix $\mathbf{M}_{\widehat{N}_{kt}L_{kt}^C}$ is a zero matrix.

Combining (2.10) and (2.11),

$$R_{kt} = \left(\mathbf{M}_{kL_{kt}} - \beta_{kt}^T\mathbf{M}_{\widehat{N}_{kt}L_{kt}}\right)\epsilon_{L_{kt}} + \mathbf{M}_{kL_{kt}^C}\epsilon_{L_{kt}^C},$$

which we will make use of in the proof for Lemma 2.B.1 below.

2.B.2.2 Lemma 2.B.1: Characterizing nodes' residuals as linear combinations of independent components

Lemma 2.B.1. *Assume that $\mathcal{A}_t$ is correct so far in the sense that for each $a \in \mathcal{A}_t$, we have $PA_a \subseteq \mathcal{A}_t$. Also assume Assumption 2.2.3 holds. We have that:*

- *If $k \in S_t$ is a valid node to continue the ordering, i.e. $PA_k \subseteq \mathcal{A}_t$, then:*

$$R_{kt} = X_k - \beta_{kt}^T X_{\widehat{N}_{kt}} = \epsilon_k.$$

- *Otherwise, if k is not a valid node, then R_{kt} is a linear combination of more than one independent component in ϵ.*

Proof of Lemma 2.B.1.

Case 1: Assume k is a valid node to continue the ordering in the sense that $PA_k \subseteq \mathcal{A}_t$. We want to show that $R_{kt} = \epsilon_k$. In this case, $\mathbf{M}_{kL_{kt}^C}$ has a non-zero entry corresponding to only

$\mathbf{M}_{kk} = 1$. This is because $AN_k = \text{support}\left(\mathbf{M}_{k\cdot}\right) \setminus \{k\} \subseteq L_{kt}$, which in turn holds because $PA_k \subseteq MB_k \cap \mathcal{A}_t \subseteq \widehat{N}_{kt}$. Thus we have

$$\mathbf{M}_{kL_{kt}^C} \epsilon_{L_{kt}^C} = \epsilon_k.$$

So we have left to show that

$$\left(\mathbf{M}_{kL_{kt}} - \beta_{kt}^T \mathbf{M}_{\widehat{N}_{kt} L_{kt}}\right) \epsilon_{L_{kt}} = 0.$$

Recall that $\mathbf{B}$ is the weighted adjacency matrix for the underlying LiNGAM. We have that $\text{support}(\mathbf{B}_{\cdot k}) = PA_k$. Let us index the entries of the column vector $\mathbf{B}_{\cdot k}$ by $\widehat{N}_{kt}$ and denote this as $\mathbf{B}_{\widehat{N}_{kt} k}$. One thing that could be helpful to prove is that if k is valid, then:

$$\beta_{kt} = \mathbf{B}_{\widehat{N}_{kt} k}.$$

Because $\text{support}(\mathbf{B}_{\cdot k}) = PA_k$ and $PA_k \subseteq \widehat{N}_{kt}$, consider that

$$X_k = X^T \mathbf{B}_{\cdot k} + \epsilon_k = X_{\widehat{N}_{kt}}^T \mathbf{B}_{\widehat{N}_{kt} k} + \epsilon_k,$$

with $\epsilon_k \perp\!\!\!\perp X_{\widehat{N}_{kt} k}$ and $\mathbb{E}[\epsilon_k] = 0$. Thus,

$$\begin{aligned} \beta_{kt} &= \left(\mathbb{E}\left[X_{\widehat{N}_{kt}} X_{\widehat{N}_{kt}}^T\right]\right)^{-1} \left(\mathbb{E}\left[X_{\widehat{N}_{kt}} X_{\widehat{N}_{kt}}^T\right] \mathbf{B}_{\widehat{N}_{kt} k} + \mathbb{E}\left[X_{\widehat{N}_{kt}} \epsilon_k\right]\right) \\ &= \left(\mathbb{E}\left[X_{\widehat{N}_{kt}} X_{\widehat{N}_{kt}}^T\right]\right)^{-1} \mathbb{E}\left[X_{\widehat{N}_{kt}} X_{\widehat{N}_{kt}}^T\right] \mathbf{B}_{\widehat{N}_{kt} k} \\ &= \mathbf{B}_{\widehat{N}_{kt} k}, \end{aligned} \tag{2.12}$$

as we wanted.

It follows that $X_k = \mathbf{B}_{\widehat{N}_{kt} k}^T X_{\widehat{N}_{kt}} + \epsilon_k = \beta_{kt}^T X_{\widehat{N}_{kt}} + \epsilon_k$. This then means that $R_{kt} = X_k - \beta_{kt}^T X_{\widehat{N}_{kt}} = \epsilon_k$, as we wanted to show.

Case 2: Assume k is not a valid node. All we need in this case for our identifiability proof is that R_{kt} is a linear combination of more than one independent component. This is the case because if k is invalid to continue the ordering, then we have that there exists at least one $j \in PA_k$ such that $j \in S_t$ (unordered) and therefore $j \in L_{kt}^C$. Recall that:

$$R_{kt} = \left(\mathbf{M}_{kL_{kt}} - \beta_{kt}^T \mathbf{M}_{\widehat{N}_{kt}L_{kt}}\right)\epsilon_{L_{kt}} + \mathbf{M}_{kL_{kt}^C}\epsilon_{L_{kt}^C}.$$

Note that it is necessarily the case that $\mathbf{M}_{kj} \neq 0$, otherwise $j \notin PA_k$. Thus, R_{kt} includes the sum $\mathbf{M}_{kj}\epsilon_j + \epsilon_k$. That is, R_{kt} in this case is a linear combination of more than one independent component in ϵ. Note that R_{kt} could be a linear combination of more entries in ϵ, in addition to ϵ_j and ϵ_k.

□

2.B.2.3 Some Information Theory Definitions and Results

We now present some straightforward information theoretic results. They are meant to help demonstrate that our surrogate optimization (now a likelihood ratio) approach for Algorithm 1 leads to the identifiability of a causal order. These lemmas are used later to prove Lemma 2.B.6, a key result that says valid nodes j in a LiNGAM are no closer to Gaussian compared to invalid nodes k, conditional on the nodes in $\widehat{N}_{jt}$ and $\widehat{N}_{kt}$, respectively.

Definition 2.B.2 (Differential Entropy)**.**
For a continuous random variable X with density $p(x)$, denote $\mathbb{E}_{p(x)}[\cdot]$ to be expectation with respect to $p(x)$. The differential entropy of X is given by:

$$\mathbf{h}(X) = \mathbb{E}_{p(x)}\left[\log \frac{1}{p(X)}\right].$$

Lemma 2.B.3 (Restatement of the entropy power inequality)**.**
Consider two independent random variables $X \sim p(x)$ and $Y \sim p(y)$, and let $X' \sim \mathcal{N}(\mathbb{E}[X'], \mathbf{V}[X'])$ and $Y' \sim \mathcal{N}(\mathbb{E}[Y'], \mathbf{V}[Y'])$ be independent random variables such that $\mathbf{h}(X) = \mathbf{h}(X')$ and $\mathbf{h}(Y) = \mathbf{h}(Y')$. Then:

$$\mathbf{h}(X + Y) \geq \mathbf{h}(X' + Y').$$

Proof. This is exactly *Theorem 17.8.1* of (Cover and Thomas 2005), so we refer the reader to their proof.

□

Lemma 2.B.4 (KL Divergence from Gaussianity)**.**
Let $X \sim p(x)$ and $q(x)$ the density for $\tilde{X} \sim \mathcal{N}(\mathbb{E}[X], Cov[X])$.

$$D_{KL}(p(x)||q(x)) = \mathbf{h}(\tilde{X}) - \mathbf{h}(X). \tag{2.13}$$

As in Lemma 2.B.3, let $X' \sim \mathcal{N}(\mathbb{E}[X'], \mathbf{V}[X'])$ such that $\mathbf{h}(X) = \mathbf{h}(X')$. We can equivalently write the KL divergence from Gaussianity as:

$$D_{KL}(p(x)||q(x)) = \frac{1}{2}\log\left(\frac{\mathbf{V}[X]}{\mathbf{V}[X']}\right).$$

Proof.
Because

$$\mathbf{h}(\tilde{X}) = \mathbb{E}_{\tilde{X}\sim q(x)}\left\{\log\frac{1}{q(\tilde{X})}\right\} = \mathbb{E}_{X\sim p(x)}\left\{\log\frac{1}{q(X)}\right\},$$

by properties of this normal distribution (namely, that $\mathbb{E}[\log q(X)] \propto \mathbf{V}[X] = \mathbf{V}[\tilde{X}]$) we have that:

$$D_{KL}(p(x)||q(x)) = \mathbf{h}(\tilde{X}) - \mathbf{h}(X).$$

Noting that the differential entropy for any $\mathcal{N}(\mu, \sigma^2)$ is $\frac{1}{2}\log(2\pi e\sigma^2)$ and our assumption that $\mathbf{h}(X) = \mathbf{h}(X')$, we arrive at the second equality:

$$D_{KL}(p(x)||q(x)) = \frac{1}{2}\log\left(\frac{2\pi e\mathbf{V}[X]}{2\pi e\mathbf{V}[X']}\right).$$

Note that also $\mathbf{V}[X] = \mathbf{V}[\tilde{X}] \geq \mathbf{V}[X'] \iff D_{KL}(p(x)||q(x)) \geq 0$, which is the case because KL-divergence is always non-negative.

□

This well known result also implies that the normal distribution is the maximum entropy distribution when we constrain the first and second order moments of each distribution to be the same.

Lemma 2.B.5 (Same distance to Gaussianity)**.**

Let $\tilde{\epsilon}_k \sim \mathcal{N}(0, \mathbf{V}[\epsilon_k])$ *with density* $q_k(\cdot)$ *for each* $k = 1, 2, \ldots, p$*. Also let* $\epsilon'_k \sim \mathcal{N}(0, \mathbf{V}[\epsilon'_k])$ *such that* $\mathbf{h}(\epsilon'_k) = \mathbf{h}(\epsilon_k)$*. If* ϵ *in our LiNGAM satisfies Assumption 2.2.1, then there exists a constant* $\gamma \geq 0$ *such that*

$$D_{KL}\left(g(\epsilon_k; \theta_k) || q_k(\epsilon_k)\right) = \gamma$$

and

$$\frac{\mathbf{V}[\epsilon_k]}{\mathbf{V}[\epsilon'_k]} = \tilde{\gamma} = \exp(2\gamma)$$

for all $k = 1, 2, \ldots, p$*.*

Proof.

From Lemma 2.B.4, we have that:

$$D_{KL}\left(g(\epsilon_k; \theta_k) || q_k(\epsilon_k)\right) = \mathbf{h}(\tilde{\epsilon}_k) - \mathbf{h}(\epsilon'_k).$$

Noting Assumption 2.2.1 and properties of differential entropy under a rescaling, it follows that for $U \sim g(\cdot; \theta_0)$:

$$\mathbf{h}(\epsilon_k) = \mathbf{h}(U) + \log(\theta_k/\theta_0).$$

Let $U' \sim \mathcal{N}(0, \mathbf{V}[U'])$ such that $\mathbf{h}(U') = \mathbf{h}(U)$. We have also that

$$\mathbf{h}(\epsilon'_k) = \mathbf{h}(U') + \log(\theta_k/\theta_0),$$

based on the construction of both ϵ'_k and U'.

Similar to $\tilde{\epsilon}_k$, let $\tilde{U} \sim \mathcal{N}(0, \mathbf{V}[U])$. Thus, regardless of $k = 1, 2, \ldots, p$, we have that:

$$\frac{1}{2}\log\left(\frac{2\pi e \mathbf{V}[\epsilon_k]}{2\pi e \mathbf{V}[\epsilon'_k]}\right) = \mathbf{h}(\tilde{\epsilon}_k) - \mathbf{h}(\epsilon'_k) = \mathbf{h}(\tilde{U}) - \mathbf{h}(U') =: \gamma.$$

□

2.B.2.4 Lemma 2.B.6: KL Divergence from Gaussianity for valid and invalid nodes' residuals

Lemma 2.B.6. *Let $X \in \mathbb{R}^p$ be a LiNGAM from Definition 2.1.2 that satisfies Assumptions 2.2.3 and 2.2.1. Assume that $\mathcal{A}_t$ is correct in the sense that $PA_a \subseteq \mathcal{A}_t$ for all $a \in \mathcal{A}_t$. Let $k \in [p]\backslash\mathcal{A}_t$ be an invalid node to continue the ordering in the sense that there exists $j \in PA_k$ such that $j \in [p]\backslash\mathcal{A}_t$. And let $\ell \in [p]\backslash\mathcal{A}_t$ be a valid node to continue the ordering in the sense that $PA(\ell) \subseteq \mathcal{A}_t$. Then the least squares residual $R_{\ell t} \sim f_{\ell t}(r_{\ell t})$ is no closer to Gaussian than $R_{kt} \sim f_{kt}(r_{kt})$ in the sense that:*

$$D_{KL}\left(f_{kt}(r_{kt})||\phi(r_{kt};\sigma_{kt})\right) \leq D_{KL}\left(f_{\ell t}(r_{\ell t})||\phi_{\ell t}(r_{\ell t})\right), \tag{2.14}$$

where ϕ_{kt} and $\phi_{\ell t}$ are the respective densities for

$$\tilde{R}_{kt} \sim \mathcal{N}(\mathbb{E}[R_{kt}], \mathbf{V}[R_{kt}]) \text{ and } \tilde{R}_{\ell t} \sim \mathcal{N}(\mathbb{E}[R_{\ell t}], \mathbf{V}[R_{\ell t}]).$$

Proof of Lemma 2.B.6.
For each $j \in [p]$, let ϵ'_j be a normally distributed random variable such that $\mathbf{h}(\epsilon'_j) = \mathbf{h}(\epsilon_j)$, while $\tilde{\epsilon}_j$ is distributed as $\mathcal{N}(\mathbb{E}[\epsilon_j], \mathbf{V}[\epsilon_j])$. Here, for all $j, k \in \{1, 2, \ldots, p\}$, $\tilde{\epsilon}_j \perp\!\!\!\perp \tilde{\epsilon}_k$ (unless $j = k$) and $\tilde{\epsilon}_j \perp\!\!\!\perp \epsilon'_k$ (even if $j = k$).

Recall also that for $j \in S_t$

$$R_{jt} = \left(\mathbf{M}_{jL_{kt}} - \beta_{jt}^T \mathbf{M}_{\widehat{N}_{jt}L_{tj}}\right)\epsilon_{L_{tj}} + \mathbf{M}_{jL_{tj}^C}\epsilon_{L_{tj}^C} = \sum_{i\in[p]} \delta_{ij}\epsilon_i,$$

where the coefficients δ_{ij} in the last equality are used for shorthand. Note that $\delta_{jj} = 1$ always. And if j is invalid to continue the ordering, then also $\delta_{ij} \neq 0$ for at least one other $i \in [p]\backslash\{j\}$, based on Lemma 2.B.1.

The relation between the quantities of interest is as follows:

$$
\begin{aligned}
D_{KL}\left(f_{kt}(r_{kt})||\phi(r_{kt};\sigma_{kt})\right) &= \mathbf{h}(\tilde{R}_{kt}) - \mathbf{h}\left(R_{kt}\right) && \text{by Lemma 2.B.4} \\
&= \mathbf{h}\left(\sum_{i\in[p]} \delta_{ik}\tilde{\epsilon}_i\right) - \mathbf{h}\left(\sum_{i\in[p]} \delta_{ik}\epsilon_i\right) && \text{Notice: } \tilde{R}_{kt} \overset{d}{=} \sum_{i\in[p]} \delta_{ik}\tilde{\epsilon}_i \\
&\leq \mathbf{h}\left(\sum_{i\in[p]} \delta_{ik}\tilde{\epsilon}_i\right) - \mathbf{h}\left(\sum_{i\in[p]} \delta_{ik}\epsilon'_i\right) && \text{by Lemma 2.B.3} \\
&= \frac{1}{2}\log\left(\frac{2\pi e\sum_{i\in[p]}\delta_{ik}^2\text{Var}\left[\epsilon_i\right]}{2\pi e\sum_{i\in[p]}\delta_{ik}^2\text{Var}\left[\epsilon'_i\right]}\right) && \text{by normality of the } \tilde{\epsilon}_i, \epsilon'_i \\
&= \frac{1}{2}\log\left(\frac{2\pi e\tilde{\gamma}\sum_{i\in[p]}\delta_{ik}^2\text{Var}\left[\epsilon'_i\right]}{2\pi e\sum_{i\in[p]}\delta_{ik}^2\text{Var}\left[\epsilon'_i\right]}\right) && \text{by Lemma 2.B.5} \\
&= \gamma \\
&= D_{KL}\left(f_{\ell t}(r_{\ell t})||\phi_{\ell t}(r_{\ell t})\right),
\end{aligned}
\tag{2.15}
$$

as we wanted (Recall $R_{\ell t} = \epsilon_\ell$ by Lemma 2.B.1).

□

2.C More Figures

2.C.1 Sorting Time for Small Networks

The general takeaway of Figure 2.C.1 is that ScoreLiNGAM is generally much faster. Consider the largest DAG, the Andes network ($p = 223$), where the sorting time of ScoreLiNGAM is typically under 1 second across all sample sizes, while for HighDimLiNGAM (parallelized across 7 threads) the sorting procedure takes between 10-1000 seconds across sample sizes. We note that ScoreLiNGAM is written with `C++` using the `Armadillo` linear algebra library and an `R` wrapper via the `Rcpp` package, while DirectLiNGAM is written in `Python` (`https://github.com/cdt15/lingam`) with a wrapper function in `R` using the `reticulate` package that is written by this paper's authors. HighDimLiNGAM is also written in `C++` (`https://github.com/ysamwang/highDNG`) with an `R` wrapper, but it searches regressor subsets when computing low-dimensional linear regressions–the likely reason for its slower time despite 7

parallel threads. All simulations were run on a Dell XPS 13 with Intel Core™ i7-8550U CPU @ 1.80GHz × 8, 8 GB RAM, and 64-bit Ubuntu 20.04.3 LTS OS.

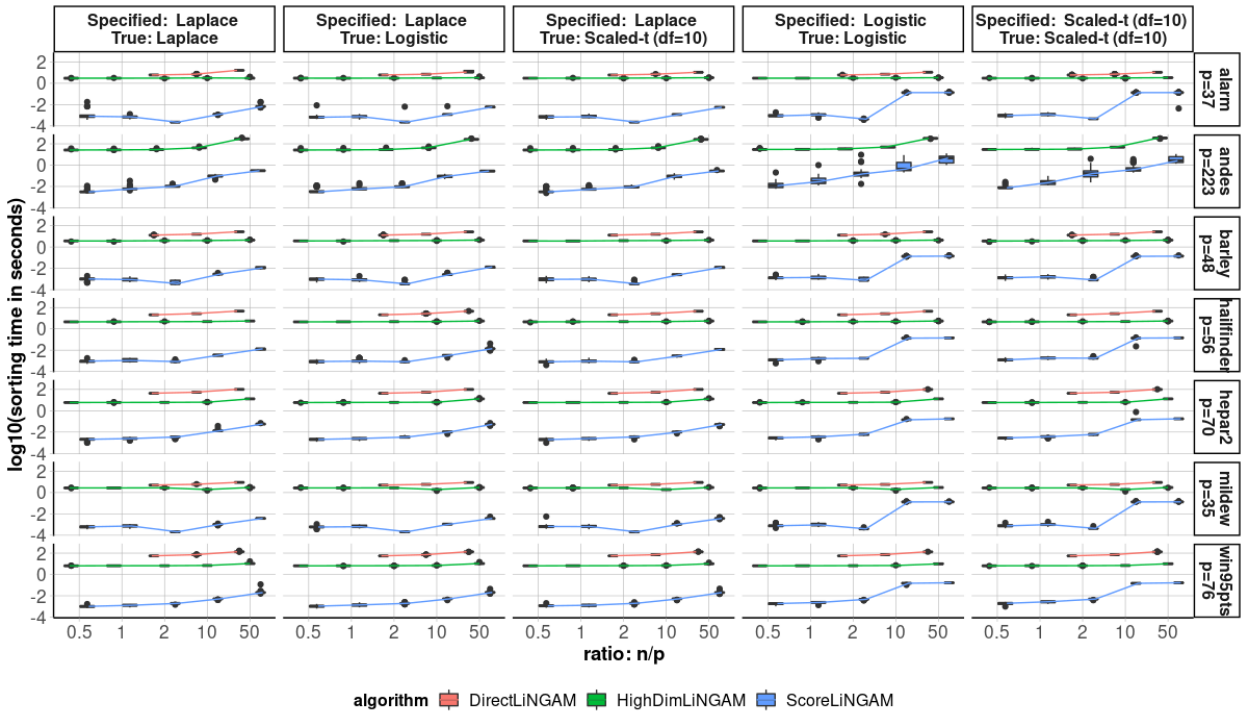

Figure 2.C.1: The simulation times for LiNGAM estimation procedures

2.C.2 Sorting Times for Large Networks

Figure 2.C.2 contains the sorting times to go along with Figure 2.3.2 in the main text.

2.D The sorting algorithm in practice

In Algorithm 2, we present further pseudo-code for ScoreLiNGAM's sorting procedure in practice, which uses partial regression.

Algorithm 2: The sorting procedure in practice

Data: $\mathbf{X} \in \mathbb{R}^{n \times p}$ (standardized), $\{\widehat{N}_k\}_{k=1}^p$

Result: $\hat{\pi}(1), \hat{\pi}(2), \ldots, \hat{\pi}(p)$

\# initialize mixing matrix

$\mathbf{M} \leftarrow \mathbb{I}_{p \times p}$

\# initialize residual matrix

$\mathbf{R} \leftarrow \mathbf{X}$

\# initialize scores

$s_k \leftarrow \mathcal{S}(k; \mathbf{R}), k = 1, 2, \ldots, p.$

\# sort the nodes

for $t = 1, 2, \ldots, p+1$ **do**

  $\hat{\pi}(t) \leftarrow \arg\max_{k \notin \mathcal{A}_t} s_k$

  \# update residuals for neighbors of selected node.

  for $k \in \widehat{N}_{\hat{\pi}(t)} \backslash \mathcal{A}_t$ **do**

    \# update residuals with partial regression.

    for $a \in \{j : \mathbf{M}_{\hat{\pi}(t)j} \neq 0, \mathbf{M}_{kj} = 0\}$ **do**

      $\mathbf{M}_{ka} \leftarrow (\mathbf{R}_{\cdot a}^T \mathbf{R}_{\cdot a})^{-1} \mathbf{R}_{\cdot a}^T \mathbf{R}_{\cdot k}$

      $\mathbf{R}_{\cdot k} \leftarrow \mathbf{R}_{\cdot k} - \mathbf{M}_{ka} \mathbf{R}_{\cdot a}$

    end

    \# update score

    $s_k \leftarrow \mathcal{S}(k; \mathbf{R})$

  end

end

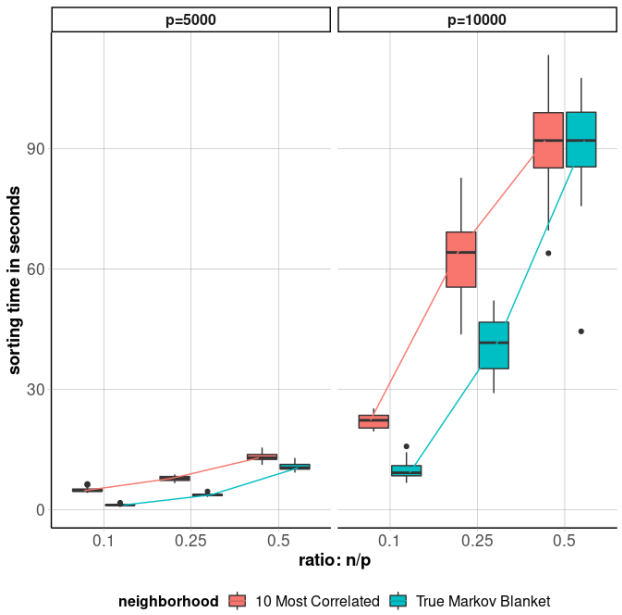

Figure 2.C.2: Sorting times for ScoreLiNGAM under $p = 5000, 10000$ and $n = 0.1p, 0.25p, 0.5p$. Color indicates how the neighborhood sets are constructed.

2.D.1 Obtaining the scale-parameter for Emprical Mean Log-likelihood in (3.1)

As discussed in the main text, our sequential algorithm at step $t \geq 1$ in practice requires the estimation of the scale parameter, η_{kt}, in Equation (3.1). Here, we discuss the estimator for the three parametric assumptions used in this chapter. We make use of the respective definitions and properties in Forbes et al. (2010).

- **Laplace Distribution:** If $\epsilon_k \sim \text{Laplace}(0, \theta_k)$, we have that $\theta_k = \mathbb{E}[|\epsilon_k|]$ is the scale parameter. When $g(\cdot; \eta_{kt})$ is specified as the density for $\text{Laplace}(0, \eta_{kt})$, the maximum likelihood estimator we use in practice is $\hat{\eta}_{kt} = \frac{1}{n} \left\| \hat{R}_{kt} \right\|_1$.
- **Logistic Distribution:** If $\epsilon_k \sim \text{Logistic}(0, \theta_k)$, then θ_k is the scale parameter. We have that $\text{Var}\left[[]\, \epsilon_k\right] = \frac{\pi^2}{3}\theta_k^2$. When $g(\cdot; \eta_{kt})$ is specified as the density for $\text{Logistic}(0, \eta_{kt})$, we find that the plug-in estimator $\hat{\eta}_{kt} = \frac{\sqrt{3}}{\pi}\hat{\sigma}_{kt}$ to work satisfactorily.

- **Scaled-t Distribution:** If $\epsilon_k \sim$ Scaled-t$(0, \nu, \theta_k)$, then we say ϵ_k is equal in distribution to the scale parameter, θ_k, times $U \sim \text{t}(0, \nu)$, a Student's t-distributed random variable having mean 0 and degrees of freedom $\nu > 0$. That is, $\epsilon_k \stackrel{d}{=} \theta_k U$. For $\nu > 2$, we have that $\text{Var}\left[[]\, \epsilon_k\right] = \theta_k^2 \left(\frac{\nu}{\nu-2}\right)$. When $g(\cdot; \eta_{kt})$ is specified as the density for Scaled-t$(0, \nu, \eta_{kt})$ with $\nu > 2$ assumed to be known, we find that the plug-in estimator $\hat{\eta}_{kt} = \hat{\sigma}_{kt}\sqrt{\frac{\nu-2}{\nu}}$ to work satisfactorily.

In equation (3.1) and in the plug-in estimators for the Logistic and Scaled-t specifications, we use

$$\hat{\sigma}_{kt}^2 = \frac{1}{n}\left\|\hat{R}_{kt}\right\|_2^2.$$

CHAPTER 3

Statistical Guarantees when Learning the Topological Ordering for the Linear Non-Gaussian Acyclic Model with Laplace Noise

3.1 Introduction

The Linear Non-Gaussian Acyclic Model (LiNGAM) of Shimizu et al. (2006) is also the focus of the present chapter. We focus on deriving statistical estimation theory for this model when the noise terms come from a Laplace distribution. Theorem 3.2.2 is our main result.

Compared to the state of the art for LiNGAM, we showed in the previous chapter that a score-based alternative allows us to identify the underlying DAG and to accurately estimate it in practice when our regularity assumptions are met. Compared to Shimizu et al. (2006), Shimizu et al. (2011), and Wang and Drton (2019) who work under a semi-parametric assumption on the LiNGAM's non-Gaussian noise terms, we provide a relatively more scale-able score-based estimation procedure under an explicit parametric assumption for the LiNGAM's noise terms. In terms of theory, Shimizu et al. (2006) and Shimizu et al. (2011) provide identifiabiliy results for the respective LiNGAM learning procedures–that is, with knowledge of the true distribution defined by the LiNGAM and an oracle for conditional independence queries in the case of latter. Meanwhile Wang and Drton (2019) provide formal statistical consistency results for their LiNGAM-learning procedure. Compared to the consistency theory of Wang and Drton (2019), which works under conditions on the higher

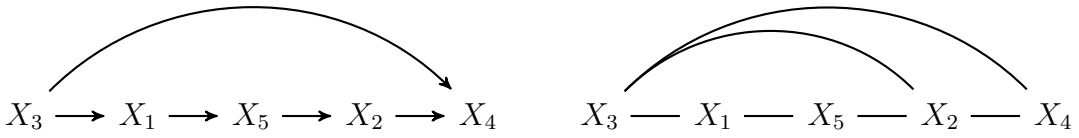

Figure 3.1.1: The target DAG (left) and the undirected graph (right) we start with. We consider the undirected graph to be known from either domain knowledge or a pre-processing step.

order moments on strategically defined least-squares residuals, we make use of sub-Exponential deviation inequalities centered on a Laplace-noise parametric assumption. The identifiability results of the previous chapter are slightly more general than a Laplace assumption: we showed how to identify a LiNGAM with noise distributions that come from any non-Gaussian scale-location family. Further, Wang and Drton (2019) discuss how one might use a priori known neighborhood estimates, e.g. the neighbors in an undirected graphical model as in Figure 3.1.1, in the LiNGAM estimation procedure, but they do not incorporate these a priori known neighborhood estimates into their consistency results; we incorporate these neighborhood sets into our theoretical discussion.

The rest of the chapter is organized as follows. In Section 3.2, we will provide the finite sample version of the algorithm for the case of Laplace errors along with a general theorem for its finite sample accuracy guarantees, a corollary for statistical consistency as the number of nodes in the underlying DAG diverges, and a corollary for finite sample accuracy when the LiNGAM is fixed. Finally, we will conclude with a summary of our findings and discussion of future work.

3.1.1 Some notation

For a positive integer m, we write [m]= {1,2,...,m}. For any set S, $|S|$ will denote its cardinality: the number of unique elements it contains. $|S| = 0$ means $S = \emptyset$, the set with no elements.

For sets $T \subseteq [m], S \subseteq [r]$, and matrix $A \in \mathbb{R}^{m \times r}$, $A_{\cdot S} \in \mathbb{R}^{m \times |S|}$ is the sub-matrix defined by indexing columns S in A, while $A_{T \cdot} \in \mathbb{R}^{|T| \times r}$ is the sub-matrix given by indexing rows T in A. Similarly, $A_{TS} \in \mathbb{R}^{|T| \times |S|}$ is the sub-matrix indexing rows T and columns S of A. Similarly, for a vector $v \in \mathbb{R}^m$, we will write $v_T \in \mathbb{R}^{|T|}$ to denote the subset of entries indexed by T. We will also sometimes write $(v_j; j \in T)$ to denote v_T.

For two sets S and T, we will make use of set operations such as their intersection: $S \cap T = \{a : a \in S \text{ and } a \in T\}$; their union: $S \cup T = \{a : a \in S \text{ or } a \in T\}$; and their difference: $S \backslash T = \{a : a \in S \text{ and } a \notin T\}$. For sets $S_1, \ldots, S_K$, we denote their intersection as $\bigcap_{j=1}^K S_j$ and their union as $\bigcup_{j=1}^K S_j$.

3.1.2 Linear Non-Gaussian Acyclic Model (LiNGAM)

We follow closely here the definition of a LiNGAM given by Shimizu et al. (2006) and in Definition 2.1.2. Moreover, we seek to estimate an ordering of nodes in the underlying LiNGAM as defined in Definition 2.1.1. As before, we will denote this estimate as $\hat{\pi}$. We will also apply Algorithm 1 at steps $t = 1, 2, \ldots, p$ until all nodes are sorted.

3.2 Finite Sample Sorting Procedure

Our main result in this section is Theorem 3.2.2, which provides a finite sample bound on the probability that $\hat{\pi}$ is accurate. Working from this result, Corollary 3.2.6 examines a condition for our sorting procedure to be statistically consistent as number of nodes diverge. On the other hand, Corollary 3.2.7 discusses the finite sample bound of Theorem 3.2.2 when the underlying LiNGAM is fixed.

Assume that we have a data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ such that $\mathbf{X}_{i\cdot}$, the i-th row, is iid across $i = 1, 2, \ldots, n$ from a distribution defined by a LiNGAM satisfying Assumptions 2.2.1 and 2.2.2. Also let Assumption 2.2.3 hold, where the $\widehat{N}_k$ are estimated with a dataset independent of $\mathbf{X}$ by an asymptotically consistent procedure, or are simply known from domain knowledge.

Analogous to Section 2.2.2, consider:

$$\hat{\beta}_{kt} = \left(\mathbf{X}^T_{\cdot \widehat{N}_{kt}} \mathbf{X}_{\cdot \widehat{N}_{kt}}\right)^{-1} \mathbf{X}^T_{\cdot \widehat{N}_{kt}} \mathbf{X}_{\cdot k} \in \mathbb{R}^{|\widehat{N}_{kt}| \times 1},$$

which exists so long as $1 \leq |\widehat{N}_{kt}| \leq n$ and $\mathbf{X}_{\cdot \widehat{N}_{kt}}$ is of full column rank almost surely. Further, we define $\hat{R}_{kt} \in \mathbb{R}^{n \times 1}$ as

$$\hat{R}_{kt} = \begin{cases} \mathbf{X}_{\cdot k} & \text{if } |\widehat{N}_{kt}| = 0 \\ \mathbf{X}_{\cdot k} - \mathbf{X}_{\cdot \widehat{N}_{kt}} \hat{\beta}_{kt} & \text{if } |\widehat{N}_{kt}| \geq 1 \end{cases},$$

the vector of residuals which we will use to estimate the pertinent parameters of (2.3). We refer the reader to Algorithm 2 in the appendix for the sorting procedure used in practice, which uses partial regression and strategic updates to the residuals to obtain a quicker procedure in practice.

3.2.1 Laplace Scale-Location Family

We now work with the Laplace scale-location family for a concrete LiNGAM model to establish finite-sample bounds. For other error distributions, similar results can be obtained by using corresponding concentration results.

Assumption 3.2.1 (Laplace errors)**.**
Let the densities $\{g(\cdot; \theta_k)\}$ correspond to

$$\epsilon_k \sim \text{Laplace}(0, \theta_k), \ k = 1, 2, \ldots, p$$

where

$$\theta_k = \mathbb{E}\left[|\epsilon_k|\right]$$

is the corresponding scale parameter, and

$$g(e; \theta_k) = \frac{1}{2\theta_k} \exp\left(-\frac{|e|}{\theta_k}\right).$$

Correspondingly, we estimate the scale parameter for $g(\cdot;\eta_{kt})$ of (2.3) using the plug-in estimate:

$$\hat{\eta}_{kt} = \frac{1}{n}\sum_{i=1}^{n} |\hat{R}_{i,kt}| = \frac{1}{n}\left\|\hat{R}_{kt}\right\|_1,$$

where $\hat{R}_{i,kt}$ is the i-th entry of $\hat{R}_{kt}$. We also estimate $\mathbf{V}[R_{kt}]$ as:

$$\hat{\sigma}^2_{kt} = \frac{1}{n}\sum_{i=1}^{n} \hat{R}^2_{i,kt} = \frac{1}{n}\left\|\hat{R}_{kt}\right\|_2^2.$$

As the sample analogue to (2.3), we select the next node to continue the ordering with respect to the sample mean log-likelihood ratio:

$$\arg\max_{k\notin\mathcal{A}_t} \frac{1}{n}\sum_{i=1}^{n} \log\left\{\frac{g(\hat{R}_{i,kt};\hat{\eta}_{kt})}{\phi(\hat{R}_{i,kt};\hat{\sigma}^2_{kt})}\right\} = \arg\max_{k\notin\mathcal{A}_t} \log\left\{\frac{\hat{\sigma}_{kt}}{\hat{\eta}_{kt}}\right\}. \tag{3.1}$$

Under the Laplace family assumption for the densities $\{g(\cdot;\theta_k)\}$, (2.3) is also the same as:

$$\arg\max_{k\notin\mathcal{A}_t} \log\left\{\frac{\sigma_{kt}}{\eta_{kt}}\right\}.$$

Intuitively, if

$$\hat{\eta}_{kt} \xrightarrow{p} \eta_{kt} \text{ and } \hat{\sigma}_{kt} \xrightarrow{p} \sigma_{kt}, \tag{3.2}$$

then it seems that the choice in (3.1) will be correct as $n \to \infty$. The remainder of this section is devoted to presenting results which tell us under what conditions on sample size, dimension p, and maximum neighborhood size we will have that using (3.1) in Algorithm 1 recovers a valid ordering satisfying Definition 2.1.1. In particular, we make use of sub-Exponential deviation inequalities.

3.2.2 Finite Sample Accuracy Based on Deviation Inequalities

For the statement of Theorem 3.2.2 and Assumptions 3.2.3 and 3.2.4 below, denote:

- The maximum cardinality of the estimated neighborhood sets:

$$d = \max_{j\in[p]} \left|\widehat{N}_j\right|.$$

- The set of permutations π_0 satisfying Definition 2.1.1 with respect to the DAG $\mathcal{G}$ underlying our LiNGAM of interest:

$$\Pi_{\mathcal{G}},$$

and the number of such permutations, i.e. the cardinalty of the set $\Pi_{\mathcal{G}}$:

$$|\Pi_{\mathcal{G}}|.$$

- The partial ordering at step $t \in [p]$ given by $\pi_0 \in \Pi_{\mathcal{G}}$:

$$\mathcal{A}_t^{\pi_0} = \begin{cases} \emptyset & t = 1 \\ \{\pi_0(j)\}_{j=1}^{t-1} & 2 \leq t \leq p \end{cases}.$$

- Node k's neighboring nodes that are sorted at step t with partial ordering according to $\pi_0 \in \Pi_{\mathcal{G}}$:

$$\widehat{N}_{kt}^{\pi_0} = \widehat{N}_k \cap \mathcal{A}_t^{\pi_0}.$$

- Node k's population-level residual from the linear regression of X_k on $X_{\widehat{N}_{kt}^{\pi_0}}$:

$$R_{kt}^{\pi_0} = X_k - \beta_{kt}^{\pi_0 T} X_{\widehat{N}_{kt}^{\pi_0}};\ \beta_{kt}^{\pi_0} = \arg \min_{\theta \in \mathbb{R}^{|\widehat{N}_{kt}^{\pi_0}|}} \mathbb{E}\left[\left(X_k - \theta^T X_{\widehat{N}_{kt}^{\pi_0}}\right)^2\right].$$

- The maximum possible sub-Exponential norm of a population-level residual:

$$\gamma_{max} = \max_{\pi_0 \in \Pi_{\mathcal{G}}, t \in [p], k \notin \mathcal{A}_t^{\pi_0}} \left\| R_{kt}^{\pi_0} \right\|_{\psi_1},$$

where $\|V\|_{\psi_1} \doteq \inf\{t > 0 : \mathbb{E}\left[\exp\left(|V|/t\right)\right] \leq 2\}$ is the sub-Exponential norm for real-valued random variable V.

- The target parameters when the partial ordering is given by $\pi_0 \in \Pi_{\mathcal{G}}$:

$$\eta_{kt}(\pi_0) \doteq \mathbb{E}[|R_{kt}^{\pi_0}|] \text{ and } \sigma_{kt}(\pi_0) \doteq (\mathbb{E}[|R_{kt}^{\pi_0}|^2])^{1/2},$$

along with

$$\eta_{max} \doteq \max_{\pi_0 \in \Pi_{\mathcal{G}}, t \in [p], k \notin \mathcal{A}_t^{\pi_0}} \eta_{kt}(\pi_0).$$

- The sets of valid and invalid nodes, respectively, to continue the partial ordering given by $\mathcal{A}_t^{\pi_0}$:

$$V(t;\pi_0), I(t;\pi_0) \subseteq [p] \backslash \mathcal{A}_t^{\pi_0}.$$

- The minimum difference in population-level scores between valid nodes $\ell \in V(t;\pi_0)$ and invalid nodes $k \in I(t;\pi_0)$ when the partial ordering is given by $\pi_0 \in \Pi_{\mathcal{G}}$ at step $t \in [p]$:

$$\delta_t^{\pi_0} = \min_{\ell \in V(t;\pi_0)} \log\left\{\frac{\sigma_{\ell t}(\pi_0)}{\eta_{\ell t}(\pi_0)}\right\} - \max_{k \in I(t;\pi_0)} \log\left\{\frac{\sigma_{kt}(\pi_0)}{\eta_{kt}(\pi_0)}\right\}.$$

Our main result is as follows.

Theorem 3.2.2 (Finite Sample Sorting Procedure Accuracy)**.**
Let $\hat{\pi} = (\hat{\pi}(1), \ldots, \hat{\pi}(p))$ *be constructed using*

$$\mathcal{S}(k, \mathcal{A}_t; \mathbf{X}) = \log\left\{\frac{\hat{\sigma}_{kt}}{\hat{\eta}_{kt}}\right\}$$

in Algorithm 1 across steps $t = 1, 2, \ldots, p$*. Then*

$$\Pr(\hat{\pi} \in \Pi_{\mathcal{G}}) \geq 1 - \frac{8|\Pi_{\mathcal{G}}|p^2}{n^{\xi}}, \tag{3.3}$$

so long as $d \leq n$ *and Assumptions 2.2.3, 3.2.1, 3.2.3, and 3.2.4 hold. Here,* $\xi > 0$ *is the constant in Assumption 3.2.3.*

We now discuss the assumptions on Theorem 3.2.2 and the lemmas where they are used. The formal proofs of Theorem 3.2.2 and all its pertinent lemmas are contained in Appendix 3.A.

Assumption 3.2.3 (Gap Condition)**.**
n is large enough so that for each $\pi_0 \in \Pi_{\mathcal{G}}$, $t \in [p]$, and each $k \in \mathcal{A}_t^{\pi_0}$ the inequalities:

$$\left(\frac{\delta_t^{\pi_0}}{\delta_t^{\pi_0} + 4}\right)\eta_{kt}(\pi_0) > \frac{2d(\gamma_{max} + \eta_{max})(1+\xi)\log(n)}{c^{1/2}n^{1/2}} \asymp \frac{d(\gamma_{max} + \eta_{max})\log(n)}{n^{1/2}} \tag{3.4}$$

and

$$\begin{aligned}\left(\frac{\delta_t^{\pi_0}}{\delta_t^{\pi_0} + 4}\right)\sigma_{kt}(\pi_0) \quad &> \frac{(1+\xi)\gamma_{max}\log^{3/2}(2n)\sqrt{32}}{cn^{1/2}} + \frac{d\gamma_{max}^2(1+\xi)\log^2(n)}{n^{1/2}c} \\ &+ \frac{d\gamma_{max}^2(1+\xi)^2\log^2(n)}{nc^2} \asymp \frac{d\gamma_{max}^2\log^2(n)}{n^{1/2}}\end{aligned} \tag{3.5}$$

hold, where c is an absolute constant.

Assumption 3.2.4 (Conditionally Zero-centered Residuals)**.**
For all $\pi_0 \in \Pi_{\mathcal{G}}$, each $t \in [p]$, and each $k \notin \mathcal{A}_t^{\pi_0}$,

$$\mathbb{E}\left[R_{kt}^{\pi_0} \middle| X_{\widehat{N}_{kt}^{\pi_0}}\right] = 0$$

when $\widehat{N}_{kt}^{\pi_0}$ is non-empty. If $\widehat{N}_{kt} = \emptyset$, we take $R_{kt}^{\pi_0} = X_k$, which is marginally zero-centered without loss of generality.

The term $|\Pi_{\mathcal{G}}|p^2$ in (3.3) comes about due to a union bound across events related to deviations in our finite sample score from the population-level score. Firstly, for $\hat{\pi}$ to be correct at any given step, it must be that $\hat{\pi}$ is the same as some $\pi_0 \in \Pi_{\mathcal{G}}$. We do not quite know which π_0 will satisfy this, so the total number of orderings contributes to the union bound. The p^2 term corresponds to the total number of steps required to sort all nodes (p) and the maximum number of unordered nodes at any given step (also p). We refer the reader to Appendix 3.A.1 for details on this union bound. Note that the right hand side of (3.3) goes to 1 so long as $|\Pi_{\mathcal{G}}|p^2/n^{\xi} \to 0$, which means that p and $|\Pi_{\mathcal{G}}|$ must both grow no faster than a polynomial in n, when $\mathcal{G}$ is not fixed (see Corollary 3.2.6). In practice, this latter scenario may come about if we are willing to add more nodes to a specified LiNGAM model as more data becomes available. But if the number of nodes stays fixed, then the requirement that $|\Pi_{\mathcal{G}}|p^2/n^{\xi} \to 0$ will be satisfied with any $\xi > 0$.

The requirement that the noise terms be Laplace distributed (Assumption 3.2.1) can be changed to any other family satisfying Assumptions 2.2.1 and 2.2.2 (both needed for identifying a valid topological ordering), but some of the arguments in Appendix 3.A will need to be changed as they make use of the sub-Exponential deviation inequalities found in Wainwright (2019).

The requirement that $d \leq n$ corresponds to our use of low-dimensional linear least squares regressions to obtain the residuals $\hat{R}_{kt}$. In practice, these regressions save quite a bit of time as they do not require tuning a penalty parameter, e.g. for a LASSO term, for each linear regression during the sorting procedure.

The right hand sides of the two inequalities in Assumption 3.2.3, which depend on d, γ_{max}, η_{max}, and n, provide the overall rate of our convergence. Theoretically, d can grow with n. If d grows with n, then γ_{max} and η_{max} may also grow (if not constants). The contribution of d to the rate shows the penalty we pay in terms of the accuracy of our estimated ordering if our neighborhood estimates are too large, while the contribution of γ_{max} shows the potential bottleneck in our accuracy if our variables are too noisy as determined the sub-Exponential norm, which measures the thickness of a distribution's tail. A similar interpretation in terms of noisy residuals holds for η_{max}. This rate is determined by the deviation bound on $\sigma_{kt}(\pi_0) - \hat{\sigma}_{kt}(\pi_0)$ in Lemma 3.A.13 of Appendix 3.A, and by the analogous bound on $\eta_{kt}(\pi_0) - \hat{\eta}_{kt}(\pi_0)$ in Lemma 3.A.9. Here, $\hat{\eta}_{kt}(\pi_0)$ and $\hat{\sigma}_{kt}(\pi_0)$ are the corresponding sample estimates when the partial ordering is given by $\pi_0 \in \Pi_{\mathcal{G}}$.

The requirement in Assumption 3.2.3 that n be large enough so that inequalities hold with respect to the target parameters $(\eta_{kt}(\pi_0), \sigma_{kt}(\pi_0))$ and $\delta_t^{\pi_0}$ is a seemingly standard requirement. In these inequalities, which correspond to the rate of convergence of our sorting procedure, notice also the appearance of ξ: the larger it is required to be, the slower the overall rate will be. These inequalities come about due to a use of the Mean Value Theorem in Lemma 3.A.2 of Appendix 3.A, which shows how to arrive at deviation bounds on $\log(\eta_{kt}(\pi_0)) - \log(\hat{\eta}_{kt}(\pi_0))$ and $\log(\sigma_{kt}(\pi_0)) - \log(\hat{\sigma}_{kt}(\pi_0))$ from deviation bounds on $\eta_{kt}(\pi_0) - \hat{\eta}_{kt}(\pi_0)$ and $\sigma_{kt}(\pi_0) - \hat{\sigma}_{kt}(\pi_0)$, respectively. Through a use of the triangle inequality, the former help bound our primary focus:

$$\log(\sigma_{kt}(\pi_0)/\eta_{kt}(\pi_0)) - \log(\hat{\sigma}_{kt}(\pi_0)/\hat{\eta}_{kt}(\pi_0)),$$

the difference between the population-level score of interest and its sample analogue.

Finally, we note that the requirement that Assumption 3.2.4 is similar to what is often taken for granted in regression analysis. It is different from the statement $\mathbb{E}\left[\hat{R}_{kt}^{\pi_0} \middle| \mathbf{X}_{\cdot \widehat{N}_{kt}^{\pi_0}}\right] = \mathbf{0}_{n\times 1}$, which is a consequence of the former and noting that $\hat{R}_{kt}^{\pi_0}$ is a projection of $\mathbf{X}_{\cdot k}$ onto the orthogonal complement of $\mathbf{X}_{\cdot \widehat{N}_{kt}^{\pi_0}}$'s column space. The use of $\mathbb{E}\left[R_{kt}^{\pi_0} \middle| X_{\widehat{N}_{kt}^{\pi_0}}\right] = 0$ corresponds to an application of Bernstein's inequality in Appendix 3.A's Lemma 3.A.7 and Lemma

3.*A*.11, two results that examine the deviation,

$$\mathbf{X}_{\cdot\widehat{N}_{kt}}\beta_{kt} - \mathbf{X}_{\cdot\widehat{N}_{kt}}\hat{\beta}_{kt}$$

in terms of the ℓ_1 and ℓ_2 vector norms, respectively. Lemma 3.A.7 helps with Lemma 3.A.9 while Lemma 3.A.11 helps with Lemma 3.A.13, the important results mentioned earlier that bound the difference between $\hat{\eta}_{kt} - \eta_{kt}$ and $\hat{\sigma}_{kt} - \sigma_{kt}$, respectively.

When $\widehat{N}_{kt}^{\pi_0} = \mathcal{A}_t^{\pi_0}$, it is readily checked that $\mathbb{E}\left[R_{kt}^{\pi_0}\middle|X_{\widehat{N}_{kt}^{\pi_0}}\right] = 0$. This is because the sub-mixing matrix, $\mathbf{M}_{\widehat{N}_{kt}^{\pi_0}\widehat{N}_{kt}^{\pi_0}}$, is of full column and row rank, so we can write:

$$X_k = \sum_{j\in[p]}\mathbf{M}_{kj}\epsilon_j = \mathbf{M}_{k\widehat{N}_{kt}^{\pi_0}}\mathbf{M}^{-1}_{\widehat{N}_{kt}^{\pi_0}\widehat{N}_{kt}^{\pi_0}}X_{\widehat{N}_{kt}^{\pi_0}} + \sum_{j\notin\widehat{N}_{kt}^{\pi_0}}\mathbf{M}_{kj}\epsilon_j, \tag{3.6}$$

a linear combination of entries in the zero-centered ϵ or, equivalently, a linear combination of $X_{\widehat{N}_{kt}^{\pi_0}} = \mathbf{M}_{\widehat{N}_{kt}^{\pi_0}\widehat{N}_{kt}^{\pi_0}}\epsilon_{\widehat{N}_{kt}^{\pi_0}}$ and the independent $(\epsilon_j; j \notin \widehat{N}_{kt}^{\pi_0})$. For any other choice of $\widehat{N}_{kt}^{\pi_0}$, the satisfaction of $\mathbb{E}\left[R_{kt}^{\pi_0}\middle|X_{\widehat{N}_{kt}^{\pi_0}}\right] = 0$ may be on a case by case basis according the underlying DAG's structure and/or the structure of the sets $\widehat{N}_k$. For more discussion on the case that $\widehat{N}_{kt}^{\pi_0} \neq \mathcal{A}_t^{\pi_0}$ and the satisfaction of this condition, we refer the reader to Appendix 3.B for a theoretical discussion on how to append other ordered nodes to the set $\widehat{N}_{kt}^{\pi_0}$ to guarantee this assumption. The gist of Appendix 3.B is that, where $L_{kt}^{\pi_0} = \bigcup_{j\in\widehat{N}_{kt}^{\pi_0}}\{j\} \cup AN_j$, we would like to guarantee that each directed path in the underlying DAG from $(X_a; a \in L_{kt}^{\pi_0}\backslash\widehat{N}_{kt}^{\pi_0})$ to X_k must be mediated by $(X_b; b \in \widehat{N}_{kt}^{\pi_0})$.

For the case that we do not modify the regression sets $\widehat{N}_{kt}^{\pi_0}$ as discussed in Appendix 3.B, consider relaxing the assumption that $\mathbb{E}[R_{kt}^{\pi_0}|X_{\widehat{N}_{kt}^{\pi_0}}] = 0$ in Assumption 3.2.4 to what is written in Assumption 3.2.5.

Assumption 3.2.5 (Loosening our Residual Assumption)**.**
Denote $P_{kt}^{\pi_0}$ as the projection matrix onto the column space of $\mathbf{X}_{\widehat{N}_{kt}^{\pi_0}}$. We require the following to hold, uniformly across $\pi_0 \in \Pi_{\mathcal{G}}$, $t \in [p]$, and $k \notin \{\pi_0(j)\}_{j=1}^{t-1}$.

1. There exists $\tau_{max} \geq 0$ such that:

$$\left\|\mathbb{E}\left[\mathbf{X}_{ik}\middle|\mathbf{X}_{i\widehat{N}_{kt}^{\pi_0}}\right] - \mathbf{X}_{i\widehat{N}_{kt}^{\pi_0}}\beta_{kt}^{\pi_0}\right\|_{\psi_1} = \left\|\mathbb{E}\left[R_{kt}^{\pi_0}\middle|X_{\widehat{N}_{kt}^{\pi_0}}\right]\right\|_{\psi_1} \leq \tau_{max},$$

2. There exists a non-negative sequence $\rho_{n,\xi}$ and $\xi > 0$ such that:

$$\frac{1}{n}\left\| P_{kt}\mathbb{E}\left[\mathbf{X}_{\cdot k}\middle|\mathbf{X}_{\cdot \widehat{N}_{kt}^{\pi_0}}\right] - \mathbf{X}_{\cdot \widehat{N}_{kt}^{\pi_0}}\beta_{kt}^{\pi_0}\right\|_1 \leq \rho_{n,\xi}$$

and

$$\frac{1}{n}\left\| P_{kt}\mathbb{E}\left[\mathbf{X}_{\cdot k}\middle|\mathbf{X}_{\cdot \widehat{N}_{kt}^{\pi_0}}\right] - \mathbf{X}_{\cdot \widehat{N}_{kt}^{\pi_0}}\beta_{kt}^{\pi_0}\right\|_2^2 \leq \rho_{n,\xi}$$

with probability at least $1 - \frac{1}{n^{\xi}}$.

As may be expected by relaxing conditions to what is given in Assumption 3.2.5, we will have a slower rate of consistency for our sorting procedure as dictated by the magnitude of τ_{max} and $\rho_{n,\xi}$. We refer the interested reader to Equations (3.13) and (3.14) in Appendix 3.A.2 for this insight. Also notice that $\mathbb{E}[R_{kt}^{\pi_0}|X_{\widehat{N}_{kt}^{\pi_0}}] = 0$ if and only if $\mathbb{E}[X_k|X_{\widehat{N}_{kt}^{\pi_0}}] = (\beta_{kt}^{\pi_0})^T X_{\widehat{N}_{kt}^{\pi_0}}$, which would imply that $\rho_{n,\xi} = 0$ and $\tau_{max} = 0$ work and we'd end up with the rate implied by (3.4) and (3.5) as a special case. After replacing (3.4) and (3.5) in Assumption 3.2.3 with (3.13) and (3.14), the statement in Theorem 3.2.2 would remain the same when we generalize Assumption 3.2.4 to Assumption 3.2.5.

3.2.3 Corollaries to Theorem 3.2.2

Given the discussion following Theorem 3.2.2 in the previous subsection, we can make the following formal corollaries.

Corollary 3.2.6 (Finite Sample Sorting Procedure is Consistent when $p \to \infty$)**.**
Let the conditions of Theoreom 3.2 hold. Assume further that p and $|\Pi_{\mathcal{G}}|$ grow, possibly to infinity, at a rate no faster than a polynomial in n. Then we have that:

$$\Pr(\hat{\pi} \in \Pi_{\mathcal{G}}) \to 1.$$

Proof.

Our condition on the growth of p and $|\Pi_{\mathcal{G}}|$ means that there exists $\xi_1, \xi_2 > 0$ such that $p = o(n^{\xi_1})$ and $|\Pi_{\mathcal{G}}| = o(n^{\xi_2})$. It follows that $\frac{p^2|\Pi_{\mathcal{G}}|}{n^{2\xi_1\xi_2}} \to 0$. Next, pick any $\xi > 0$ satisfying

Assumption 3.2.3 such that $\xi \geq 2\xi_1\xi_2$. With such a choice of ξ, we have also $\frac{p^2|\Pi_{\mathcal{G}}|}{n^{\xi}} \to 0$. This means the right hand side of (3.3) goes to 1 and therefore $\Pr(\hat{\pi} \in \Pi_{\mathcal{G}}) \to 1$ as we wanted. □

Note that Corollary 3.2.6 does not require $p \to \infty$, necessarily. If the underlying LiNGAM is fixed, then we can say (trivially) that p and $\Pi_{\mathcal{G}}$ grow as a polynomial in n of degree zero. Moreover, the growth of p and $\mathcal{G}$ can be of any polynomial in n, no matter how large, for us to achieve convergence in probability of our sorting procedure.

Corollary 3.2.7 (Finite sample accuracy guarantee for fixed LiNGAM)**.**
If the underlying LiNGAM is fixed, and that $\widehat{N}_k = [p]\backslash\{k\}$ for each $k \in [p]$, then

$$\Pr(\hat{\pi} \in \Pi_{\mathcal{G}}) \gtrsim 1 - \frac{1}{n^{\xi}}$$

for any $\xi > 0$ and large enough n.

Proof.
Note that Theorem 2.2.5 on the identifiability of a topological ordering guarantees that $\delta_t^{\pi_0} > 0$, while $\eta_{kt}(\pi_0) > 0$ holds due to our assumption of a continuous distribution for the noise terms in our LiNGAM (Assumption 3.2.1). Overall the left hand sides of (3.4) and (3.5) in Assumption 3.2.3 will be positive. Moreover, the terms d, γ_{max}, and η_{max} on the right hand side of (3.4) and (3.5) in Assumption 3.2.3 will remain fixed, due to our assumption that the LiNGAM is fixed. It follows that for any $\xi > 0$, a large enough n will satisfy Assumption 3.2.3. Assumption 3.2.4 is satisfied because $\widehat{N}_k = [p]\backslash\{k\}$ means that $\widehat{N}_{kt}^{\pi_0} = \mathcal{A}_t^{\pi_0}$; we simply invoke the discussion surrounding (3.6) for this case. Moreover, $|\Pi_{\mathcal{G}}|p^2$ is also a constant, due to our assumption of a fixed LiNGAM. Thus, we make our desired claim that $\hat{\pi} \in \Pi_{\mathcal{G}}$ with probability at least $1 - \frac{1}{n^{\xi}}$, up to a constant factor. □

Corollary 3.2.7 is of potential interest for the following reason related to a statistical power analysis. Suppose we would like to be at least 95% confident that our topological ordering is accurate. This requires us to find a minimal $\xi > 0$ and minimal n such that

$\Pr(\hat{\pi} \in \Pi_{\mathcal{G}}) \geq 1 - \frac{8|\Pi_{\mathcal{G}}|p^2}{n^{\xi}} \approx 0.95$. In turn, this requires us to specify plausible values of $\delta_t^{\pi_0}, \gamma_{max}$, and η_{max} in Assumption 3.2.3 and plausible values of $|\Pi_{\mathcal{G}}|p^2$ in a manner similar to specifying plausible effect sizes in the power analysis for a 2-sample t-test. As an example, this all seems quite feasible when $\mathcal{G}$ is a chain graph for which only one topological ordering exists. In this case, feasible values γ_{max} and η_{max} may be upper bounded by the sub-exponential norm estimate and L_1 norm estimate, respectively, of X_k for each $k \in [p]$. Moreover, feasible $\delta_t^{\pi_0}$ values may be found by estimating LiNGAMs for several random orderings and computing the gap between the scores in (3.1). We leave the further study of this idea to future work.

3.3 Discussion

In this chapter, we discussed a estimation theory for a score-based alternative to the state of the art for the Linear Non-Gaussian Acyclic Model (LiNGAM) of Shimizu et. al (2006). Under the belief that a data mining procedure, such as those of causal discovery, cannot be useful in practice without good theoretical foundations for some underlying assumed model, our contributions are consistency and finite sample results, including for the case that $p > n$.

As a topic for future work, it would be interesting to study whether identifiability of a LiNGAM's topological ordering with the score $\log(\sigma_{kt}/\eta_{kt})$ can hold with respect to any non-Gaussian sub-Exponential noise distribution, not just the Laplace distribution, so that the argument for the proof of Theorem 3.2.2 can remain essentially unchanged. Generalizing the identification of LiNGAMs with the score $\log(\sigma_{kt}/\eta_{kt})$ is a similar, yet different claim to that of Theorem 2.2.5 which holds when the noise term densities come from an arbitrary non-Gaussian scale-location family. Further, given the simplicity of the score $\log(\sigma_{kt}/\eta_{kt})$, which takes the ratio of the L_1 and L_2 norms of R_{kt}'s distribution, it would be interesting to examine in future work whether identification of a valid topological ordering can also hold for population-level residuals arising from an interesting class of non-linear regressions. Should this be the case, Theorem 3.2.2's results will need to be modified accordingly at Appendix

3.A's Lemma 3.A.7 and Lemma 3.A.11, which deal with the deviation between the estimated conditional expectation and the true conditional expectation as functions of the random regressors.

APPENDIX

3.A Finite Sample Sorting Procedure Accuracy Lemmas and Proofs

3.A.1 Proof of Theorem 3.2.2

Recalling Assumption 2.2.3, we will condition our inference on the event:

$$\mathcal{B} = \bigcap_{k \in [p]} \{\widehat{N}_k \supseteq MB_k\}, \tag{3.7}$$

which says that all neighborhood estimates, $\widehat{N}_k$, contain the true Markov blanket, MB_k. Note that we may trivially take $\widehat{N}_k = [p] \backslash \{k\}$ for each k so that $\mathcal{B}$ holds true. In general, the sets $\widehat{N}_{kt}$ can be random, but they are independent of the data $\mathbf{X}$ used to obtain $\hat{\pi}$.

For each $t = 1, 2, \ldots, p+1$ in the sequential node ordering procedure, we are interested in the event:

$$\mathcal{E}_t \doteq \begin{cases} \emptyset & \text{if } t = 1 \\ \bigcap_{j=1}^{t-1} \{PA\left(\hat{\pi}(j)\right) \subseteq \mathcal{A}_j\} & \text{if } 2 \leq t \leq p+1 \end{cases}.$$

$\mathcal{E}_t$ states that the partial ordering to this point is correct in the sense that the parents of each node $\hat{\pi}(j) \in \mathcal{A}_t$ were already contained in the ordering before $\hat{\pi}(j)$ was appended.

Let $\Pi_{\mathcal{G}}$ be the set of permutations π satisfying Definition 2.1.1 for the DAG $\mathcal{G}$ underlying our LiNGAM of interest. The event $\mathcal{E}_t$ $(t \geq 2)$ equates to the event:

$$\text{There exists } \pi_0 \in \Pi_{\mathcal{G}} \text{ such that } \hat{\pi}(j) = \pi_0(j) \text{ for each } j = 1, \ldots, t-1. \tag{3.8}$$

Let us use the notation $(\eta_{kj}(\pi), \sigma_{kj}(\pi))$ and $(\hat{\eta}_{kj}(\pi), \hat{\sigma}_{kj}(\pi))$ to denote the target parameters and the corresponding sample estimates that we obtain by regressing node k at step j onto the set of nodes

$$\widehat{N}_{kj}^{\pi_0} \doteq \{\pi(i)\}_{i=1}^{j-1} \cap \widehat{N}_k,$$

where π is some permutation of interest. The corresponding population-level least squares residual will be denoted as:

$$R_{kj}^{\pi_0} \doteq \begin{cases} X_k & \text{if } \widehat{N}_{kj}^{\pi_0} = \emptyset \\ X_k - \left(\left(\mathbb{E}\left[X_{\widehat{N}_{kj}^{\pi_0}} X_{\widehat{N}_{kj}^{\pi_0}}^T \right] \right)^{-1} \mathbb{E}\left[X_{\widehat{N}_{kj}^{\pi_0}} X_k \right] \right)^T X_{\widehat{N}_{kj}^{\pi_0}} & \text{otherwise.} \end{cases}$$

Now define the event:

$$\mathcal{Q}_j(\ell, k; \pi) \doteq \left\{ \log\left\{ \frac{\hat{\sigma}_{\ell j}(\pi)}{\hat{\eta}_{\ell j}(\pi)} \right\} > \log\left\{ \frac{\hat{\sigma}_{kj}(\pi)}{\hat{\eta}_{kj}(\pi)} \right\} \right\},$$

which says that node ℓ has a higher finite sample score than node k in step j when the partial ordering is $\{\pi(i)\}_{i=1}^{j-1}$. Also consider the sets of nodes:

$$V(j; \pi_0) = \{\ell \notin \mathcal{A}_j : PA(\ell) \subseteq \{\pi(i)\}_{i=1}^{j-1}\} \text{ and } I(j; \pi_0) = \{k \notin \mathcal{A}_j : PA_k \backslash \{\pi(i)\}_{i=1}^{j-1} \neq \emptyset\},$$

which are the set of valid and invalid nodes to continue the partial ordering defined by permutation π at step j.

We now make use of an implication given by Theorem 2.2.5 which says:

$$\arg \max_{k \notin \{\pi_0(i)\}_{i=1}^{j-1}} \log\left\{ \frac{\sigma_{kj}(\pi_0)}{\eta_{kj}(\pi_0)} \right\} = V(j; \pi_0).$$

That is, all nodes in $V(j; \pi_0)$ will give the largest mean log-likelihood ratio at the population level, because the partial ordering $\{\pi_0(i)\}_{i=1}^{j-1}$ is correct for $\pi_0 \in \Pi_{\mathcal{G}}$. Define the gap at step j between the population-level score of interest for any node $\ell \in V(j; \pi_0)$ and the maximum population-level score among nodes $k \in I(j; \pi_0)$ as:

$$\delta_j^{\pi_0} = \log\left\{ \frac{\sigma_{\ell j}(\pi_0)}{\eta_{\ell j}(\pi_0)} \right\} - \max_{k \in I(j;\pi_0)} \log\left\{ \frac{\sigma_{kj}(\pi_0)}{\eta_{kj}(\pi_0)} \right\}.$$

Making use of this gap, which Theorem 2.2.5 gaurantees will be strictly positive, consider the events:

$$\mathcal{F}_j(k, \sigma; \pi_0) = \left\{ |\log(\sigma_{kj}(\pi_0)) - \log(\hat{\sigma}_{kj}(\pi_0))| < \frac{\delta_j^{\pi_0}}{4} \right\}$$

and

$$\mathcal{F}_j(k,\eta;\pi_0) \;=\; \left\{|\log(\eta_{kj}(\pi_0)) - \log(\hat{\eta}_{kj}(\pi_0))| < \frac{\delta_j^{\pi_0}}{4}\right\}.$$

By the triangle inequality, we have that:

$$\begin{aligned}\mathcal{F}_j(k,\sigma;\pi_0) \cap \mathcal{F}_j(k,\eta;\pi_0) \;&\subseteq\; \mathcal{H}_j(k;\pi_0) \\ &\doteq\; \left\{|\log(\sigma_{kj}(\pi_0)/\eta_{kj}(\pi_0)) - \log(\hat{\sigma}_{kj}(\pi_0)/\hat{\eta}_{kj}(\pi_0))| < \frac{\delta_j^{\pi_0}}{2}\right\}.\end{aligned}$$

Importantly, should the right hand side event, $\mathcal{H}_j(k;\pi_0)$, occur for each $k \notin \{\pi_0(i)\}_{i=1}^{j-1}$, the finite sample version of our sorting procedure will make the correct choice at step j when $\{\pi_0(i)\}_{i=1}^{j-1}$ is the partial ordering. That is,

$$\bigcap_{k\in V(j;\pi_0)\cup I(j;\pi_0)} \mathcal{F}_j(k,\sigma;\pi_0)\cap\mathcal{F}_j(k,\eta;\pi_0) \subseteq \bigcap_{k\in V(j;\pi_0)\cup I(j;\pi_0)} \mathcal{H}_j(k;\pi_0) \subseteq \bigcup_{\ell\in V(j;\pi_0)} \bigcap_{k\in I(j;\pi_0)} \mathcal{Q}_j(\ell,k;\pi_0). \tag{3.9}$$

Taking it a step further, we have the following containment statements:

$$\begin{aligned}\bigcap_{j=1}^{t-1} \bigcap_{\pi_0\in\Pi_\mathcal{G}} \bigcap_{k\in V(j;\pi_0)\cup I(j;\pi_0)} \mathcal{F}_j(k,\sigma;\pi_0) \cap \mathcal{F}_j(k,\eta;\pi_0) \;&\subseteq \bigcap_{j=1}^{t-1} \left[\bigcap_{\pi_0\in\Pi_\mathcal{G}} \bigcup_{\ell\in V(j;\pi_0)} \bigcap_{k\in I(j;\pi_0)} \mathcal{Q}_j(\ell,k;\pi_0)\right] \\ &\subseteq \mathcal{E}_t.\end{aligned} \tag{3.10}$$

The first containment follows from (3.9), after taking the intersection across $\pi_0 \in \Pi_\mathcal{G}$ and across $j = 1, \ldots, t-1$. The second containment holds by induction. At $t = 1$, the partial ordering is empty and the event in square brackets guarantees that $\hat{\pi}(1)$ is equal to the first element of some valid topological ordering. Next, suppose for induction that $\hat{\pi}(i) = \pi_0'(i)$ across $i = 1, 2, \ldots, j-2$ for some $\pi_0' \in \Pi_\mathcal{G}$. The fact that $Q_j(\ell, k; \pi_0)$ holds for all $\pi_0 \in \Pi_\mathcal{G}$ and all valid and invalid nodes ℓ and k, respectively, guarantees that $\hat{\pi}(j-1)$ will be valid, and therefore our overall topological ordering at step j will also be given by some valid topological ordering in $\Pi_\mathcal{G}$.

From (3.10), it follows that:

$$\Pr\left(\bigcap_{j=1}^{t-1}\bigcap_{\pi_0\in\Pi_{\mathcal{G}}}\bigcap_{k\in V(j;\pi_0)\cup I(j;\pi_0)}\mathcal{F}_j(k,\sigma;\pi_0)\cap\mathcal{F}_j(k,\eta;\pi_0)\middle|\mathcal{B}\right)\leq\Pr\left(\mathcal{E}_t|\mathcal{B}\right), \tag{3.11}$$

where we condition on the event $\mathcal{B}$ of (3.7). To lower bound the right hand side, it is sufficient to lower bound the left hand side. By rule of complements and union bound, we can derive a lower bound on $\Pr(\mathcal{E}_t|\mathcal{B})$ by upper bounding:

$$\sum_{j=1}^{t-1}\sum_{\pi_0\in\Pi_{\mathcal{G}}}\sum_{k\in V(j;\pi_0)\cup I(j;\pi_0)}\left\{\Pr\left(\mathcal{F}_j^C(k,\sigma;\pi_0)\middle|\mathcal{B}\right)+\Pr\left(\mathcal{F}_j^C(k,\eta;\pi_0)\middle|\mathcal{B}\right)\right\}, \tag{3.12}$$

where B^C denotes the complement of an event B.

Therefore, the key for our argument is to upper bound:

$$\Pr\left(\mathcal{F}_j^C(k,\sigma;\pi_0)\middle|\mathcal{B}\right)+\Pr\left(\mathcal{F}_j^C(k,\eta;\pi_0)\middle|\mathcal{B}\right)$$

for any $j=1,\ldots,t-1$, any $k\notin\mathcal{A}_j^{\pi_0}$, and any $\pi_0\in\Pi_{\mathcal{G}}$. We will show that this upper bound goes to zero at an exponentially decaying rate, thus making the union bound in (3.12) less severe to our overall rate. This result is given by Lemma 3.A.2 which builds on Lemma 3.A.1. Lemma 3.A.1 makes use of the linear regression results in Section 3.A.3.

Recall our definition of event $\mathcal{E}_t$ for $2\leq t\leq p+1$:

$$\mathcal{E}_t\doteq\bigcap_{j=1}^{t-1}\{PA\left(\hat{\pi}(j)\right)\subseteq\mathcal{A}_j\}.$$

To finish our proof of Theorem 3.2.2, we need to lower bound $Pr\left(\mathcal{E}_{p+1}|\mathcal{B}\right)$. The key will be to derive this lower bound while making use of Lemma 3.A.2 as well as the statement about the containment of our events of interest in (3.10).

Recall that (3.10) implies (3.11), which together with (3.12) (the place where we used the union bound) says that:

$$\Pr(\mathcal{E}_{p+1}|\mathcal{B})\geq 1-\sum_{j=1}^{p}\sum_{\pi_0\in\Pi_{\mathcal{G}}}\sum_{k\in V(j;\pi_0)\cup I(j;\pi_0)}\left\{\Pr\left(\mathcal{F}_j^C(k,\sigma;\pi_0)\middle|\mathcal{B}\right)+\Pr\left(\mathcal{F}_j^C(k,\eta;\pi_0)\middle|\mathcal{B}\right)\right\}$$

Combining Lemma 3.A.2 with the previous inequality, we have that:

$$\Pr(\mathcal{E}_{p+1}|\mathcal{B}) \geq 1 - \sum_{\pi_0 \in \Pi_\mathcal{G}} \sum_{j=1}^{p} \sum_{k \in V(j;\pi_0) \cup I(j;\pi_0)} \frac{8}{n^\xi} \asymp 1 - \frac{8|\Pi_\mathcal{G}|p^2}{n^\xi},$$

where $|\Pi_\mathcal{G}|$ is the cardinality of the set $\Pi_\mathcal{G}$, i.e. to the total number of valid permutations.

This concludes the proof of Theorem 3.2.2. The rest of this section contains the key lemmas used to derive the upper bounds on the deviation between the scale parameter estimates and the true counterparts, which were used in the proof of Lemma 3.A.1.

3.A.2 Lemmas for Theorem 3.2.2

The task of Lemma 3.A.1 is to place the generic results of Section 3.A.3 in the notation of our LiNGAM's sorting procedure, including all relevant assumptions. In the previous subsection, we apply Lemma 3.A.2 in the special case that $\mathbb{E}[R_{kt}^{\pi_0}|X_{\widehat{N}_{kt}^{\pi_0}}] = 0$, i.e. in the case that Assumption 3.2.5 is satisfied with $\rho_{n,\xi} = 0$ and $\tau_{max} = 0$. Lemma 3.A.2 is a strategic application of Lemma 3.A.1 through a use of the Mean Value Theorem.

Lemma 3.A.1 (Deviation between Scale Parameter Estimates and Truth)**.**
Let $\pi_0 \in \Pi_\mathcal{G}$ Assumption 2.2.3, denoted as the event $\mathcal{B}$ in this subsection.
Let:

- $d \doteq \max_{k\in[p]} \left|\widehat{N}_k\right|$, *the maximum cardinality for the estimated neighborhood sets.*
- $\gamma_{max} \doteq \max_{j\in[p],k\notin\{\pi_0(i)\}_{i=1}^{j-1}} \left\|R_{kt}^{\pi_0}\right\|_{\psi_1}$, *the maximum possible sub-Exponential norm for the population-level least squares residual across permutations $\pi_0 \in \Pi_\mathcal{G}$.*
- $\eta_{max} \doteq \max_{\pi_0\in\Pi_\mathcal{G},j\in[p],k\notin\{\pi_0(i)\}_{i=1}^{t-1}} \eta_{kj}(\pi_0)$, *the maximum possible Laplace scale parameter for the population-level least squares residual across permutations $\pi_0 \in \Pi_\mathcal{G}$.*

If $d \leq n$ and Assumption 3.2.5 holds, then for arbitrary $\xi > 0$ and n large enough:

- *With probability at least*

$$1 - \frac{3}{n^\xi} \to 1,$$

we have that:

$$|\hat{\eta}_{kj}(\pi_0) - \eta_{kj}(\pi_0)| \quad < \frac{2d(\gamma_{max} + \tau_{max} + \eta_{max}/\log(2))(1+\xi)\log(n)}{c^{1/2}n^{1/2}} + \rho_{n,\xi}.$$

- *With probability at least*

$$1 - \frac{5}{n^{\xi}} \to 1,$$

we have that:

$$\begin{aligned}|\hat{\sigma}^2_{kj}(\pi_0) - \sigma^2_{kj}(\pi_0)| \quad \leq & \frac{(1+\xi)\gamma_{max}\log^{3/2}(2n)\sqrt{32}}{cn^{1/2}} + \frac{d\gamma_{max}(\gamma_{max}+\tau_{max})(1+\xi)\log^2(n)}{n^{1/2}c} \\ & \frac{d\gamma_{max}(1+\xi)\log^2(n)}{n^{1/2}c}\rho_{n,\xi} + \frac{d\gamma^2_{max}(1+\xi)^2\log^2(n)}{n^{1/2}c^2} + \rho_{n,\xi}.\end{aligned}$$

Proof.

The general forms of Lemma 3.A.9 and Lemma 3.A.13 give us the desired conclusion. Note that Assumption 3.2.5 in the main text defines $\rho_{max,\xi}$ and τ_{max} in Analogy to Section 3.A.3's Assumptions 3.A.4 and 3.A.5.

□

Lemma 3.A.2 (Showing Events $\mathcal{F}^C_j(k,\eta;\pi_0)$ and $\mathcal{F}^C_j(k,\sigma;\pi_0)$ have probability going to 0)**.** *Let $\pi_0 \in \Pi_{\mathcal{G}}$. Also let:*

- $d \doteq \max_{k\in[p]} \left|\widehat{N}_k\right|$, *the maximum cardinality for the estimated neighborhood sets as.*
- $\gamma_{max} \doteq \max_{j\in[p],k\notin\{\pi_0(i)\}_{i=1}^{j-1}} \|R^{\pi_0}_{kt}\|_{\psi_1}$, *the maximum possible sub-Exponential norm for the population-level least squares residual across permutations $\pi_0 \in \Pi_{\mathcal{G}}$.*
- $\eta_{max} \doteq \max_{\pi_0\in\Pi_{\mathcal{G}},j\in[p],k\notin\{\pi_0(i)\}_{i=1}^{t-1}} \eta_{kj}(\pi_0)$, *the maximum possible Laplace scale parameter for the population-level least squares residual across permutations $\pi_0 \in \Pi_{\mathcal{G}}$.*

If $d \leq n$, Assumption 3.2.5 holds, and n large enough so that:

$$\left(\frac{\delta^{\pi_0}_j}{\delta^{\pi_0}_j + 4}\right)\eta_{kj}(\pi_0) \quad > \frac{2d(\gamma_{max} + \tau_{max} + \eta_{max}/\log(2))(1+\xi)\log(n)}{c^{1/2}n^{1/2}} + \rho_{n,\xi}, \tag{3.13}$$

then for arbitrary $\xi > 0$:

$$Pr\left(\mathcal{F}_j^C(k,\eta;\pi_0)\middle|\mathcal{B}\right) \leq \frac{3}{n^\xi} \to 0.$$

And if additionally, n large enough so that:

$$\begin{aligned}\left(\frac{\delta_j^{\pi_0}}{\delta_j^{\pi_0}+4}\right)\sigma_{kj}(\pi_0) \;\; > \;\; & \frac{(1+\xi)\gamma_{max}\log^{3/2}(2n)\sqrt{32}}{cn^{1/2}} + \frac{d\gamma_{max}(\gamma_{max}+\tau_{max})(1+\xi)\log^2(n)}{n^{1/2}c} \\ & \frac{d\gamma_{max}(1+\xi)\log^2(n)}{n^{1/2}c}\rho_{n,\xi} + \frac{d\gamma_{max}^2(1+\xi)^2\log^2(n)}{n^{1/2}c^2} + \rho_{n,\xi}.\end{aligned} \tag{3.14}$$

then:

$$Pr\left(\mathcal{F}_j^C(k,\sigma;\pi_0)\middle|\mathcal{B}\right) \leq \frac{5}{n^\xi} \to 0.$$

Proof.

Consider the mean value theorem which says that for $f : [a,b] \to \mathbb{R}$ which is continuous and at least once differentiable on the open interval (a,b), there exists $c \in (a,b)$ such that:

$$f(b) - f(a) = (b-a)f'(c).$$

This implies that:

$$|f(b) - f(a)| \leq |b-a| \sup_{a<x<b} |f'(x)|. \tag{3.15}$$

Now let $f(t) = log(t)$ with t restricted to be in $[a,b]$ where:

$$a < \min\left\{\eta_{kj}(\pi_0), \hat{\eta}_{kj}(\pi_0)\right\} \text{ and } b > \max\left\{\eta_{kj}(\pi_0), \hat{\eta}_{kj}(\pi_0)\right\}.$$

It follows from (3.15) that:

$$|\log(\eta_{kj}(\pi_0)) - \log(\hat{\eta}_{kj}(\pi_0))| \leq \frac{|\eta_{kj}(\pi_0) - \hat{\eta}_{kj}(\pi_0)|}{a}, \tag{3.16}$$

since $\sup_{a<x<b}|f'(x)| \leq \frac{1}{a}$ and

$$|\log(\eta_{kj}(\pi_0)) - \log(\hat{\eta}_{kj}(\pi_0))| \leq |\log(b) - \log(a)|,$$

due to monotonicity of $t \mapsto \log(t)$.

From Lemma 3.A.1, we know that with probability at least $1 - \frac{3}{n^{\xi}}$:

$$|\eta_{kj}(\pi_0) - \hat{\eta}_{kj}(\pi_0)| < r_n < \kappa, \tag{3.17}$$

where κ is a constant we'd like to derive and r_n is shorthand for:

$$r_n \doteq \frac{2d(\gamma_{max} + \tau_{max} + \eta_{max}/\log(2))(1+\xi)\log(n)}{c^{1/2}n^{1/2}} + \rho_{n,\xi}.$$

We have that:

$$\hat{\eta}_{kj}(\pi_0) \in (\eta_{kj}(\pi_0) - \kappa, \eta_{kj}(\pi_0) + \kappa).$$

So we can take:

$$a = \eta_{kj}(\pi_0) - \kappa \text{ and } b = \eta_{kj}(\pi_0) + \kappa.$$

With this choice of a, the implication of (3.16) and (3.17) is that:

$$|\log(\eta_{kj}(\pi_0)) - \log(\hat{\eta}_{kj}(\pi_0))| < \frac{r_n}{\eta_{kj}(\pi_0) - \kappa}$$

with probability at least $1 - \frac{3}{n^{\xi}}$.

Recall that $\mathcal{F}_j(k, \eta; \pi_0)$ is defined as the occurrence of the inequality

$$|\log(\eta_{kj}(\pi_0)) - \log(\hat{\eta}_{kj}(\pi_0))| < \delta_j^{\pi_0}/4.$$

We can thus see that:

$$\frac{r_n}{\eta_{kj}(\pi_0) - \kappa} \leq \delta_j^{\pi_0}/4 \iff \kappa \leq \eta_{kj}(\pi_0) - \frac{4r_n}{\delta_j^{\pi_0}}.$$

Setting $\kappa = \eta_{kj}(\pi_0) - \frac{4r_n}{\delta_j^{\pi_0}}$, from (3.17), we require that:

$$r_n < \eta_{kj}(\pi_0) - \frac{4r_n}{\delta_j^{\pi_0}} \iff r_n < \left(\frac{\delta_j^{\pi_0}}{\delta_j^{\pi_0} + 4}\right)\eta_{kj}(\pi_0),$$

which is satisfied for large enough n.

Thus, for n large enough so that $r_n < \left(\frac{\delta_j^{\pi_0}}{\delta_j^{\pi_0}+4}\right)\eta_{kj}(\pi_0)$,

$$Pr(\mathcal{F}_j^C(k, \eta; \pi_0)) \leq \frac{3}{n^{\xi}},$$

as we wanted.

By a similar argument, so long as n is large enough, we have that:

$$Pr(\mathcal{F}_j^C(k, \sigma; \pi_0)) \leq \frac{5}{n^\xi}.$$

□

3.A.3 Full Column Rank Linear Regression with sub-Exponential Noise

This section contains the core lemmas for bounding the finite sample performance of our sorting procedure. We make use of the regression setup in Assumption 3.A.3.

Assumption 3.A.3 (A Regression Setup with sub-Exponential Noise)**.**
Let:

- $\mathbf{X} \in \mathbb{R}^{n \times m}$ with $\text{rank}(\mathbf{X}) = m$ almost surely, and $m \leq d \leq n$.
- $\beta \in \mathbb{R}^m$.
- $Y = \mathbf{X}\beta + U \in \mathbb{R}^n$.
- $U = Y - \mathbf{X}\beta \in \mathbb{R}^n$ such that $U_i \overset{i.i.d.}{\sim}$ sub-Exponential for $i = 1, 2, \ldots, n$ with $\mathbb{E}[U_i] = 0$.
- $\|U_i - \mathbb{E}[U_i|\mathbf{X}_{i\cdot}]\|_{\psi_1} \leq s$, where $\|V\|_{\psi_1} \doteq \inf\{t > 0 : \mathbb{E}\left[\exp\right](|V|/t) \leq 2\}$ is known as the sub-Exponential norm of scalar random variable V.
- $\mathbb{E}[|U_i|^4] < \infty$.
- $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y = \beta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T U$ the linear least squares estimate.
- $P = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, the projection matrix onto column space of $\mathbf{X}$.

Assumption 3.A.4 (Projection Assumption)**.**
There exists a non-negative sequence $r_{n,\xi}$ and $\xi > 0$ such that for n large enough,

$$\frac{1}{n}\left\|P\mathbb{E}\left[Y|\mathbf{X}\right] - X\beta\right\|_1 \leq r_{n,\xi}$$

and

$$\frac{1}{n}\left\|P\mathbb{E}\left[Y|\mathbf{X}\right]-X\beta\right\|_2^2\leq r_{n,\xi}$$

with probability at least $1-\frac{1}{n^{\xi}}$.

These two inequalities say that $r_{n,\xi}$ is, with a certain probability, the worst case sample average deviation between $\mathbf{X}_{i\cdot}\beta$ and the projection of $\mathbb{E}[Y_i|\mathbf{X}_{i\cdot}]$ onto the column space of $\mathbf{X}$.

Assumption 3.A.5 (sub-exponential norm assumption)**.**
There exists $\tau\geq 0$ such that:

$$\left\|\mathbb{E}[U_i|\mathbf{X}_{i\cdot}]\right\|_{\psi_1}=\left\|\mathbb{E}[Y_i|\mathbf{X}_{i\cdot}]-\mathbf{X}_{i\cdot}\beta\right\|_{\psi_1}\leq\tau.$$

Remark 3.A.6 (When Y is linear in $\mathbf{X}$)**.**
Importantly, if $\mathbb{E}[Y|\mathbf{X}]=\mathbf{X}\beta$, then $r_{n,\xi}=0$ is a valid choice for Assumption 3.A.4 while $\tau=0$ is a valid choice for Assumption 3.A.5. The former is because $P\mathbf{X}\beta=\mathbf{X}\beta$ by properties of P, while the latter holds by noting that $\mathbb{E}[U_i|\mathbf{X}_{i\cdot}]=0$ in this case.

This section how to bound the differences between:

- Prediction estimate $\hat{Y}=\mathbf{X}\hat{\beta}$ and $\mathbf{X}\beta$ in terms of the ℓ_1-norm and in terms of the ℓ_2-norm in Lemmas 3.A.7 and 3.A.11.
- Plugin estimate $\frac{1}{n}\left\|Y-\mathbf{X}\hat{\beta}\right\|_1$ and $\mathbb{E}[|U_i|]$ in Lemma 3.A.9.
- Plugin estimate $\frac{1}{n}\left\|Y-\mathbf{X}\hat{\beta}\right\|_2^2$ and $\mathbf{V}[U_i]$ in Lemma 3.A.13.

The rest of the lemmas help to prove these results.

Lemma 3.A.7 (Exponential decay probability for ℓ_1-norm deviation of sample linear least squares prediction from population linear least squares prediction)**.**

For the setup given by Assumption 3.A.3 and Assumption 3.A.4, we have that for n large enough and absolute constant c:

$$\frac{1}{n}\left\|\mathbf{X}\hat{\beta}-\mathbf{X}\beta\right\|_1=\frac{1}{n}\left\|PU\right\|_1\leq\frac{ds(1+\xi)\log(n)}{\sqrt{n}c}+r_n$$

with probability at least $1-\frac{3}{n^\xi}\to 1$.

Proof.

Let $v_1, v_2, \ldots, v_d$ be an orthonormal basis for $\text{range}(P)=\text{col}(\mathbf{X})$, the column space of $\mathbf{X}$ such that the inner product

$$\langle v_j, v_i\rangle=\delta_{ij}=\begin{cases}1 & \text{if } i=j\\ 0 & \text{o/w}\end{cases},$$

which exists if we take any basis for $\text{col}(\mathbf{X})$ and pass it through a Gram-Schmidt process.

Let $\tilde{U}=U-\mathbb{E}[U|\mathbf{X}]$. For our quantity of interest, we have:

$$\begin{aligned}
\frac{1}{n}\left\|PU\right\|_1 &= \frac{1}{n}\left\|P\tilde{U}\right\|_1+\frac{1}{n}\left\|P\mathbb{E}[U|\mathbf{X}]\right\|_1 && \text{triangle inequality}\\
&\leq \frac{1}{n}\left\|P\tilde{U}\right\|_1+r_n && \text{Assumption 3.A.4, with probability } \geq 1-\frac{1}{n^\xi}\\
&= \frac{1}{n}\left\|\sum_{j=1}^{d}\langle v_j,\tilde{U}\rangle v_j\right\|_1+r_n && \text{span}\{v_1,\ldots,v_d\}=\text{col}(\mathbf{X})\\
&\leq \frac{1}{n}\sum_{j=1}^{d}\left|\langle v_j,\tilde{U}\rangle\right|\left\|v_j\right\|_1+r_n && \text{triangle inequality}\\
&\leq \frac{1}{n}\sum_{j=1}^{d}\left|\langle v_j,\tilde{U}\rangle\right|\sqrt{n}\left\|v_j\right\|_2+r_n && \text{b/c } \left\|\cdot\right\|_1\leq\sqrt{n}\left\|\cdot\right\|_2 \text{ in } \mathbb{R}^n\\
&= \frac{1}{\sqrt{n}}\sum_{j=1}^{d}\left|\langle v_j,\tilde{U}\rangle\right|+r_n && \text{b/c } \left\|v_j\right\|_2=1\\
&\leq \frac{d}{\sqrt{n}}\max_{j=1,\ldots,d}\left|\langle v_j,\tilde{U}\rangle\right|+r_n
\end{aligned} \tag{3.18}$$

Now consider each $\langle v_j, \tilde{U} \rangle$, which is zero-centered conditional on $\mathbf{X}$. Theorem 2.8.2 of Vershynin (2018), Bernstein's inequality for a linear combination of independent zero-centered sub-Exponential random variables, tells us that for $t \geq 0$

$$\begin{aligned} Pr\left(\left|\langle v_j, \tilde{U} \rangle\right| \geq t \middle| \mathbf{X}\right) &\leq 2\exp\left[-c\min\left(\frac{t^2}{\left[\max_{i=1,\ldots,n}\left\|\tilde{U}_i\right\|_{\psi_1}\right]^2 \|v_j\|_2^2}, \frac{t}{\max_{i=1,\ldots,n}\left\|\tilde{U}_i\right\|_{\psi_1}\|v_j\|_\infty}\right)\right] \\ &\leq 2\exp\left[-c\min\left(\frac{t^2}{s^2}, \frac{t}{s}\right)\right], \end{aligned}$$

where c is an absolute constant. The second inequality holds because $\|v_j\|_\infty \leq \|v_j\|_2 = 1$, and because $\max_{i=1,\ldots,n} \|U_i\|_{\psi_1} \leq s$ by assumption. Note that the second inequality does not depend on $\mathbf{X}$, so we also have unconditionally

$$Pr\left(\left|\langle v_j, \tilde{U} \rangle\right| \geq t\right) \leq 2\exp\left[-c\min\left(\frac{t^2}{s^2}, \frac{t}{s}\right)\right] = \begin{cases} \exp\left(-c\frac{t^2}{s^2}\right) & \text{if } t \leq s \\ \exp\left(-c\frac{t}{s}\right) & \text{o/w} \end{cases}.$$

So by union bound,

$$Pr\left(\max_{j=1,\ldots,d}\left|\langle v_j, \tilde{U} \rangle\right| \geq t\right) \leq \begin{cases} 2d\exp\left(-c\frac{t^2}{s^2}\right) & \text{if } t \leq s \\ 2d\exp\left(-c\frac{t}{s}\right) & \text{o/w} \end{cases}.$$

Letting $\xi > 0$ be arbitrary and $t^* = s(1+\xi)\log(n)/c$ and $n > N_{\xi,s,c}$ large enough such that $t^* > s$, we have:

$$Pr\left(\max_{j=1,\ldots,d}\left|\langle v_j, \tilde{U} \rangle\right| \geq t^*\right) \leq \frac{2d}{(n)^{1+\xi}} \leq \frac{2}{(n)^{\xi}} \to 0.$$

It follows that

$$\frac{1}{n}\left\|\mathbf{X}\hat{\beta} - \mathbf{X}\beta\right\|_1 \leq \frac{d}{\sqrt{n}}\max_{j=1,\ldots,d}\left|\langle v_j, \tilde{U} \rangle\right| + r_n \leq \frac{ds(1+\xi)\log(n)}{\sqrt{n}c} + r_n$$

with probability at least $1 - \frac{2}{n^\xi} \to 1$.

□

Lemma 3.A.8 (Sample Estimate of $\|\cdot\|_{L_1}$)**.**

Consider the setup given by Assumption 3.A.3 and Assumption 3.A.5. Denote $b = \|U_i\|_{L_1} = \mathbb{E}[|U_i|]$*. Also let* $\xi > 0$ *be arbitrary.*

Then for n *large enough and absolute constant* c*:*

$$\left|\frac{1}{n}\sum_{i=1}^{n}(|U_i| - b)\right| \leq \frac{(s + \tau + b/log(2))\log(n)}{c^{1/2}n^{1/2}}$$

with probability at least $1 - \frac{1}{n^{\xi}} \to 1$*.*

Proof.

Consider the sub-Exponential norm of the zero-centered r.v. $|U_i| - b$:

$$\begin{aligned}
\||U_i| - b\|_{\psi_1} &\leq \||U_i|\|_{\psi_1} + \|b\|_{\psi_1} \\
&= \|U_i\|_{\psi_1} + \frac{b}{\log(2)} \\
&= \|U_i - \mathbb{E}[U_i|\mathbf{X}_{i\cdot}]\|_{\psi_1} + \|\mathbb{E}[U_i|\mathbf{X}_{i\cdot}]\|_{\psi_1} + \frac{b}{\log(2)} \\
&\leq s + \tau + \frac{b}{\log(2)} \doteq \check{s},
\end{aligned} \tag{3.19}$$

with the second line holding by $\||U_i|\|_{\psi_1} = \|U_i\|_{\psi_1}$, while $\|c\|_{\psi_1} = \frac{|c|}{\log(2)}$ for any constant c. The fourth line holds by Assumptions 3.A.3 and 3.A.5.

Theorem 2.8.2 of Vershynin (2018), Bernstein's inequality for a linear combination of independent zero-centered sub-Exponential random variables, tells us that for $t \geq 0$:

$$\begin{aligned}
Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}(|U_i| - b)\right| \geq t\right) &\leq 2\exp\left[-cn\min\left(\frac{t^2}{\left[\max_{i=1,\ldots,n}\||U_i| - b\|_{\psi_1}\right]^2}, \frac{t}{\max_{i=1,\ldots,n}\||U_i| - b\|_{\psi_1}}\right)\right] \\
&\leq 2\exp\left[-cn\min\left(\frac{t^2}{\check{s}^2}, \frac{t}{\check{s}}\right)\right],
\end{aligned}$$

where c is an absolute constant.

Set $t = \frac{\check{s}\xi \log(n)}{c^{1/2}n^{1/2}}$, and assume n large enough so that $t \leq \check{s}$. Thus, $\min\left(\frac{t^2}{\check{s}^2}, \frac{t}{\check{s}}\right) = \frac{t^2}{\check{s}^2}$. We have:

$$Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}(|U_i| - b)\right| \geq \frac{\check{s}\xi \log(n)}{\sqrt{cn}}\right) \leq \frac{2}{n^{\xi}} \to 0.$$

□

Lemma 3.A.9 (Laplace Shape Parameter Estimation with OLS Residuals)**.**

Consider the setup given by Assumptions 3.A.3, 3.A.4, and 3.A.5. Let our estimate of $b \doteq \|Y_i - \mathbf{X}_{i\cdot}\beta\|_{L_1}$ *be:*

$$\hat{b} \doteq \frac{1}{n}\left\|Y - \mathbf{X}\hat{\beta}\right\|_1.$$

Then for n *large enough and absolute constant* c*:*

$$|\hat{b} - b| \leq \frac{2d(s + \tau + b/log(2))(1+\xi)\log(n)}{c^{1/2}n^{1/2}} + r_n$$

with probability at least

$$1 - \frac{2}{n^{\xi}} \to 0.$$

Proof.

Note that:

$$\begin{aligned} \hat{b} &\leq \frac{1}{n}\left\|\mathbf{X}\beta - \mathbf{X}\hat{\beta}\right\|_1 + \frac{1}{n}\|U\|_1 \\ &= \frac{1}{n}\|PU\|_1 + \frac{1}{n}\|U\|_1 \end{aligned} \tag{3.20}$$

by triangle inequality. Similarly,

$$\begin{aligned} \hat{b} &= \frac{1}{n}\|U + PU\|_1 \\ &\geq \left|\frac{1}{n}\|U\|_1 - \frac{1}{n}\|-PU\|_1\right| \\ &\geq \frac{1}{n}\|U\|_1 - \frac{1}{n}\|PU\|_1 \end{aligned} \tag{3.21}$$

by reverse triangle inequality. Thus,

$$\hat{b} - b \leq \frac{1}{n} \|PU\|_1 + \frac{1}{n} \|U\|_1 - b \leq \frac{1}{n} \|PU\|_1 + \left| \frac{1}{n} \|U\|_1 - b \right|.$$

Similarly,

$$b - \hat{b} \leq b - \frac{1}{n} \|U\|_1 + \frac{1}{n} \|PU\|_1 \leq \frac{1}{n} \|PU\| + \left| \frac{1}{n} \|U\|_1 - b \right|.$$

So overall,

$$|\hat{b} - b| \leq \frac{1}{n} \|PU\| + \left| \frac{1}{n} \|U\|_1 - b \right|.$$

By Lemmas 3.A.7 and 3.A.8, we have that

$$|\hat{b}-b| \leq \frac{(s+\tau+b/log(2))\log(n)}{c^{1/2}n^{1/2}} + \frac{ds(1+\xi)\log(n)}{\sqrt{n}c} + r_n \leq \frac{2d(s+\tau+b/log(2))(1+\xi)\log(n)}{\sqrt{n}c} + r_n$$

with probability at least

$$1 - \frac{2}{n^{\xi}} \to 1.$$

□

Lemma 3.A.10 (Variance Estimate for Sub-Exponential Random Variables)**.**

Let Assumptions 3.A.3 and 3.A.5 hold. Also assume $\mathbb{E}[U_i^4] < \infty$ *(implying that* $\mathbf{V}[U_i^2] < \infty$*).*

Consider the deviation of the variance estimate $\frac{1}{n}\sum_{i=1}^{n} U_i^2$ *from* $\mathbf{V}[U_i] = \mathbb{E}[U_i^2] = \sigma^2$*.*

For arbitrary $\xi > 0$*, n large enough, and abslute constant c, we have that:*

$$\left| \left(\frac{1}{n} \sum_{i=1}^{n} U_i^2 \right) - \sigma^2 \right| \leq \frac{(1+\xi)(s+\tau)\log^{3/2}(2n)\sqrt{32}}{cn^{1/2}} \to 0$$

with probability at least

$$1 - \frac{1}{(2n)^{\xi}} \to 1.$$

Proof.

Indexing across $f \in \mathcal{F}$, we will make use of the Rademacher process

$$f \mapsto \sum_{i=1}^{n} \rho_i Z_i(f),$$

where $(\rho_1, \rho_2, \ldots, \rho_n) \perp\!\!\!\perp (U_1, U_2, \ldots, U_n)$, $\rho_i \overset{i.i.d.}{\sim} \text{Unif}\{-1, 1\}$, and $Z_i(f) \doteq \frac{1}{n} f(U_i)$. In our case, we will take $\mathcal{F}$ to be a singleton set with element $t \mapsto t - \sigma^2$.

Lemma 2.3.7 (Symmetrization for probabilities) of van der Vaart and Wellner (1996) tells us that for our iid stochastic processes $Z_1(f), \ldots, Z_n(f)$ and arbitrary functionals $\mu_1, \ldots, \mu_n : \mathcal{F} \mapsto \mathbb{R}$,

$$\beta_n(x) Pr\left(\sup_{f \in \mathcal{F}} \left|\sum_{i=1}^{n} Z_i(f)\right| > x\right) \leq 2Pr\left(4 \sup_{f \in \mathcal{F}} \left|\sum_{i=1}^{n} \rho_i (Z_i(f) - \mu_i(f))\right| > x\right), \tag{3.22}$$

for every $x > 0$ and

$$\beta_n(x) = \inf_f Pr\left(\left|\sum_{i=1}^{n} Z_i(f)\right| < x/2\right). \tag{3.23}$$

We will take $\mu_i(f) = -\sigma^2$ for each $i = 1, 2, \ldots, n$.

Noting that in our case we have a supremum across a singleton set $\mathcal{F}$, we can re-write the inequality in (3.22) as:

$$Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left(U_i^2 - \sigma^2\right)\right| > x\right) \leq \left[\frac{1}{\beta_n(x)}\right] 2Pr\left(4\left|\frac{1}{n}\sum_{i=1}^{n}\rho_i U_i^2\right| > x\right). \tag{3.24}$$

Now, consider the event

$$\mathcal{E}_\gamma = \{\max_{i=1,\ldots,n} |U_i| \leq \gamma\} = \{\max_{i=1,\ldots,n} U_i^2 \leq \gamma^2\}.$$

Note that

$$Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}\rho_i U_i^2\right| > x/4\right) \leq Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}\rho_i U_i^2\right| > x/4 \middle| \mathcal{E}_\gamma\right) + Pr(\mathcal{E}_\gamma^C),$$

since $Pr(\cdot) = Pr(\cdot|\mathcal{E}_\gamma)Pr(\mathcal{E}_\gamma) + Pr(\cdot|\mathcal{E}_\gamma^C)Pr(\mathcal{E}_\gamma^C) \leq Pr(\cdot|\mathcal{E}_\gamma) + Pr(\mathcal{E}_\gamma^c)$.

We want for the upper bound

$$Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}\rho_i U_i^2\right| > x/4 \middle| \mathcal{E}_\gamma\right) + Pr(\mathcal{E}_\gamma^C)$$

to go to zero.

By union bound and properties of sub-Exponential random variables, we have that:

$$Pr(\mathcal{E}_\gamma^C) \leq nPr(|U_i| > \gamma) \leq 2n\exp\left[-c\min\left(\frac{\gamma^2}{(s+\tau)^2}, \frac{\gamma}{s+\tau}\right)\right],$$

Here, we used that $\|U_i\|_{\psi_1} \leq \|U_i - \mathbb{E}[U_i|\mathbf{X}_{i\cdot}]\|_{\psi_1} + \|\mathbb{E}[U_i|\mathbf{X}_{i\cdot}\|_{\psi_1} \leq s + \tau$.

Let $\gamma = (1+\xi)(s+\tau)\log(2n)/c$ and n large enough so that $\min\left(\frac{\gamma^2}{(s+\tau)^2}, \frac{\gamma}{s+\tau}\right) = \frac{\gamma}{s+\tau}$. Thus,

$$Pr(\mathcal{E}_\gamma^C) \leq nPr(|U_i| > \gamma) \leq \frac{2n}{(2n)^{1+\xi}} = \frac{1}{(2n)^\xi} \to 0.$$

Now consider that, conditional on $\mathcal{E}_\gamma$, $\rho_i U_i^2 \in [-\gamma^2, \gamma^2]$ almost surely. Hoeffding's inequality for sums of bounded random variables thus tells us that:

$$Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}\rho_i U_i^2\right| > x/4 \middle| \mathcal{E}_\gamma\right) = Pr\left(\left|\sum_{i=1}^{n}\rho_i U_i^2\right| > nx/4 \middle| \mathcal{E}_\gamma\right) \leq \exp\left[-\frac{2(nx/4)^2}{n(2\gamma)^2}\right] = \exp\left[-\frac{nx^2}{32\gamma^2}\right].$$

Letting $x = \frac{\gamma\sqrt{32\xi\log(2n)}}{n^{1/2}}$, we have that conditional on $\mathcal{E}_\gamma$,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\rho_i U_i^2\right| > \frac{(1+\xi)(s+\tau)\log^{3/2}(2n)}{cn^{1/2}} \to 0$$

with probability at most:

$$\frac{1}{(2n)^\xi} \to 0.$$

So unconditionally:

$$Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}\rho_i U_i^2\right| > x/4\right) \leq Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}\rho_i U_i^2\right| > x/4 \middle| \mathcal{E}_\gamma\right) + Pr(\mathcal{E}_\gamma^C) \leq \frac{2}{(2n)^\xi} \to 0. \quad (3.25)$$

Now recall our choice of $\beta_n(x)$. Base on $Z_i(f) = \frac{1}{n}(U_i^2 - \sigma^2)$, we have:

$$\begin{aligned}
\beta_n(x) &= 1 - Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}(U_i^2 - \sigma^2)\right| \geq x/2\right) \\
&\geq 1 - \frac{4}{x^2}\mathbf{V}\left[\frac{1}{n}\sum_{i=1}^{n}(U_i^2 - \sigma^2)\right] && \text{by Chebyshev's Inequality} \\
&= 1 - \frac{4}{nx^2}\mathbf{V}[U_i^2] \\
&= 1 - \frac{c^2}{8\xi(1+\xi)^2(s+\tau)^2\log^3(2n)} && \text{based on our choice of } x.
\end{aligned} \tag{3.26}$$

For large enough n, note that $\beta_n(x) \geq 1/2$. Thus, for this large enough n and our choice of x, (3.24) and (3.25) tell us that:

$$\left|\left(\frac{1}{n}\sum_{i=1}^{n}U_i^2\right) - \sigma^2\right| \leq \frac{(1+\xi)(s+\tau)\log^{3/2}(2n)\sqrt{32}}{cn^{1/2}} \to 0$$

with probability at least

$$1 - \frac{1}{(2n)^{\xi}} \to 1.$$

□

Lemma 3.A.11 (Exponential decay probability for ℓ_2-norm deviation of sample linear least squares prediction from population linear least squares prediction)**.**

For the setup given by Assumptions 3.A.3, 3.A.4, we have that for n large enough and absolute constant c:

$$\frac{1}{n}\left\|\mathbf{X}\hat{\beta} - \mathbf{X}\beta\right\|_2^2 \leq \frac{ds^2(1+\xi)^2\log^2(n)}{nc^2} + r_n$$

with probability at least $1 - \frac{3}{n^{\xi}} \to 1$.

Proof.

Let $v_1, v_2, \ldots, v_d$ be an orthonormal basis for $\text{range}(P) = \text{col}(\mathbf{X})$, the column space of $\mathbf{X}$ such that the inner product

$$\langle v_j, v_i\rangle = \delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{o/w} \end{cases},$$

which exists if we take any basis for col($\mathbf{X}$) and pass it through a Gram-Schmidt process.

Let $\tilde{U} = U - \mathbb{E}[U|\mathbf{X}]$.

For our quantity of interest, we have:

$$\begin{aligned}
\frac{1}{n}\left\|PU\right\|_2^2 &\leq \frac{1}{n}\left\|P\tilde{U}\right\|_2^2 + \frac{1}{n}\left\|P\mathbb{E}[U|\mathbf{X}]\right\|_2^2 \\
&\leq \frac{1}{n}\left\|P\tilde{U}\right\|_2^2 + r_n && \text{Assumption 3.A.4, with probability } \geq 1 - \frac{1}{n^\xi} \\
&= \frac{1}{n}\left\|\sum_{j=1}^{d}\langle v_j, \tilde{U}\rangle v_j\right\|_2^2 + r_n \\
&= \frac{1}{n}\sum_{j=1}^{d}\left|\langle v_j, \tilde{U}\rangle\right|^2\left\|v_j\right\|_2^2 + r_n && \text{Pythagoras' theorem} \\
&= \frac{1}{n}\sum_{j=1}^{d}\left|\langle v_j, \tilde{U}\rangle\right|^2 + r_n && \text{because } \left\|v_j\right\|_2 = 1 \\
&\leq \frac{d}{n}\max_{j=1,\ldots,d}\left|\langle v_j, \tilde{U}\rangle\right|^2 + r_n \\
&= \frac{d}{n}\left(\max_{j=1,\ldots,d}\left|\langle v_j, \tilde{U}\rangle\right|\right)^2 + r_n.
\end{aligned} \tag{3.27}$$

Now consider an arbitrary $\langle v_j, \tilde{U}\rangle$, which is zero-centered conditional on $\mathbf{X}$. Theorem 2.8.2 of Vershynin (2018), Bernstein's inequality for a linear combination of independent zero-centered sub-Exponential random variables, and a similar argument to Lemma 3.A.7 tell us the following. For arbitrary $\xi > 0$ and $t^* = s(1+\xi)\log(n)/c$ and $n > N_{\xi,s,c}$ large enough such that $t^* > s$, we have:

$$Pr\left(\max_{j=1,\ldots,d}\left|\langle v_j, U\rangle\right| \geq t^*\right) \leq \frac{2d}{(n)^{1+\xi}} \leq \frac{2}{n^\xi} \to 0.$$

It follows that

$$\frac{1}{n}\left\|\mathbf{X}\hat{\beta} - \mathbf{X}\beta\right\|_2^2 \leq \frac{d}{n}\left[\max_{j=1,\ldots,d}\left|\langle v_j, U\rangle\right|\right]^2 \leq \frac{ds^2(1+\xi)^2\log^2(n)}{nc^2}$$

with probability at least $1 - \frac{3}{n^\xi} \to 1$.

□

Lemma 3.A.12 (ℓ_∞-norm of Sub-Exponential Vector)**.**

Consider the setup in Assumption 3.A.3 and 3.A.5. Let $\xi > 0$ be arbitrary. We have that

$$\|U\|_\infty \leq (s+\tau)(1+\xi)\log(n)$$

with probability at least $1 - \frac{2}{n^\xi} \to 1$.

Proof.

Let $t > 0$. Consider:

$$\begin{aligned}
Pr(\|U\|_\infty > t) &\leq nPr(|U_i| > t) \\
&= nPr(\exp(|U_i|/\|U_i\|_{\psi_1}) > \exp(t/\|U_i\|_{\psi_1})) \\
&\leq n\mathbb{E}\left[\exp\left(\frac{|U_i|}{\|U_i\|_{\psi_1}}\right)\right]\exp(-t/\|U_i\|_{\psi_1}) && \text{by Markov Inequality} \\
&\leq n2\exp(-t/\|U_i\|_{\psi_1}) && \text{by defn. of } \|\cdot\|_{\psi_1} \\
&\leq 2\exp(-t/(s+\tau)+\log(n)) && \text{b/c } \|U_i\|_{\psi_1} \leq s+\tau \\
&\leq 2\exp(-t/(s+\tau)+\log(n))
\end{aligned}$$

where the first inequality follows from union bound.

Let $\xi > 0$ be arbitrary. Set $t = (s+\tau)(1+\xi)\log(n)$. Thus,

$$\|U\|_\infty \leq (s+\tau)(1+\xi)\log(n)$$

with probability at least

$$1 - 2\exp\left(-\xi\log(n)\right) = 1 - \frac{2}{n^\xi} \to 1.$$

□

Lemma 3.A.13 (Plug-in Variance Estimation with OLS Residuals)**.**

Consider the setup given by Assumptions 3.A.3, 3.A.4, and 3.A.5. Denote $\hat{\sigma} = \frac{1}{\sqrt{n}} \left\| Y - \mathbf{X}\hat{\beta} \right\|_2$ *and* $\sigma = \|Y_i - \mathbf{X}_{i\cdot}\beta\|_{L_2}$*, where* $\|U\|_{L_p} = (\mathbb{E}[U^p])^{1/p}$*.*

Then for arbitrary $\xi > 0$ *and absolute constant* $c > 0$*:*

$$\begin{aligned} |\hat{\sigma}^2 - \sigma^2| \ &> \frac{(1+\xi)(s+\tau)\log^{3/2}(2n)\sqrt{32}}{cn^{1/2}} + \frac{ds(s+\tau)(1+\xi)^2\log^2(n)}{n^{1/2}c} \\ &\quad + \frac{(s+\tau)(1+\xi)\log(n)}{n^{1/2}c} r_n + \frac{ds^2(1+\xi)^2\log^2(n)}{nc^2} + r_n \end{aligned}$$

with probability at least

$$1 - \frac{5}{n^{\xi}} \to 0.$$

Proof.

By the definition of $\hat{\sigma}^2$ and the fact that $Y = \mathbf{X}\beta + U$, we obtain

$$\hat{\sigma}^2 = \frac{1}{n}U^TU + \frac{2}{n}U^T\mathbf{X}(\beta - \hat{\beta}) + \frac{1}{n}\left\|\mathbf{X}(\beta - \hat{\beta})\right\|_2^2.$$

Denoting $\vec{\sigma} = (\sigma, \sigma, \ldots, \sigma)^T \in \mathbb{R}^n$, we can see that:

$$|\hat{\sigma}^2 - \sigma^2| \leq \frac{1}{n}|U^TU - \vec{\sigma}^T\vec{\sigma}| + \frac{2}{n}|U^T\mathbf{X}(\beta - \hat{\beta})| + \frac{1}{n}\|PU\|_2^2.$$

For the middle term on the right of the inequality, we can apply Cauchy-Schwarz inequality and have further that:

$$|\hat{\sigma}^2 - \sigma^2| \leq \frac{1}{n}|U^TU - \vec{\sigma}^T\vec{\sigma}| + \frac{2}{n}\|U\|_\infty\|PU\|_1 + \frac{1}{n}\|PU\|_2^2.$$

Note that we have controlled all the terms on the right hand side already in Lemmas 3.A.7, 3.A.10, 3.A.11, and 3.A.12. That is, these lemmas imply that:

$$\begin{aligned} |\hat{\sigma}^2 - \sigma^2| \ &> \frac{(1+\xi)(s+\tau)\log^{3/2}(2n)\sqrt{32}}{cn^{1/2}} + \frac{ds(s+\tau)(1+\xi)^2\log^2(n)}{n^{1/2}c} \\ &\quad + \frac{(s+\tau)(1+\xi)\log(n)}{n^{1/2}c} r_n + \frac{ds^2(1+\xi)^2\log^2(n)}{nc^2} + r_n \end{aligned}$$

with probability at most

$$\frac{5}{n^{\xi}} \to 0.$$

for arbitrary $\xi > 0$ and absolute constant $c > 0$.

□

3.B A modified neighborhood set which guarantees residuals have conditional mean zero

The question of this subsection is, "for every possible $\pi_0 \in \Pi_{\mathcal{G}}$ and each k which is not in the partial ordering given by π_0 at step t, how can we modify the regression sets $\widehat{N}_{kt}^{\pi_0}$ to guarantee the requirement in Theorem 3.2.2 that $\mathbb{E}\left[R_{kt}^{\pi_0}\middle|X_{\widehat{N}_{kt}^{\pi_0}}\right] = 0$ will hold?" The main results of this section are Algorithm 3 and Theorem 3.B.1. Algorithm 3 tells us how to modify the sets $\widehat{N}_{kt}^{\pi_0}$, while Theorem 3.B.1 tells us why $\mathbb{E}[R_{kt}^{\pi_0}|X_{\widehat{N}_{kt}^{\pi_0}}] = 0$ will hold after applying Algorithm 3.

3.B.1 Algorithm 3 for Appending Regression nodes

Consider that our use of the sets $\widehat{N}_k$ coincides with the definition of an undirected graph where an edge between node j and k exists if either $j \in \widehat{N}_k$ or $k \in \widehat{N}_j$. Because of Assumption 2.2.3, it follows that d-separation queries on this undirected graph imply conditional independence statements. For example, if $\widehat{N}_k \neq [p]\backslash\{k\}$ (the trivial neighborhood set), we have that $X_k \perp\!\!\!\perp X_{[p]\backslash(\widehat{N}_k \cup \{k\})}|X_{\widehat{N}_k}$. For more details, we refer the reader to *Proposition 4.8* in §4.5 of Koller and Friedman (2009) on moral graphs: undirected graphs formed by adding undirected edges between all co-parents in a DAG and removing the orientation from every directed edge. See also Figure 3.B.1 for an example of a moral graph. In our case, due to Assumption 2.2.3, the (undirected) moral graph corresponding to our LiNGAM's underlying DAG contains a subset of the edges in our undirected graph of interest.

To modify $\widehat{N}_{kt}^{\pi_0}$ according to our requirement in Theorem 3.2.2 that $\mathbb{E}\left[R_{kt}^{\pi_0}\middle|X_{\widehat{N}_{kt}^{\pi_0}}\right]=0$, we will make strategic use of first and higher order neighbors to define sequences of $m\geq 3$ nodes $(a_1,\ldots,a_m)$ such that $a_l\in\widehat{N}_{a_{l+1}}$ or $a_{l+1}\in\widehat{N}_{a_l}$ $(1\leq l\leq m-1)$. We will take a_1 to be a sorted node that is not in the set $\widehat{N}_{kt}^{\pi_0}$ already, and a_l for $2\leq l\leq m$ to be unsorted nodes with $a_m=k$. The logic is that the nodes a_l $(1\leq l\leq m-1)$ in Algorithm 3 are candidate ancestors for node k, if any. In particular, a_{m-1} is a prospective parent of k, and we need to account for paths which go through it because it is unsorted. This translates to finding additional sorted nodes a_1 that can be k's ancestors and appending them to the set $\widehat{N}_{kt}^{\pi_0}$. As we demonstrate shortly, appending these nodes to $\widehat{N}_{kt}^{\pi_0}$ guarantees that $R_{kt}^{\pi_0}$ will be independent of $X_{\widehat{N}_{kt}^{\pi_0}}$. In practice, the ancestor candidates a_l $(2\leq l\leq m-1)$ may include node k's descendants in terms of the underlying DAG $\mathcal{G}$, but we must take these ancestor candidates into account because at step $t\geq 3$ we technically only know the partial ordering $\mathcal{A}_t^{\pi_0}=\{\pi_0(j)\}_{j=1}^{t-1}$.

For use in Algorithm 3, consider:

$$L_{kt}^{\pi_0}\doteq\bigcup_{j\in\widehat{N}_{kt}^{\pi_0}}\{j\}\cup AN_j^{\pi_0},$$

which contains each $j\in\widehat{N}_{kt}^{\pi_0}$ and its ancestors, as determined by the partial ordering at steps $\pi_0^{-1}(j)$ and the a priori known neighborhood sets. Note that $\widehat{N}_{kt}^{\pi_0}\subseteq L_{kt}^{\pi_0}\subseteq\mathcal{A}_t^{\pi_0}:=\{\pi_0(j)\}_{j=1}^{t-1}$. The candidate nodes to append to $\widehat{N}_{kt}^{\pi_0}$ will be those in $L_{kt}^{\pi_0}\backslash\widehat{N}_{kt}^{\pi_0}$.

3.B.2 Example Output of Algorithm 3

As an example to what Algorithm 3 outputs, consider Figure 3.B.1 and π_0 given by the natural ordering, $\pi_0(j)=j$ for $j=1,2,\ldots,5$, which is unique in this case (this need not be the case in general):

1. At step $t=1$, the partial ordering is empty, so we need not apply Algorithm 3 since each X_j $(j=1,2,\ldots,5)$ is, without loss of generality, marginally zero-centered.

Algorithm 3: Appending regression nodes to $\widehat{N}_{kt}^{\pi_0}$

Data: $\pi_0 \in \Pi_{\mathcal{G}}$, $t \in [3, p]$, $k \notin \mathcal{A}_t^{\pi_0}$, $\{\widehat{N}_j\}_{j=1}^p$

Result: $\widehat{N}_{kt}^{\pi_0} \cup S_{kt}^{\pi_0}$

#initialize set of nodes to append

$S_{kt}^{\pi_0} \leftarrow \emptyset$

if $|\widehat{N}_{kt}^{\pi_0}| = 0$ **then**

  #no need to append nodes

else

  #decide what nodes to append

  #the candidate ancestors to append

  for $a \in L_{kt}^{\pi_0} \backslash \widehat{N}_{kt}^{\pi_0}$ **do**

    # check whether node k is reachable from node a along paths for which only $a \in \mathcal{A}_t^{\pi_0}$

    if *there exists a sequence of* $m \geq 3$ *nodes* $(a_1, \ldots, a_m)$ *such that*

- $a_1 = a, a_m = k$.
- $a_l \in \widehat{N}_{a_{l+1}}$ *or* $a_{l+1} \in \widehat{N}_{a_l}$ *for* $1 \leq l \leq m-1$.
- $a_l \notin \mathcal{A}_t^{\pi_0}$ *for* $2 \leq l \leq m$.

    then

      $S_{kt}^{\pi_0} \leftarrow S_{kt}^{\pi_0} \cup \{a\}$.

    else

      continue

    end

  end

end

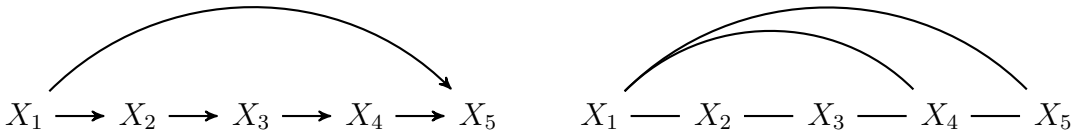

Figure 3.B.1: The original DAG (left) and the undirected graph (right) used for Algorithm 3, which we can arrive at if $\widehat{N}_k = MB_k$, node k's Markov Blanket, for each $k = 1, 2, \ldots, 5$. Compared to the DAG on the left, notice the extra edge between X_1 and X_4, since a markov blanket contains co-parents.

2. At step $t = 2$, the partial ordering is $\{1\}$. The corresponding sets of ordered neighbors are $\widehat{N}_{22}^{\pi_0} = \widehat{N}_{41}^{\pi_0} = \widehat{N}_{52}^{\pi_0} = \{1\}$ and $\widehat{N}_{32}^{\pi_0} = \emptyset$. Because there are no more nodes to append to $\widehat{N}_{22}^{\pi_0}$, $\widehat{N}_{41}^{\pi_0}$, and $\widehat{N}_{52}^{\pi_0}$, we need not apply Algorithm 3 for nodes 2, 4, and 5. For node 3, we also need not run Algorithm 3 because we assume X_3 is marginally zero-centered.

3. At step $t = 3$, the partial ordering is $\{1, 2\}$. The corresponding sets of ordered neighbors are $\widehat{N}_{33}^{\pi_0} = \{2\}$ and $\widehat{N}_{43}^{\pi_0} = \widehat{N}_{53}^{\pi_0} = \{1\}$.

    - In the undirected graph in Figure 3.B.1, there exists a path from the sorted node 1 to the unsorted node 4, a neighbor of node 3. So $S_{33}^{\pi_0} = \{1\}$ in Algorithm 3.
    - $S_{43}^{\pi_0} = \{2\}$ because of the path that exists in the undirected graph from sorted node 2 to unsorted node 3, a neighbor of node 4.
    - Finally, $S_{53}^{\pi_0} = \{2\}$ because there is a path in the undirected graph from sorted node 2, through unsorted node 3, then to the unsorted node 4, a neighbor of node 5.

4. At step $t = 4$, the partial ordering is $\{1, 2, 3\}$. The sets of ordered neighbors are $\widehat{N}_{44}^{\pi_0} = \{1, 3\}$ and $\widehat{N}_{54}^{\pi_0} = \{1\}$.

    - We cannot append node 2 to $\widehat{N}_{44}^{\pi_0}$ in Algorithm 3, because no path from node 2 to an unsorted neighbor of node 4 satisfies the constraint that is must contain only one sorted node.

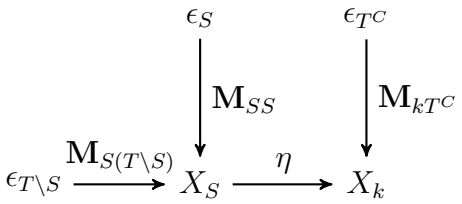

Figure 3.B.2: The DAG that results from Algorithm 3.

- $S_{54}^{\pi_0} = \{3\}$ due to the path from node 3 to node 4 in the undirected graph. We similarly do not add node 2 to $S_{54}^{\pi_0}$ because the path from node 2 to node 4 contains two sorted nodes.

5. At step $t = 5$, the partial ordering is $\{1, 2, 3, 4\}$ and we have only one node left to order, so we need not run Algorithm 3.

3.B.3 Theoretical Grounding of Algorithm 3

Recycling notation, now take $\widehat{N}_{kt}^{\pi_0} = \widehat{N}_{kt}^{\pi_0,old} \cup S_{kt}^{\pi_0}$ as the output of Algorithm 3. In Theorem 3.B.1, we formally show that with our modified regression set $\widehat{N}_{kt}^{\pi_0}$,

$$\mathbb{E}[R_{kt}^{\pi_0} | X_{\widehat{N}_{kt}^{\pi_0}}] = 0.$$

Figure 3.B.2 provides the corresponding DAG with $S = \widehat{N}_{kt}^{\pi_0}$ as outputted by Algorithm 3 and $T = L_{kt}^{\pi_0}$.

Theorem 3.B.1 (X_k is linear in $X_{\widehat{N}_{kt}^{\pi_0}}$ after applying Algorithm 3)**.**
For simplicity of notation, let $S = \widehat{N}_{kt}^{\pi_0}$ as outputted by Algorithm 3 and $T = L_{kt}^{\pi_0}$, which contains nodes in $\widehat{N}_{kt}^{\pi_0}$ and their ancestors. We have that

$$X_k = \eta^T X_S + \mathbf{M}_{kT^C} \epsilon_{T^C},$$

with $\mathbf{M}_{kT^C} \epsilon_{T^C} \perp\!\!\!\perp X_S$ and $R_{kt}^{\pi_0} = \mathbf{M}_{kT^C} \epsilon_{T^C}$. That is, the DAG of Figure 3.B.2 holds with each child being linearly related to its parents.

Proof.

Consider the entry (b, a) of the mixing matrix $\mathbf{M}$. It corresponds to the signal X_b obtains from ϵ_a. Intuitively, this signal is related to all directed paths from node a to node b in the underlying DAG. These directed paths can be defined as sequences of nodes $(a_1, \ldots, a_m)$ with $m \geq 1$ such that:

- $a_1 = a$ and $a_m = b$.
- $a_l \in PA_{a_{l+1}}$ for $1 \leq l \leq m-1$ when $m \geq 2$.

Based on the linear relationship each node has with its parents, we have that:

$$\mathbf{M}_{ba} = \mathbf{1}_{\{a=b\}} + \sum_{\substack{a_1,\ldots,a_m \in [p]: \\ 2 \leq m \leq p, a_1 = a, a_m = b}} \prod_{l=1}^{m-1} \mathbf{B}_{a_l a_{l+1}}. \tag{3.28}$$

Note that in (3.28), we are technically summing across $\sum_{2 \leq m \leq p} p^{m-2}$ combinations of nodes. But because $\mathbf{B}$ is an acyclic weighted adjacency matrix, the terms $\mathbf{B}_{a_l a_{l+1}}$ will conveniently be zero if $a_l \notin PA_{a_{l+1}}$. This implies that $\mathbf{M}_{ba} = 0$ if $a \notin \{b\} \cup AN_b$, $\mathbf{M}_{ba} = 1$ if $b = a$, and overall that $\mathbf{M}_{ba}$ will only be a sum of non-zero terms corresponding to valid directed paths from a to b.

Consider two sets of nodes S and T such that $S \subseteq T \subseteq [p] \backslash \{k\}$ and every directed path from a node $a \in T$ must go through a node $b \in S$ in order to reach k (including the case that $a = b$). This is exactly the case with $S = \widehat{N}_{kt}^{\pi_0}$ and $T = L_{kt}^{\pi_0}$ as a property of Algorithm 3. For any $a \in T$, it follows that all directed paths $(a_1, \ldots, a_m)$ from $a_1 = a$ to $a_m = k$ are such that the index $\lambda = \arg\max\{1 \leq l \leq m-1 : \ a_l \in S\}$ exists: a_λ is the final node in the directed path which was an element of S. An important property of λ is that $a_l \in \{k\} \cup AN_k \backslash \mathcal{A}_t^{\pi_0}$ for $\lambda < l \leq m$ due to our reconstruction of S in Algorithm 3. Incorporating this additional

structure to (3.28), we have:

$$
\begin{aligned}
\mathbf{M}_{ka} &= \sum_{\substack{a_1,\ldots,a_m\in[p]:\\ 2\le m\le p, a_1=a, a_m=k\\ \exists\ \lambda=\arg\max\{1\le l\le m-1:\ a_l\in S\}}} \prod_{l=1}^{m-1}\mathbf{B}_{a_l a_{l+1}} \\
&= \sum_{b\in S}\left[\sum_{\substack{a_1,\ldots,a_m\in[p]:\\ 2\le m\le p, a_1=a, a_m=k\\ \exists\ \lambda=\arg\max\{1\le l\le m-1:\ a_l\in S\}\\ a_\lambda=b}} \prod_{l=1}^{m-1}\mathbf{B}_{a_l a_{l+1}}\right] \\
&= \sum_{b\in S}\left[\left[\mathbf{1}_{\{a=b\}} + \sum_{\substack{a_1,\ldots,a_\lambda\in[p]:\\ 2\le\lambda\le p, a_1=a, a_\lambda=b}} \prod_{l=1}^{\lambda-1}\mathbf{B}_{a_l a_{l+1}}\right] \sum_{\substack{b_1,\ldots,b_\gamma\in[p]:\\ 2\le\gamma\le p, b_1=b, b_\gamma=k\\ b_l\notin\mathcal{A}_t^{\pi_0}\text{ for }2\le l\le\lambda}} \left[\prod_{l=1}^{\gamma-1}\mathbf{B}_{b_l b_{l+1}}\right]\right] \\
&= \sum_{b\in S}\mathbf{M}_{ba} \sum_{\substack{b_1,\ldots,b_\gamma\in[p]:\\ 2\le\gamma\le p, b_1=b, b_\gamma=k\\ b_l\notin\mathcal{A}_t^{\pi_0}\text{ for }2\le l\le\lambda}} \prod_{l=1}^{\gamma-1}\mathbf{B}_{b_l b_{l+1}} \\
&= \eta^T\mathbf{M}_{Sa}.
\end{aligned}
\tag{3.29}
$$

Note that the sum in the first line is across a non-empty set of node sequences, based on $a \in T$ and our discussion around a_λ. The second line is essentially the same as the first, but in the inner sum we are specifying which node $b \in S$ is equal to a_λ, while in the outer sum we are iterating through the various possible $b \in S$. The third line in (3.29) holds because for $\lambda \geq 2$:

$$(a_1, \ldots, a_\lambda, \ldots, a_m) = (a_1, \ldots, b_1, \ldots, b_\gamma).$$

The indicator $\mathbf{1}_{\{a=b\}}$ incorporates the possibility that the path, $(a_1, \ldots, a_m)$, starts at $b \in S$ if $a = b$. The fourth line in (3.29) holds based on an application of (3.28) for $\mathbf{M}_{ba}$. In the fifth line, we take $\eta \in \mathbb{R}^{|S|\times 1}$ such that

$$\eta_j = \sum_{\substack{b_1,\ldots,b_\gamma\in[p]:\\ 2\le\gamma\le p, b_1=(S)_j, b_\gamma=k\\ b_l\notin\mathcal{A}_t^{\pi_0}\text{ for }2\le l\le\lambda}} \prod_{l=1}^{\gamma-1}\mathbf{B}_{b_l b_{l+1}},$$

where $(S)_j$ is the j-th element of the set S.

Note that η does not depend on $a \in T$, so it follows that from (3.29) that:

$$\mathbf{M}_{kT} = \eta \mathbf{M}_{ST}.$$

Consider that:

$$\begin{aligned} X_k &= \mathbf{M}_{kT}\epsilon_T + \mathbf{M}_{kT^C}\epsilon_{T^C} \\ &= \eta \mathbf{M}_{ST}\epsilon_T + \mathbf{M}_{kT^C}\epsilon_{T^C} \quad \text{because } \mathbf{M}_{kT} = \eta \mathbf{M}_{ST} \\ &= \eta X_S + \mathbf{M}_{kT^C}\epsilon_{T^C} \quad \text{because } X_S = \mathbf{M}_{ST}\epsilon_T, \end{aligned} \tag{3.30}$$

which is what we wanted to show. Note that $\mathbf{M}_{kT^C}\epsilon_{T^C} \perp\!\!\!\perp X_S$ because X_S is a deterministic function of ϵ_T and $\epsilon_T \perp\!\!\!\perp \epsilon_{T^C}$ by our LiNGAM assumption. Also, the least squares residual satisfies $R^{\pi_0}_{kt} = \mathbf{M}_{kT^C}\epsilon_{T^C}$ because

$$\eta = \arg\min_\theta \left\{ \mathbb{E}[(\eta^T X_S - \theta^T X_S)^2] + \mathbb{E}[(\mathbf{M}_{kT^C}\epsilon_{T^C})^2] \right\} = \arg\min_\theta \mathbb{E}[(X_k - \theta^T X_S)^2].$$

□

CHAPTER 4

Non-asymptotic Confidence Bands on the Probability an Individual Benefits from Treatment (PIBT)

4.1 Introduction

This chapter presents our work found in the pre-print Ruiz and Padilla (2022).

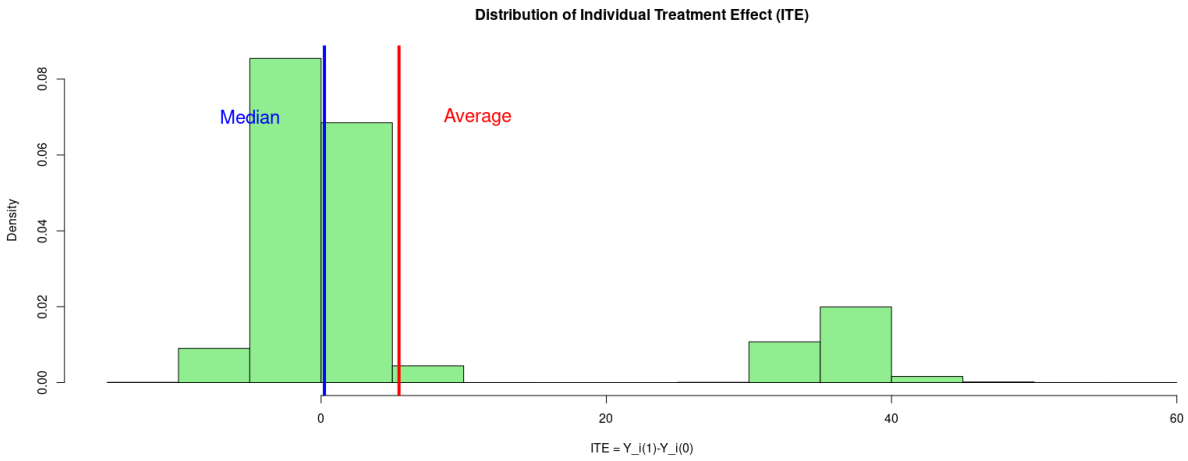

Figure 4.1.1: A hypothetical distribution for the Individual Treatment Effects. Here, the mean is positive yet the probability an individual's treatment outcome is better (larger in value) than their control outcome is approximately 50%.

We are interested in the individual treatment effect (ITE):

$$Y_i(1) - Y_i(0), \ i = 1, 2, \ldots, N,$$

where $Y_i(w)$ for $w = 0, 1$ is known as a potential outcome (Splawa-Neyman et al. 1990, Rubin 1974) and equivalently as a counterfactual (Pearl 2009, Hernán and Robins 2020). Here, $Y_i(w)$ corresponds to the outcome Y_i of interest when a hypothetical experimenter intervenes with nature to force the binary exposure indicator W_i for individual i to be little $w \in \{0, 1\}$, typically through random assignment in an experiment. This intervention can also be denoted by (Pearl 2009)'s "do" operator: $Y_i(w)$ is the outcome we observe under $\text{do}(W_i = w)$.

Suppose a large value of the observed outcome $Y_i = W_i Y_i(1) + [1 - W_i]Y_i(0)$ is "good" for an individual. Assuming we wish to show $W_i = 1$ is effective at accomplishing this, we will consider an individual such that

$$Y_i(1) - Y_i(0) > \delta \tag{4.1}$$

to have benefited from treatment. We can take δ to be any relevant threshold we'd like, such as $\delta = 0$. We may also reverse the inequality if more appropriate. Should the potential outcomes be binary, we may also use the inequality in (4.1) with $\delta = 0$. Moreover, if the outcome of interest is strictly positive, we may alternatively define benefiting from treatment in terms of the ratio of an individual's potential outcomes being above a threshold as we discuss in Section 4.2.4. For simplicity of presentation and without any loss of generality for these alternative definitions which our results can be applied to, we will say an individual benefits from treatment when the inequality in (4.1) holds. Just as well, the vocabular and notational semantics of the inequality in (4.1) can instead be with respect to whether an individual is harmed by an intervention as we discuss in Section 4.4.

As can be appreciated from the hypothetical distribution of the individual treatment effects in Figure 4.1.1, it is very well possible that the average of the individual treatment effect distribution is pulled by outliers and therefore may mislead us. This example motivates our interest in the probability an individual benefits from treatment (PIBT) as provided in Definition 4.1.1.

Definition 4.1.1 (The probability an individual benefits from treatment (PIBT))**.**

To understand heterogeneity in treatment effect for individuals across differing strata of a pre-treatment covariate, X_i, denote

$$\theta(\delta, x) := \Pr\left(Y_i(1) - Y_i(0) > \delta | X_i = x\right)$$

as the conditional probability an individual benefits from treatment in pre-treatment covariate stratum x. To understand the effect of treatment for all individuals regardless of strata, denote

$$\theta(\delta) := \Pr\left(Y_i(1) - Y_i(0) > \delta\right)$$

as the marginal probability an individual benefits from treatment.

Here, X_i is a pre-treatment covariate that is thought to deconfound variability in the observed outcome Y_i that is not due to W_i and exogenous noise alone (see Assumption 4.1.7 and Example 4.1.8) (Rubin 1974, Imbens and Rubin 2015). One can see that $\theta(\delta) = \mathbb{E}\left[\theta(X_i, \delta)\right]$, where the expectation is taken with respect to the confounder X_i. Consider now two very similar quantities in Definition 4.1.2.

Definition 4.1.2 (Probability a treatment outcome is better than an independently drawn control outcome)**.**
To understand heterogeneity in treatment's effect across differing individuals and across differing strata of a pre-treatment covariate, denote

$$\eta(\delta, x) := \Pr\left(Y_i(1) - Y_j(0) > \delta | X_i = X_j = x; i \neq j\right)$$

as the probability that a randomly selected individual in pre-treatment covariate stratum x has a treatment potential outcome that is better than the control potential outcome of a differing randomly selected individual also in stratum x. To understand treatment's effect across differing individuals overall, denote

$$\eta(\delta) := \Pr\left(Y_i(1) - Y_j(0) > \delta; i \neq j\right)$$

as the overall probability that a randomly selected individual has a treatment outcome that is better than a differing randomly selected individual's control outcome.

Of crucial importance, $\theta(\delta, x)$ and $\theta(\delta)$ are not in general the same as $\eta(\delta, x)$ and $\eta(\delta)$ in Definition 4.1.2 (Hand 1992, Greenland et al. 2020). This is because $Y_i(1)$ and $Y_i(0)$ are in general *dependent* random variables, while $Y_i(1)$ is independent of $Y_j(0)$ when $i \neq j$ under the standard Stable Unit Treatment Value Assumption (see Assumption 4.1.3). Under appropriate identifiability assumptions, such as Assumption 4.1.5 or Assumption 4.1.7 below, $\eta(\delta, x)$ and $\eta(\delta)$ can be identified because we can generally sample from the distributions $Y_i|W_i = w, X_i = x$ and $Y_i|W_i = w$ for $w = 0, 1$. It is impossible to sample from the joint distributions of $(Y_i(0), Y_i(1))$ marginally or given $X_i = x$, because an individual cannot be in both the treatment group and the control group simultaneously. So we cannot generally identify $\theta(\delta, x)$ nor $\theta(\delta)$. Not even if we are able to perfectly match individuals in opposite treatment groups based on pre-treatment covariates. This is what is known as the fundamental problem of causal inference.

The goal nonetheless is to reason about $\theta(\delta, x)$ and $\theta(\delta)$ through estimated bounds on these quantities. Building from the work of Fan and Park (2010), our focus is on deriving closed-form, non-asymptotic margins of error on the estimated bounds for an overall confidence band on PIBT of the form: PIBT is contained between

$$[\text{estimator for lower bound on PIBT}] - [\text{margin of error for PIBT bounds}] \tag{4.2}$$

and

$$[\text{estimator for upper bound on PIBT}] + [\text{margin of error for PIBT bounds}] \tag{4.3}$$

with some target frequentist confidence level. This interpretation with respect to PIBT is motivated by that of Fay et al. (2018) for the bootstrap confidence intervals they calculate using the PIBT bound estimators of Fan and Park (2010) in the randomized experiment setting.

4.1.0.1 Overview of our contributions

The contributions of the present work are as follows.

1. For the bounds on the marginal probability an individual benefits from treatment which are estimated with data from a randomized experiment (RE), we derive a closed-form concentration inequality depending on only the sample size and the desired frequentist confidence level. As discussed in Section 4.2.3, this allows for a formal statistical power analysis, albeit conservative, but notably without the requirement of an asymptotic limiting distribution nor the specification of any unknown parameters (e.g. plausible effect sizes). Different from the non-asymptotic margin of error that can be obtained with bootstrap re-sampling (Efron and Tibshirani 1994, Bickel et al. 1997), our non-asymptotic margin of error will be closed-form and simultaneous for all thresholds δ that can be used to define PIBT, thus allowing for a form of sensitivity analysis on its definition.

2. Making strategic use of regression residuals, we also discuss how to estimate, possibly in an observational setting, the PIBT conditional on strata of an individual's pre-treatment co-variates. We accompany the proposed approach to study heterogeneity in PIBT with a simple but general theorem that suggests how to extend or obviate from this approach with regression residuals. For the approach with regression residuals, we provide tailored versions of the general statement that allow for a frequentist confidence interpretation simultaneously at all pre-treatment covariate strata. In Section 4.3.2.1, we provide an extended discussion of the application of this result to the canonical linear regression model.

3. We include in Section 4.4 an extended discussion on the scope of our results. For binary potential outcomes, we show in Proposition 4.4.1 that the general approach we take to bound PIBT is equivalent to using the sharp Boole-Fréchet bounds (Rüschendorf 1981).

4. We include in Section 4.5 an example application to a real-life randomized experiment dataset, Criteo AI Lab's benchmark data for uplift prediction (Diemert Eustache, Betlei Artem et al. 2018). In particular, this section points toward a useful combination

of conditional average treatment effect (CATE) estimation and inferring Individual Treatment Effects: through a partitioning of individuals in a sample based on their similar CATE prediction, we can estimate PIBT in each of these strata to better understand the implication of the CATE estimate. This is related to recent work on more interpretable causal analysis given by stratifying treatment effects on an informative univariate score, such as a prognostic or propensity score (Abadie et al. 2018, Padilla et al. 2021, Ye et al. 2021b, Yadlowsky et al. 2021, Xu and Yadlowsky 2022).

The existing mathematical results we exploit include the following. Key to establishing the population-level target bounds on PIBT, we use the Makarov bounds first introduced in Makarov (1982) and later generalized in (Frank et al. 1987, Williamson and Downs 1990). These works establish a distribution-free bound on the cumulative distribution function (CDF) on the sum (or difference or product) of two or more random variables having any unknown joint distribution and fixed marginal distributions. For the non-asymptotic concentration results (the margin of error derivations), the novel contribution of this chapter, we use the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality (Dvoretzky et al. 1956, Massart 1990, Naaman 2021). The DKW inequality gives a non-parametric, non-asymptotic deviation inequality for the supremum difference at any evaluation point between a target CDF and its empirical analogue estimated with a sample of independent and identically distributed (i.i.d.) random variables. Importantly, Massart (1990), Naaman (2021) show that this inequality is tight under no additional assumptions.

4.1.1 Existing work

4.1.1.1 Bounding the Distribution of Individual Treatment Effects

Statistical inference on the CDF of the ITE distribution, $1 - \theta(\delta)$, has been of interest before the present work. In particular, we are decidedly not unique in applying Makarov (1982)'s

bounds to study the ITE's CDF. Fan and Park (2010) estimate bounds on $1-\theta(\delta)$ under a randomized experiment (RE) setting in which the distributions of $Y_i(1)$ and $Y_i(0)$ are each marginally identified. These authors present a straightforward plug-in estimation approach that uses the empirical marginal CDFs to estimate the lower bound and upper bound on $1-\theta(\delta)$ in practice. We make use of this same plug-in estimation approach for the RE setting as well.

The contribution of this work, relative to the contribution of Fan and Park (2010) in the RE setting, is the concentration inequality for the estimated PIBT bounds. Under regularity conditions, the authors show asymptotically that the plug-in bound estimators follow either a normal distribution (centered at the target bound), or a truncated normal distribution, or a point mass. Exactly which distribution this is depends on the supremum difference between the two potential outcomes' CDFs, which is unkown. Even if we knew that the asymptotic distribution of the estimator is Gaussian, a prospective power analysis further requires an estimator for the standard error (not provided in their work) to guarantee a target confidence level and margin of error (e.g. a maximum deviation of 0.05). This points to a strength of the main concentration result in this setting: despite the possibility that the plug-in estimator can have a non-trivial, possibly biased, sampling distribution in a finite sample, the confidence level we can have for a target margin of error depends only on sample size (see the discussion around Theorem 4.2.1 and Figure 4.2.1 for details).

Fay et al. (2018) further discuss the statistical inference technique of Fan and Park (2010) in conjunction with the quantity $1-\eta(\delta)$ in Definition 4.1.2. They discuss how $1-\eta(\delta)$ (plus a tie correction term allowing for discrete outcomes) is related to the Wilcoxon-Mann-Whitney U test (Wilcoxon 1945, Mann and Whitney 1947), often used as a non-parametric alternative to the 2 sample t-test (Fay and Proschan 2010). Through extensive synthetic examples and an application to study vaccine efficacy, these authors demonstrate that $\eta(\delta)$ and $\theta(\delta)$ can be related, for example should certain parametric models hold. But they caution that one must remain very wary of Hand (1992)'s paradox: $\theta(\delta) < 1/2$ can be the case, i.e. treatment

is ineffective for a majority of individuals, yet $\eta(\delta) > 1/2$ may lead us to believe otherwise. Similarly, $\eta(\delta) < 1/2$ and $\theta(\delta) > 1/2$ can instead be the case. Relatedly, Greenland et al. (2020) cautions about the danger of conflating $\eta(\delta)$ with $\theta(\delta)$.

Interestingly, it has been established that the Makarov bounds for the marginal CDF of $Y_i(1) - Y_i(0)$ studied in (Fan and Park 2010) are point-wise but not uniformly sharp (Firpo and Ridder 2010, 2019). This means that Makarov (1982)'s bounds, evaluated at a point $\delta \in \text{support}(Y_i(1) - Y_i(0))$, arise from a joint distribution on the support of $(Y_i(0), Y_i(1))$ which itself may not satisfy the constraint of (Makarov 1982) with respect to the fixed CDFs of one outcome. These authors then show how one can tighten the *population-level* Makarov bounds on the marginal ITE CDF evaluated at only a finite set of points $\delta_1, \ldots, \delta_m \in \text{support}(Y_i(1) - Y_i(0))$ $(m \geq 2)$. While promising, we consider the estimation of these tightened bounds for $\delta_1, \ldots, \delta_m$ beyond the scope of this chapter as it is not immediately clear that it is amenable to our analysis.

For continuous outcomes in a randomized experiment setting, Frandsen and Lefgren (2021) works under a condition known as mutual stochastic increasing-ness of the potential outcomes $(Y_i(0), Y_i(1))$ (Lehmann 1966). The authors write that this condition, a more general way to define positive correlation, "means that individuals with higher potential outcomes in one treatment state draw from a more favorable–in the first order stochastic sense–conditional distribution of outcomes in the other state." The plug-in estimation approach we use in the RE case does not make the assumption of positive correlation: it works for any joint distribution on $(Y_i(0), Y_i(1))$ (Fan and Park 2010), including those with any type of negative association. Given that their numerical results suggest greater precision in the point estimates of the bounds on (one minus) PIBT compared to Fan and Park (2010)'s approach, an interesting avenue for further extensions of the power analysis we propose here is to incorporate their estimation approach for greater precision in settings where we believe positive correlation between potential outcomes is justified.

Also in the context of a randomized experiment, Caughey et al. (2021) study PIBT under

a randomization inference setup that is traditionally used to test the sharp null hypothesis that all individual treatment effects are constant (Fisher 1935). The authors extend this framework to test whether individual treatment effects are bounded, and they also present a strategic use of order statistics to reason about PIBT. The approach we take to bound PIBT assumes the existence of an infinite super-population that subjects in our sample at hand are drawn i.i.d. from and for which our plug-in estimators provide inference for. Caughey et al. (2021) appears to be a nice alternative under the differing assumption that randomness is solely due to random assignment of subjects to a treatment.

Of special note, the quantity $\theta(\delta)$ in Definition 4.1.1 when $0 \leq \delta < 1$ is equivalent to what is known as the "probability of necessity and sufficiency (PNS)" when $Y_i(0)$ and $Y_i(1)$ take on binary values (Pearl 1999, Tian and Pearl 2000, Pearl 2009). In this case, PNS and what we call "marginal PIBT" are given by the joint probability $\Pr(Y_i(1) = 1, Y_i(0) = 0)$. As suggested by the intriguing use of prepositional logic terminology in its name, PNS informs us of an intervention's effectiveness at achieving a strictly better outcome. Bounding and estimation approaches different from the approach taken by Fan and Park (2010) (our focus) are provided in Pearl (1999), Tian and Pearl (2000). In particular, both experimental and observational data can be used to bound PNS. See also Cinelli and Pearl (2021) for a recent discussion on PNS and related quantities that inform about treatment's efficacy. We discuss more on PIBT for binary potential outcomes in Section 4.4's Proposition 4.4.1.

Related to the study of PIBT, Makarov's bounds can also be used to obtain sharp bounds on the quantiles of the marginal distribution of the individual treatment effects in randomized experiments (Fan and Park 2010, 2012). In particular, Fan and Park (2012) discusses the statistical inference on the estimators of these sharp bounds. Also related to reasoning about individual treatment effects, Ding et al. (2019) study bounds on the variance of the ITE. These bounds can also be conditional on pre-treatment covariates, which helps us understand whether covariates help explain away the original variation in ITE. An interesting avenue of future work may be to relate ATE (or CATE) with the ITE distribution via a Chebyshev's

inequality using the results in Ding et al. (2019).

4.1.1.2 PIBT conditional on pre-treatment covariates

Fan and Park (2010) also discuss the conditional bounds for the conditional PIBT, $1-\theta(\delta, x)$, at the population-level along with a brief discussion of possible estimation approaches. The appendix of Frandsen and Lefgren (2021) also discuss a generalization of the mutual stochastic increasing-ness assumption in order to arrive at bounds for $1-\theta(\delta, x)$. In the context of ordinal outcomes, using Makarov's bounds to study the ITE, Lu et al. (2015) also consider the case we would like to condition on covariates. All three suggest some form of distributional regression, the semi-parametric estimation of a conditional CDF (Koenker et al. 2013, Chernozhukov et al. 2013, Kneib et al. 2021). We extend their discussion on covariate conditioning with a discussion on theoretical guarantees and how to conduct statistical inference with the bound estimators.

Related to our use of pre-treatment covariates, Lei and Candès (2021) develops prediction intervals for the individual treatment effect based on quantile regression (Koenker and Bassett 1978) with strategic calibration using conformal inference (Vovk et al. 2005, Shafer and Vovk 2008, Tibshirani et al. 2019). Moreover, this work is extended to scenarios where unobserved confounding is possible (Yin et al. 2021, Jin et al. 2021). Our work here is complementary to these advances, in analogy to the inverse relation between quantiles and the CDF of a distribution. With respect to theoretical guarantees, Theorem 4.3.4 below is with respect to the supremum deviation across inputted pre-treatment covariate levels, whereas the guarantees of Lei and Candès (2021) are with respect to any single randomly generated covariate level. This is a subtle but important difference: one may like inference about heterogeneity in individual treatments effects to extend simultaneously to multiple individuals with fixed (non-random) covariate levels, not necessarily a single random individual. On the other hand, our work does not necessarily extend to a target population beyond that represented by our training sample, and one of the main results here (Theorem 4.3.4) makes use of a regularity

condition on regression residuals that quantile regression generally avoids.

4.1.2 Assumptions

Throughout this work, we will assume the stable unit treatment value assumption (SUTVA) in Assumption 4.1.3.

Assumption 4.1.3 (Stable Unit Treatment Value Assumption (SUTVA))**.**
We quote Imbens and Rubin (2015): "The potential outcomes for any unit do not vary with the treatments assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes."

We will also assume consistency of the observed outcome throughout as given in Assumption 4.1.4.

Assumption 4.1.4 (Consistency)**.**
The observed outcome is dictated by treatment receipt indicator W_i:

$$Y_i := W_i Y_i(1) + [1 - W_i] Y_i(0).$$

Depending on the setting in which we would like to bound PIBT, we will also work with differing identification assumptions which are discussed now. Assumption 4.1.5 and 4.1.7 are in line with the Neyman-Rubin potential outcome model. To demonstrate how these assumptions can come up, we provide Examples 4.1.6 and 4.1.8 which make use of Pearl (2009)'s structural causal model (SCM).

4.1.2.1 The randomized experiment case

Assumption 4.1.5 (Strong Ignorability (Rubin 1974, Imbens and Rubin 2015))**.**
For $i = 1, 2, \ldots, n$, assume $(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i$ and $\Pr(W_i = 1)$ is bounded away from 0 and 1.

Example 4.1.6 next gives a situation in which the marginal independence component of Assumption 4.1.5 holds: noting that $Y_i(w) := h_Y(w, X_i, \epsilon_{iY})$ for non-random $w = 0, 1$

and that $(X_i, \epsilon_{iY}) \perp\!\!\!\perp W_i$, it follows that $(Y_i(0), Y_i(1))$ is independent of random treatment assignment indicator W_i.

Example 4.1.6.

Let h_X, h_W, and h_Y be fixed functions. For $i = 1, 2, \ldots, n$, assume the random variables of interest are generated i.i.d. according to:

$$\begin{cases} X_i &= h_X(\epsilon_{iX}) \\ W_i &= h_W(\epsilon_{iW}) \\ Y_i &= h_Y(W_i, X_i, \epsilon_{iY}) \end{cases} .$$

Here, $(\epsilon_{iX}, \epsilon_{iW}, \epsilon_{iY})$ are the latent causes for variation in (X_i, W_i, Y_i), and they are mutually independent.

4.1.2.2 The pre-treatment covariate adjusted case

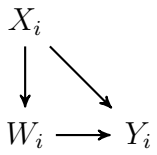

Figure 4.1.2: A Directed Acyclic Graph (DAG) with common cause X_i between W_i and Y_i.

Assumption 4.1.7 (Strong Conditional Ignorability (Rubin 1974, Imbens and Rubin 2015))**.**

For $i = 1, 2, \ldots, n$, assume $(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i | X_i$ and $\Pr(W_i = 1 | X_i)$ is bounded away from 0 and 1 almost surely.

Example 4.1.8 which goes along with Figure 4.1.2 gives an example in which the conditional independence component of Assumption 4.1.7 holds. Conditional on $X_i = x$, a non-random value, we have that $Y_i(w) = h_Y(w, x, \epsilon_{iY})$ for non-random $w = 0, 1$. From this and the

fact that $\epsilon_{iY} \perp\!\!\!\perp W_i$, it follows that $(Y_i(0), Y_i(1))$ is independent of random W_i conditional on $X_i = x$, for any value of x.

Example 4.1.8.

Let h_X, h_W, and h_Y be fixed functions. For $i = 1, 2, \ldots, n$, assume the random variables of interest are generated i.i.d. according to:

$$\begin{cases} X_i & = h_X(\epsilon_{iX}) \\ W_i & = h_W(X_i, \epsilon_{iW}) \\ Y_i & = h_Y(W_i, X_i, \epsilon_{iY}) \end{cases} .$$

Here, $(\epsilon_{iX}, \epsilon_{iW}, \epsilon_{iY})$ are the latent causes for variation in (X_i, W_i, Y_i), and they are mutually independent.

4.2 PIBT bounds in a randomized experiment

Here, we aim to estimate bounds on $\theta(\delta) = \Pr(Y_i(1) - Y_i(0) > \delta)$, the marginal PIBT in Definition 4.1.1. In order for us to identify these bounds, we will work under Assumption 4.1.5. Denote Makarov (1982)'s lower bound and upper bound on PIBT as $\theta^L(\delta)$ and $\theta^U(\delta)$, which are such that

$$\theta^L(\delta) \ \leq \ \theta(\delta) = \Pr(Y_i(1) - Y_i(0) > \delta) \ \leq \ \theta^U(\delta).$$

Denote their corresponding estimators based on i.i.d. data as $\hat{\theta}^L(\delta)$ and $\hat{\theta}^U(\delta)$, respectively. Theorem 4.2.1 is the main result in this section, providing a guarantee for the accuracy of these estimators, which we now formally define.

4.2.1 The target bounds on PIBT and their estimators

We refer the reader to Lemma 4.A.1 in Appendix 4.A for the formal statement of the Makarov bounds. The target parameters $\theta^L(\delta)$ and $\theta^U(\delta)$ are in terms of the potential outcomes'

marginal CDFs. Here, the marginal CDF of $Y_i(w)$ is:

$$F_w(y) := \Pr\left(Y_i(w) \leq y\right),$$

which is identified under Assumption 4.1.5. That is, $F_w(y) = \Pr\left(Y_i \leq y | W_i = w\right)$, the marginal CDF of the observed outcomes in treatment group $w = 0, 1$. Denote the empirical cumulative distribution function (eCDF) for $Y_i(w)$, a natural estimator for $F_w(y)$, as:

$$\hat{F}_{wn}(y) := \frac{1}{n_w} \sum_{i:\ W_i = w} \mathbf{1}\left\{Y_i \leq y\right\} \text{ for } w = 0, 1. \tag{4.4}$$

Here, $\mathbf{1}\{\cdot\}$ is the indicator function. Using Lemma 4.A.1 in Appendix 4.A, the target parameters to bound $\theta(\delta)$ across any joint distribution of $(Y_i(0), Y_i(1))$ are:

$$\theta^L(\delta) \ = \ -\min\left(\inf_y \left\{F_1(y + \delta/2) - F_0(y - \delta/2)\right\}, 0\right)$$

for the lower bound, while for the upper bound we have:

$$\theta^U(\delta) \ = 1 - \max\left(\sup_y \left\{F_1(y + \delta/2) - F_0(y - \delta/2)\right\}, 0\right).$$

Correspondingly, we can obtain the bound estimators by plugging in the CDF estimators as in Fan and Park (2010):

$$\hat{\theta}^L(\delta) := -\min\left(\inf_y \left[\hat{F}_{1n}(y + \delta/2) - \hat{F}_{0n}(y - \delta/2)\right], 0\right)$$

and

$$\hat{\theta}^U(\delta) := 1 - \max\left(\sup_y \left[\hat{F}_{1n}(y + \delta/2) - \hat{F}_{0n}(y - \delta/2)\right], 0\right)$$

for the lower bound and upper bound, respectively.

4.2.2 The main result in the RE setting

Given the choice of our estimators $(\hat{\theta}^L(\delta), \hat{\theta}^U(\delta))$, the question becomes how accurate they are for a given sample size n. Theorem 4.2.1 provides us with this understanding.

Theorem 4.2.1 (Concentration inequality for the bounds on PIBT in an RE)**.** *If* $(Y_1(0), Y_1(1)), \ldots, (Y_n(0), Y_n(1))$ *are i.i.d. and Assumption 4.1.5 (strong ignorability) holds, then for any* $\alpha \in (0, 1)$*, we have that:*

$$Pr\left(\sup_{\delta}\left\{\left|\hat{\theta}^L(\delta) - \theta^L(\delta)\right| \vee \left|\hat{\theta}^U(\delta) - \theta^U(\delta)\right|\right\} \leq \left(\frac{\log(4/\alpha)}{2}\right)^{\frac{1}{2}}\left(n_0^{-\frac{1}{2}} + n_1^{-\frac{1}{2}}\right)\right) \;\geq\; 1-\alpha.$$

In Theorem 4.2.1, "$\vee$" is the maximum between the left and right arguments. Moreover, n_0 is the number of control group units, while n_1 is the number of treatment group units. The proof of Theorem 4.2.1 is contained in Appendix 4.B.2. The core idea is to first show that the bound estimators' joint deviation can be understood in terms of the deviation between the potential outcome's CDFs and eCDFs uniformly across each $y \in \text{support}(Y_i)$ and $w = 0, 1$. That is, we must show that it is sufficient to bound:

$$\max_{w=0,1}\sup_{y}\left|\hat{F}_{wn}(y) - F_w(y)\right|$$

with high probability. Conveniently, this second part of the proof to Theorem 4.2.1 is given by the DKW inequality (Dvoretzky et al. 1956, Massart 1990, Naaman 2021) under the mild assumption that Y_i for each i such that $W_i = w$ must be i.i.d..

In Remark 4.2.2, we can see the practical implication of Theorem 4.3.4. The validity of Remark 4.2.2 follows from the definition of the target parameters, $\theta^L(\delta)$ and $\theta^U(\delta)$, and the derived margin of error for their estimation. In Section 4.2.3, we further discuss its practical implication with respect to a statistical power analysis.

Remark 4.2.2. Using Theorem 4.2.1, we can say that with confidence at least $(1-\alpha)\times 100\%$, the probability an individual represented by our randomized experiment will benefit from treatment,

$$\theta(\delta) = \Pr\left(Y_i(1) - Y_i(0) > \delta\right),$$

for any threshold δ of interest, is between

$$\hat{\theta}^L(\delta) - \left(\frac{\log(4/\alpha)}{2}\right)^{\frac{1}{2}}\left(n_0^{-\frac{1}{2}} + n_1^{-\frac{1}{2}}\right)$$

and

$$\hat{\theta}^U(\delta) + \left(\frac{\log(4/\alpha)}{2}\right)^{\frac{1}{2}} \left(n_0^{-\frac{1}{2}} + n_1^{-\frac{1}{2}}\right).$$

4.2.3 A power analysis with Theorem 4.2.1

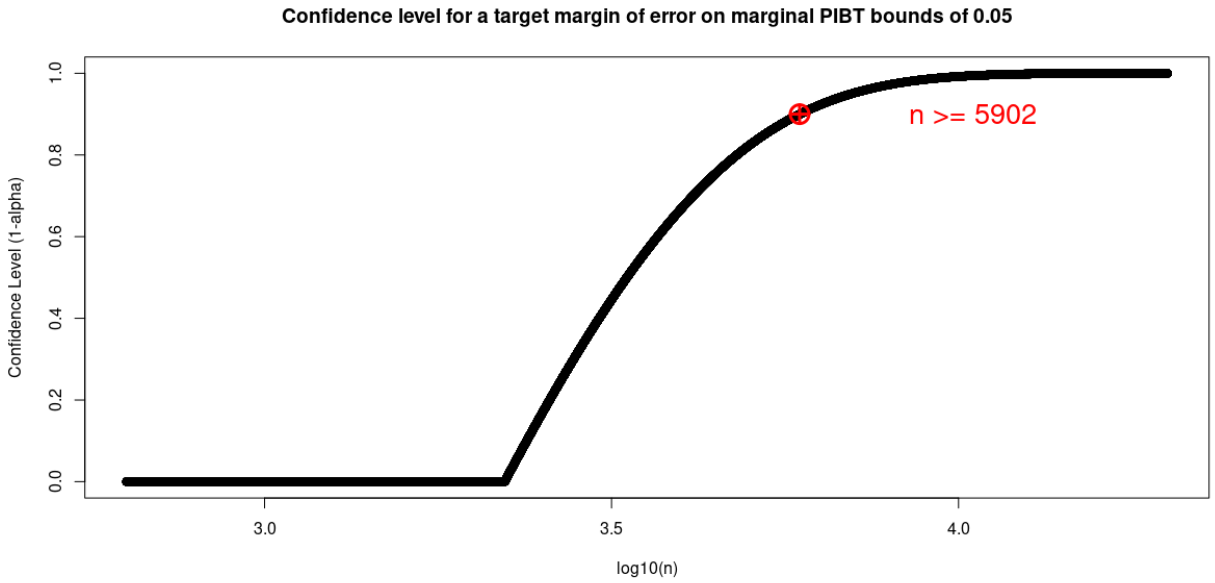

Figure 4.2.1: An example power analysis based on Theorem 4.2.1. Here, $n_0 = n_1$ and $n = n_0 + n_1$. The target margin of error is 0.05, while the target confidence level is 90%. A sample size of $n \geq 5902$ guarantees this margin of error with at least a 90% confidence level.

Figure 4.2.1 presents an example power analysis making use of Theorem 4.2.1 for the case that $n_0 = n_1$ and a target margin of error of 0.05. In general, suppose our target margin of error for

$$\sup_{\delta} \left\{ \left| \hat{\theta}^L(\delta) - \theta^L(\delta) \right| \vee \left| \hat{\theta}^U(\delta) - \theta^U(\delta) \right| \right\}$$

is $\varepsilon \in (0, 1)$. Solving for the significance level α_ε when we set ε equal to the margin of error in Theorem 4.2.1:

$$\varepsilon \doteq \left(\frac{\log(4/\alpha_\varepsilon)}{2}\right)^{\frac{1}{2}} \left(n_0^{-\frac{1}{2}} + n_1^{-\frac{1}{2}}\right)$$

gives:

$$\alpha_\varepsilon = 4\exp\left(-2\left(n_0^{-\frac{1}{2}} + n_1^{-\frac{1}{2}}\right)^{-2}\varepsilon^2\right).$$

The confidence level we can thus have for the target margin of error ε at any given sample size (n_0, n_1) is at least:

$$\begin{cases} 1-\alpha_\varepsilon & \text{if } 0 \leq \alpha_\varepsilon \leq 1 \\ 0 & \text{otherwise} \end{cases}.$$

4.2.4 Differing definition of benefiting from treatment in terms of the ratio of potential outcomes

We also have the following simple extension to ratios of potential outcomes. It is motivated by *Theorem 2* of Williamson and Downs (1990), which derives bounds for the CDF of a sum, difference, product, or ratio of two random variables with an unknown joint distribution.

Suppose that we are interested in strictly positive potential outcomes $\tilde{Y}_i(0)$ and $\tilde{Y}_i(1)$. For example, this can be in a setting where the time to an event is the outcome of interest (Cox 1972, Stitelman and van der Laan 2010, Austin 2014, Schober and Vetter 2018, Cai and van der Laan 2020). Using a threshold $\tilde{\delta} > 0$ (e.g. $\tilde{\delta} = 1$), one may alternatively consider an individual to have benefited from treatment should the inequality

$$\tilde{Y}_i(1)/\tilde{Y}_i(0) > \tilde{\delta}$$

occur. That is, individual i is deemed to have benefited from treatment should their treatment outcome be larger than their control outcome by a factor larger than $\tilde{\delta}$. Correspondingly, one may be interested in bounding the unidentifiable probability

$$\tilde{\theta}(\tilde{\delta}) := \Pr\left(\tilde{Y}_i(1)/\tilde{Y}_i(0) > \tilde{\delta}\right).$$

To do so, one can work with the variable transformation

$$Y_i(w) \doteq \log(\tilde{Y}_i(w)); w = 0, 1.$$

Given our definition of $\theta(\delta)$ in Definition 4.1.1 and the one-to-one nature of the log transformation, we get that

$$\tilde{\theta}(\tilde{\delta}) = \theta(\delta)$$

when we set $\delta \doteq \log(\tilde{\delta})$. It follows that:

$$\theta^L(\delta) \leq \tilde{\theta}(\tilde{\delta}) \leq \theta^U(\delta).$$

We will simply have to work with the eCDF of $\log(\tilde{Y}_i(w))$ when obtaining the estimators $\hat{\theta}^L(\delta)$ and $\hat{\theta}^U(\delta)$. Moreover, Theorem 4.2.1 here is still useful to conduct inference on $\tilde{\theta}(\tilde{\delta})$ under the i.i.d. assumption.

4.3 PIBT bounds with pre-treatment covariates

We now seek to estimate bounds on the unidentifiable probability an individual benefits from treatment (PIBT) in pre-treatment stratum $X_i = x$:

$$\theta(\delta, x) = \Pr\left(\Delta_i > \delta | X_i = x\right).$$

Could it be known, this quantity is helpful to understand whether the benefit of receiving treatment varies across pre-treatment covariate strata. Denote the target bounds as $\theta^L(\delta, x)$ and $\theta^U(\delta, x)$, the lower and upper bound, respectively. They satisfy:

$$\theta^L(\delta, x) \leq \theta(\delta, x) \leq \theta^U(\delta, x).$$

And denote the corresponding estimators as $\hat{\theta}^L(\delta, x)$ and $\hat{\theta}^U(\delta, x)$, respectively. We would like a guarantee about how close $(\hat{\theta}^L(\delta, x), \hat{\theta}^U(\delta, x))$ is to $(\theta^L(\delta, x), \theta^U(\delta, x))$.

Remark 4.3.1 (Large enough sample at a covariate stratum?)**.**
Importantly, we note that should a large enough sample be collected at stratum x of the pre-treatment covariates, Theorem 4.2.1 can be applied for a frequentist confidence statement about $\theta(\delta, x)$ using the analogous interpretation in Remark 4.2.2. The rest of this section

is useful for the case that a large enough sample is not collected for some or all of the pre-treatment covariate strata of interest.

Theorem 4.3.2 is the main non-asymptotic, non-parametric result for this setting. Theorem 4.3.4 is the adaptation of Theorem 4.3.2 to a case where we strategically use regression residuals to estimate the bounds on PIBT. For this approach with regression residuals, we show how a confidence statement about the conditional PIBT bound estimators can be written in terms of a target confidence level that is adjusted according to how accurate the regression function estimator is. In Corollary 4.3.3, we demonstrate how the statement written in this manner implies that the conditional bound estimators are as statistically efficient as the regression function estimator of choice. Moreover, Proposition 4.3.7 adopts the more general Theorem 4.3.4 to the canonical linear regression case. We demonstrate how to use this result to conduct a power analysis for the simultaneous inference on PIBT at all pre-treatment covariate strata.

4.3.1 The target bounds on conditional PIBT, their estimators, and the main result

The bounds $\theta^L(\delta, x)$ and $\theta^U(\delta, x)$ make use of the conditional CDFs ($w = 0, 1$):

$$F_w(y|x) := \Pr\left(Y_i(w) \leq y | X_i = x\right) = \Pr\left(Y_i \leq y | X_i = x, W_i = w\right),$$

with the second equality due to Assumption 4.1.7 and consistency. Explicitly, due to Lemma 4.A.1 in Appendix 4.A, we have:

$$\theta^L(\delta, x) = -\min\left(\inf_y \left\{F_1(y + \delta/2|x) - F_0(y - \delta/2|x)\right\}, 0\right)$$

along with

$$\theta^U(\delta, x) = 1 - \max\left(\sup_y \left\{F_1(y + \delta/2|x) - F_0(y - \delta/2|x)\right\}, 0\right)$$

at the population-level. Denote

$$G(y, \delta, x) := F_1(y + \delta/2|x) - F_0(y - \delta/2|x),$$

and its corresponding estimator as $\hat{G}_n(y, \delta, x)$ based on the training sample. In practice, one may specify $\hat{G}_n(y, \delta, x)$ as the difference of two conditional CDF estimators as in Corollary 4.3.3 below. The plug-in estimators for the lower bound and upper bounds, respectively, will be:

$$\hat{\theta}^L(\delta, x) = -\min\left(\inf_y \left\{\hat{G}(y, \delta, x)\right\}, 0\right) \tag{4.5}$$

along with

$$\hat{\theta}^U(\delta, x) = 1 - \max\left(\sup_y \left\{\hat{G}_n(y, \delta, x)\right\}, 0\right). \tag{4.6}$$

With respect to this choice of $(\hat{\theta}^L(\cdot), \hat{\theta}^U(\cdot))$, Theorem 4.3.2 is a result under the most general conditions. Corollary 4.3.3 and Theorem 4.3.4 give further concreteness for how exactly to guarantee the premise of Theorem 4.3.2 with respect to $\hat{G}(y, \delta, x)$. The idea behind the generic statement in Theorem 4.3.2 is to encourage extensions, especially those with the possibility of being more statistically efficient, with less restricted conditions than those in Theorem 4.3.4, or with modeling assumptions that are tailored to the application at hand.

Theorem 4.3.2 (A non-parametric inequality about the conditional bound estimators' deviation)**.**
If Assumption 4.1.7 (strong conditional ignorability) holds, we have for all $\delta \in \mathbb{R}$ and all $x \in$ support(X_i):

$$\left(\left|\hat{\theta}^L(\delta, x) - \theta^L(\delta, x)\right| \vee \left|\hat{\theta}^U(\delta, x) - \theta^U(\delta, x)\right|\right) \leq \sup_y \left|\hat{G}_n(y, \delta, x) - G(y, \delta, x)\right|. \tag{4.7}$$

If, additionally, $\hat{G}_n(y, \delta, x)$ is such that there exists a value $t_\alpha(\delta, x)$ such that:

$$Pr\left(\sup_y \left|\hat{G}_n(y, \delta, x) - G(y, \delta, x)\right| \leq t_\alpha(\delta, x)\right) \geq 1 - \alpha,$$

then we have that:

$$Pr\left(\left|\hat{\theta}^L(\delta, x) - \theta^L(\delta, x)\right| \vee \left|\hat{\theta}^U(\delta, x) - \theta^U(\delta, x)\right| \leq t_\alpha(\delta, x)\right) \geq 1 - \alpha.$$

The proof of Theorem 4.3.2 is contained in Appendix 4.B.3. The implications of the non-parametric, deterministic inequality in (4.7) are interesting and perhaps a bit surprising. This

inequality is stating that in a finite sample, the conditional bound estimators $(\hat{\theta}^L(\delta,x),\hat{\theta}^U(\delta,x))$ in (4.5) and (4.6) are jointly no less accurate at estimating $(\theta^L(\delta,x),\theta^U(\delta,x))$ as the choice of $\hat{G}_n(y,\delta,x)$ is for $G(y,\delta,x)$, whatever the choice may be.

As the second part of Theorem 4.3.2 suggests, we can turn (4.7) into a statement of frequentist confidence provided such a $t_\alpha(\delta,x)$ exists. Note that $t_\alpha(\delta,x)$ need not vary with x or δ; it can also be with respect to the concentration of $\hat{G}_n(y,\delta,x)$ uniformly across x or across δ (or both) if more appropriate. Theorem 4.3.4 below is an example with such a uniform guarantee.

With regard to specifying $\hat{G}_n(y,\delta,x)$, one choice is to plug-in estimators of $F_w(y|x)$ for $w=0,1$ to arrive at a concentration inequality for the conditional bound estimators as Corollary 4.3.3 suggests. In doing so, provided the appropriate guarantee exists for the conditional CDF estimators, we actually get a strong guarantee for the bounds estimators $\hat{\theta}^L(\delta,x)$ and $\hat{\theta}^U(\delta,x)$ that is simultaneous across all threshold values δ used to define PIBT in Definition 4.1.1.

Corollary 4.3.3 (Conditional bound estimators' concentration when plugging in conditional CDF estimators)**.**
If Assumption 4.1.7 holds, and there exists estimators of $F_0(y|x)$ and $F_1(y|x)$ such that for $w=0,1$ and $\alpha\in(0,1)$ there exists a value $t_{w,\alpha}(x)$ such that:

$$Pr\left(\sup_y\left|\hat{F}_{wn}(y|x)-F_w(y|x)\right|\leq t_{w,\alpha}(x)\right)\geq 1-\alpha/2,$$

then we have that:

$$Pr\left(\sup_\delta\left(\left|\hat{\theta}^L(\delta,x)-\theta^L(\delta,x)\right|\vee\left|\hat{\theta}^U(\delta,x)-\theta^U(\delta,x)\right|\right)\leq\sum_{w=0,1}t_{w,\alpha}(x)\right)\geq 1-\alpha.$$

The proof of Corollary 4.3.3 is found in Appendix 4.B.4.

4.3.2 More explicit conditional bounds with strategic use of regression residuals

Given that we are after confidence bands on PIBT in this chapter, the question now becomes how exactly we should specify $\hat{F}_{wn}(y|x)$ in Corollary 4.3.3, while guaranteeing the closeness between $(\hat{\theta}^L(\delta, x), \hat{\theta}^U(\delta, x))$ and $(\theta^L(\delta, x), \theta^U(\delta, x))$ at some target confidence level. We explore one such choice using regression residuals for which such a high confidence guarantee is possible as summarized in Theorem 4.3.4. The motivation is that we would like something very similar to the plug-in estimator of $\Pr(Y_i(w) \leq y)$ given in Equation (4.4) for the RE case.

Assume $(X_1, W_1, Y_1(0), Y_1(1)), (X_2, W_2, Y_2(0), Y_2(1)), \ldots, (X_n, W_n, Y_n(0), Y_n(1))$ are i.i.d. copies from a joint distribution. Denote the training data as:

$$\mathcal{T} := \{(X_i, W_i, Y_i)\}_{i=1}^n,$$

where $Y_i = W_i Y_i(1) + [1 - W_i]Y_i(0)$. Consider partitioning $\mathcal{T}$ into two independent splits $\mathcal{T}_1$ and $\mathcal{T}_2$. Denote the corresponding training indices as $\mathcal{I}_1, \mathcal{I}_2 \subseteq \{1, \ldots, n\}$ for $\mathcal{T}_1$ and $\mathcal{T}_2$, respectively. Denote $S_w := \{i : \ W_i = w\}$, the index set of individuals in the sample in treatment group $w = 0, 1$. Let $n_w := |S_w \cap \mathcal{I}_2|$, the sample size in treatment group $w = 0, 1$ coming from data split $\mathcal{T}_2$.

Further, denote

$$\mu_w(x) := \mathbb{E}[Y_i(w)|X_i = x],$$

the conditional expectation of the potential outcome as a function of the pre-treatment covariates. Denote the regression estimate using $\mathcal{T}_1$ as $\hat{\mu}_w(x)$. Importantly, we will be able to reason about the counterfactual quantity $\mu_w(x)$ under Assumption 4.1.7, because:

$$\mu_w(x) = \mathbb{E}[Y_i|W_i = w, X_i = x],$$

the conditional expectation of the observed outcome in treatment group $w = 0, 1$. Now consider the population-level residuals:

$$R_i(w) := Y_i(w) - \mu_w(X_i).$$

Denote the approximation of $R_i(w)$ using $\hat{\mu}_w(\cdot)$ as

$$\hat{R}_i(w) := Y_i(w) - \hat{\mu}_w(X_i).$$

Motivated by the use of i.i.d. draws from the distribution $Y_i|W_i = w$ used to define $\hat{F}_{wn}(y)$ in Equation (4.4) for the marginal PIBT in a RE setting, we would like to approximate draws from the distribution $Y_i|X_i = x, W_i = w$. With this in mind, we will specify:

$$\hat{F}_{wn}(y|x) := \frac{1}{n_w} \sum_{i \in S_w \cap \mathcal{I}_2} \mathbf{1}\left\{\hat{\mu}_w(x) + \hat{R}_i(w) \leq y\right\}. \tag{4.8}$$

Considering that the definition of $R_i(w)$ means that $Y_i(w) = \mu_w(x) + R_i(w)$ conditional on $X_i = x$, it seems that using $\hat{\mu}_w(x) + \hat{R}_i(w)$ will make this choice of $\hat{F}_{wn}(y|x)$ a reasonable approximation to $F_w(y|x)$.

Noting the liberal use of the plug-in principle en route to the choice of $(\hat{\theta}^L(\cdot), \hat{\theta}^U(\cdot))$ using the conditional CDF estimators in (4.8), a concern is now what the regularity conditions must be so that overall the conditional bound estimators are close to their true values. For any given value of x, $\hat{F}_{wn}(y|x)$ is reusing residuals for indices in the sample (split $\mathcal{I}_2$) corresponding to subjects that are not necessarily in stratum x. Implicit in this use is that the distribution of $(R_i(0), R_i(1))$ is the same across values of x. That is, we are using the independence assumption:

$$(R_i(0), R_i(1)) \perp\!\!\!\perp X_i.$$

Beyond this regularity condition, we also require that the distribution of $\hat{R}_i(w)$ approximates well the distribution of $R_i(w)$, which in turn requires that $\hat{\mu}_w(\cdot)$ be close to $\mu_w(\cdot)$. This explains the correction to the confidence level in Theorem 4.3.4 with respect to how likely a deviation, in a uniform sense, is to occur between the true regression curve and the estimated regression curve based on random training data.

In Theorem 4.3.4, $\mathbf{X} \in \mathbb{R}^{|\mathcal{I}_1| \times p}$ is such that its rows are comprised of $X_i^T \in \mathbb{R}^{1 \times p}$ for $i \in \mathcal{I}_1$.

Theorem 4.3.4 (Concentration Inequality for the Conditional Bounds on PIBT using regression residuals)**.**

For the bound estimators in Equations (4.5) *and* (4.6)*, let us specify* $\hat{G}(y|x) = \hat{F}_{1n}(y|x) - \hat{F}_{0n}(y|x)$ *using the conditional CDF estimators of* (4.8)*. Also let Assumption 4.1.7 (strong conditional ignorability) hold and assume further that the arbitrary joint distribution of* $(R_i(0), R_i(1), X_i)$ *is such that*

$$(R_i(0), R_i(1)) \perp\!\!\!\perp X_i.$$

Conditional on $\mathbf{X}$*, we have for any appropriate*[1] $\alpha \in (0,1)$*:*

$$\begin{aligned}
&\sup_{\delta,x}\left(\left|\hat{\theta}^L(\delta,x) - \theta^L(\delta,x)\right| \vee \left|\hat{\theta}^U(\delta,x) - \theta^U(\delta,x)\right|\right) \\
\leq &\sum_{w=0,1} \sup_r \left\{Pr(r < R_i(w) \leq r + 2t_w|\mathbf{X}) \vee Pr(r - 2t_w < R_i(w) \leq r|\mathbf{X})\right\} \\
&+ \left(\frac{\log(4/\alpha)}{2}\right)^{\frac{1}{2}} \left(n_0^{-\frac{1}{2}} + n_1^{-\frac{1}{2}}\right)
\end{aligned}$$

with probability at least

$$1 - \alpha - \sum_{w=0,1} Pr\left(\sup_x |\hat{\mu}_w(x) - \mu_w(x)| > t_w \middle| \mathbf{X}\right).$$

Here, $t_0, t_1 \geq 0$ *may depend on* $\mathbf{X}$*. If they do not, we may remove the conditional statements.*

The proof of Theorem 4.3.4 is contained in Appendix 4.B.5. Given Corollary 4.3.3, the task in this proof is to characterize the high probability concentration between the conditional CDF estimators of (4.8) and the true conditional CDFs. This involves a strategic application of the DKW inequality that is tailored to the imputed draws from the conditional potential outcome distribution, as well as incorporating the deviation between the estimated regression curves and the true regression curves.

Consider the following corollary to Theorem 4.3.4. Corollary 4.3.5 means that, with respect to statistical efficiency, we lose nothing with the plug-in estimation approach by building on an estimator of $(\mu_0(\cdot), \mu_1(\cdot))$. One can apply Theorem 4.3.4 with any meta-learning algorithms

[1] *Here, "appropriate" values of* α *are those such that* $1 - \alpha - 2\max_{w=0,1} Pr(\sup_x |\hat{\mu}_w(x) - \mu_w(x)| > t)$ *is between* 0 *and* 1*.*

that are used to estimate the conditional average treatment effect function (CATE) (Künzel et al. 2019, Nie and Wager 2020, Athey et al. 2019, Wager and Athey 2018, Kennedy 2020):

$$\tau(x) := \mu_1(x) - \mu_0(x).$$

Corollary 4.3.5 is stating that in learning $(\theta^L(\cdot), \theta^U(\cdot))$, we retain the same rate as any one of these methods.

Corollary 4.3.5 (Efficiency of Conditional Bound Estimators in Theorem 4.3.4)**.**
Let $\mathcal{F}$ be the function class containing our regression estimator, $\hat{\mu}_w(\cdot)$. Assume there exists a sequence $g_{n,\mathcal{F}}$, depending on n and the complexity of $\mathcal{F}$ (e.g. feature dimension, regularization parameters)[2]*, such that*

$$\max_{w=0,1} \sup_x |\hat{\mu}_w(x) - \mu_w(x)| \lesssim g_{n,\mathcal{F}}$$

with probability at least $1-\alpha$. Then:

$$\sup_{\delta,x} \left(\left| \hat{\theta}^L(\delta, x) - \theta^L(\delta, x) \right| \vee \left| \hat{\theta}^U(\delta, x) - \theta^U(\delta, x) \right| \right) \lesssim g_{n,\mathcal{F}} + \left(\frac{\log(4/\alpha)}{2} \right)^{\frac{1}{2}} \left(n_0^{-\frac{1}{2}} + n_1^{-\frac{1}{2}} \right)$$

holds with probability at least $1-2\alpha$.

Proof.
Denote $f_w(r)$ as the marginal density of $R_i(w)$ for $w = 0, 1$. The key here is that

$$\begin{aligned} &\sup_r \left\{ \Pr\left(r < R_i(w) \leq r + 2t_w | \mathbf{X}\right) \vee \Pr\left(r - 2t_w < R_i(w) \leq r | \mathbf{X}\right) \right\} \\ &\leq 2t_w \max_{w=0,1} \sup_r f_w(r), \end{aligned}$$

using that $\int_a^b h(u)du \leq \sup_u |h(u)||b-a|$ for any integrable function h. We also used that $R_i(w) \perp\!\!\!\perp \mathbf{X}$. Under the regularity condition that the density $f_w(r)$ is non-negative and bounded away from infinity, we set $t_w \doteq g_{n,\mathcal{F}}$ for $w = 0, 1$. This inequality and Theorem 4.3.4 allow us to arrive at the desired result. □

[2] *It decreases down to 0 as n increases under proper specification.*

4.3.2.1 An example power analysis using Theorem 4.3.4 and a restricted regression setup

Theorem 4.3.2 and Theorem 4.3.4 provide generic moulds for a statement about inference with the estimated bounds $(\hat{\theta}^L(\cdot), \hat{\theta}^U(\cdot))$. Theorem 4.3.4 provides this inference simultaneously at all pre-treatment covariate strata x. A tall task. For illustrative purposes, we now consider a simple regression setup to give further concreteness to Theorem 4.3.2 and Theorem 4.3.4. With knowledge about how the pre-treatment covariates are distributed and the restricted regression setup in Assumption 4.3.6 below, we would like to understand the behavior of the margin of error for the bound estimators in Theorem 4.3.4. That is, we would like to conduct a statistical power analysis.

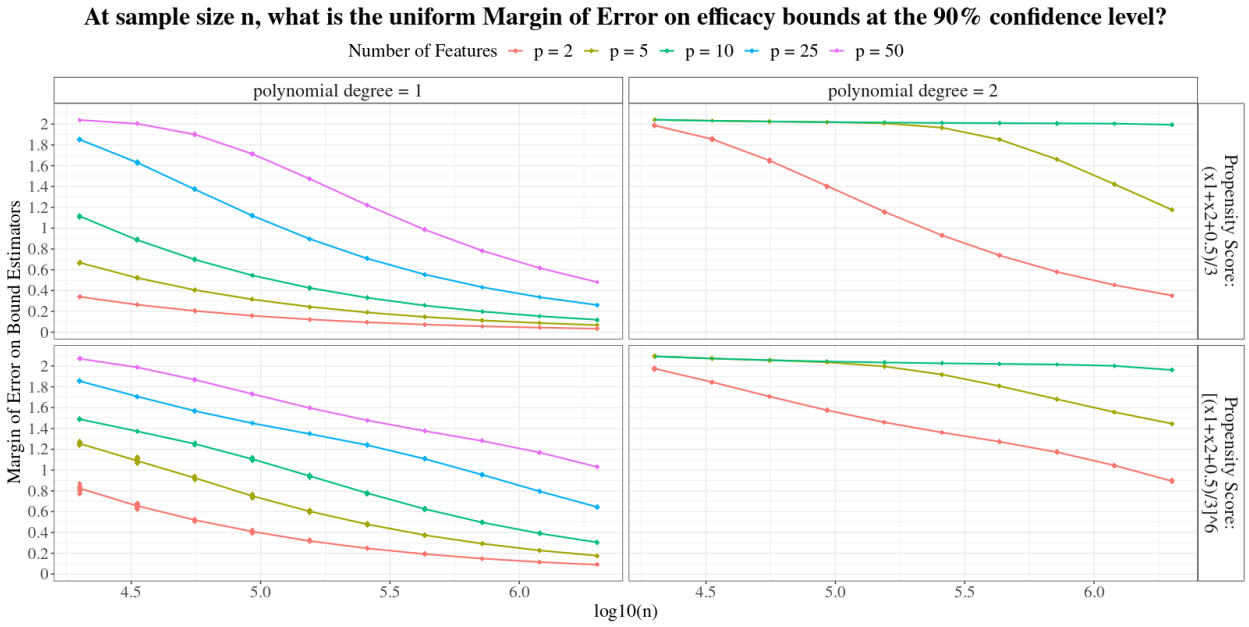

Figure 4.3.1: Power analysis based on Proposition 4.3.7. Each curve is given by the median calculated margin of error across 30 Monte Carlo simulations at the points that are also plotted.

Though somewhat idealistic, the attraction of Proposition 4.3.7 below for the power analysis under Assumption 4.3.6 is that there are no unknowable parameters.

Assumption 4.3.6 (Restricted data generating mechanism)**.**
Across $i = 1, 2, \ldots, n$ and $w = 0, 1$, we will assume the i.i.d. data generating mechanism to be as follows.

1. X_i a vector, possibly random, in $\mathbb{R}^p$.
2. $\mu_w(X_i) = \beta_w^T \Psi_w(X_i)$, where $\Psi_w : \mathbb{R}^p \to \mathbb{R}^d$ is a fixed mapping such that $\|\Psi_w(x)\|_2 \leq 1$ and $d < n$.
3. $Y_i(w) = \mu_w(X_i) + R_i(w)$.
4. $(R_i(0), R_i(1))$ have any joint distribution satisfying $(R_i(0), R_i(1)) \perp\!\!\!\perp X_i$.
5. Marginally, $R_i(w) \sim \mathcal{N}(0, \sigma_w^2)$.

Let $\mathbf{\Lambda}_w \in \mathbb{R}^{|\mathcal{I}_1 \cap S_w| \times d}$ be such that its rows are made up by stacking $\Psi_w(X_i)^T$ for each $i \in \mathcal{I}_1 \cap S_w$. Further, let $\mathbf{Y}_w \in \mathbb{R}^{|\mathcal{I}_1 \cap S_w| \times 1}$ contain entries for the corresponding observed outcome Y_i for each $i \in \mathcal{I}_1 \cap S_w$. In applying Theorem 4.3.4, we will estimate β_w separately for $w = 0, 1$ using ordinary least squares regression, so that:

$$\hat{\mu}_w(x) := \Psi_w(x)^T \hat{\beta}_w; \ \hat{\beta}_w = (\mathbf{\Lambda}_w^T \mathbf{\Lambda}_w)^{-1} \mathbf{\Lambda}_w^T \mathbf{Y}_w.$$

We note that these separate regressions are an instance of the "two-learner" meta learning of the conditional average treatment effect (CATE); CATE's estimate with a two-learner is given by $\hat{\mu}_1(x) - \hat{\mu}_0(x)$ (Künzel et al. 2019).

For Proposition 4.3.7, let Φ denote the CDF of the standard normal distribution. For a matrix $A \in \mathbb{R}^{m \times q}$, denote its operator norm as:

$$|||A|||_{op} := \sup_{v \in \mathbb{R}^q: \ \|v\|_2 = 1} \|Av\|_2 .$$

Proposition 4.3.7 (Uniform confidence bands for the linear case with homoscedastic, Gaussian residuals)**.**
Assume rank$(\mathbf{\Lambda}_w) = d$ *for* $w = 0, 1$ *almost surely. Let* $v_{d,\alpha}$ *denote the* $(1 - \alpha/2)^{th}$ *quantile for*

the χ^2_d distribution. Under Assumption 4.3.6, we have with confidence at least $(1-2\alpha)\times 100\%$ that uniformly across all pre-treatment covariate strata x,

$$\theta(\delta, x) = Pr\left(Y_i(1) - Y_i(0) > \delta | X_i = x\right)$$

is contained in the interval with starting point

$$\hat{\theta}^L(\delta, x) - \sum_{w=0,1} \left\{ \Phi\left(\sqrt{v_{d,\alpha}} \left| \left| \left| (\mathbf{\Lambda}_w^T \mathbf{\Lambda}_w)^{-1/2} \right| \right| \right|_{op} \right) - \Phi\left(-\sqrt{v_{d,\alpha}} \left| \left| \left| (\mathbf{\Lambda}_w^T \mathbf{\Lambda}_w)^{-1/2} \right| \right| \right|_{op} \right) \right\}$$
$$- \left(\frac{\log(4/\alpha)}{2} \right)^{\frac{1}{2}} \left(n_0^{-\frac{1}{2}} + n_1^{-\frac{1}{2}} \right)$$

and end point

$$\hat{\theta}^U(\delta, x) + \sum_{w=0,1} \left\{ \Phi\left(\sqrt{v_{d,\alpha}} \left| \left| \left| (\mathbf{\Lambda}_w^T \mathbf{\Lambda}_w)^{-1/2} \right| \right| \right|_{op} \right) - \Phi\left(-\sqrt{v_{d,\alpha}} \left| \left| \left| (\mathbf{\Lambda}_w^T \mathbf{\Lambda}_w)^{-1/2} \right| \right| \right|_{op} \right) \right\}$$
$$+ \left(\frac{\log(4/\alpha)}{2} \right)^{\frac{1}{2}} \left(n_0^{-\frac{1}{2}} + n_1^{-\frac{1}{2}} \right).$$

The proof of Proposition 4.3.7 is contained in Appendix 4.B.7. The idea is to tailor Theorem 4.3.4 to the parametric assumptions and the two-learner. Moreover, the conclusion that $\theta(\delta, x)$ is contained in the specified interval follows because $\theta^L(\delta, x) \leq \theta(\delta, x) \leq \theta^U(\delta, x)$ by definition, and because the quantity added/subtracted to the estimated bounds is the form taken on by their margin of error at the $1-2\alpha$ confidence level.

Under Assumption 4.3.6, Figure 4.3.1 illustrates the behavior of the margin of error for

$$\sup_{\delta, x} \left(\left| \hat{\theta}^L(\delta, x) - \theta^L(\delta, x) \right| \vee \left| \hat{\theta}^U(\delta, x) - \theta^U(\delta, x) \right| \right)$$

at the 90% confidence level. As one can imagine, the distribution of X_i, the transformation Ψ_w in Assumption 4.3.6, and the propensity score matter for an application of Proposition 4.3.7. That is, there may very well exist cases where the operator norm of $(\mathbf{\Lambda}_w^T \mathbf{\Lambda}_w)^{-1/2}$ does not decrease with n, along with cases of the propensity score where this norm decreases slowly due to insufficient treated or control units in the sample. To study this, the example summarized in Figure 4.3.1 generates data as follows:

1. We sample from a population in which each $X_{ij} \overset{i.i.d.}{\sim} \text{Uniform}(0,1)$ across $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$.

2. The degree $q = 1, 2$ polynomial transformation of X_i into $\Psi_w(X_i)$ includes all possible interaction terms of degree $1 \leq k \leq q$, and each entry is re-scaled by $1/\sqrt{d}$ so that $\|\Psi_w(X_i)\|_2 \leq 1$ as in Assumption 4.3.6.

    - For example, $\Psi_w(X_i) = (X_{i1}, X_{i2}, X_{i1}X_{i2}, X_{i1}^2, X_{i2}^2)/\sqrt{5}$ when $p = q = 2$.

3. The propensity score is $\Pr(W_i = 1|X_i) = \left[\frac{1}{3}(X_{i1} + X_{i2} + 0.5)\right]^m$ for $m = 1, 6$.

    - The case with $m = 1$ makes it so that the number of treated and control units are very close to each other in a random sample, while the case with $m = 6$ will make it so that control units are typically much more represented.

4. Moreover, the sample splitting is such that half of the observations are used to estimate $(\mu_0(\cdot), \mu_1(\cdot))$, while the other half of the observations' residuals are used to estimate $(\theta^L(\cdot), \theta^U(\cdot))$.

We believe the example application of Proposition 4.3.7 in Figure 4.3.1 may be regarded as a microcosm of what can occur in practice while applying the estimation procedure outlined in Section 4.3.2 (or similar) for the bounds on $\Pr(Y_i(1) - Y_i(0) > \delta|X_i = x)$ uniformly across x. The primary concerns for satisfactory margins of error are the typical concerns of regression: parsimony, multicollinearity, and feature dimension. In terms of parsimony, a more complex model specification as understood by the polynomial degree in Figure 4.3.1 requires much more data for reasonable margins of error. Regarding multicollinearity, because the features X_i are generated with independent entries, the case of polynomial degree 1 generally has sharper decreases in margins of error compared to the more complex models where the entries in $\Psi_w(X_i)$ can become correlated. In terms of feature dimension, we also generally see slower rates of decrease in the margins of error when p or d is larger.

4.4 More on the scope of our results

The contribution of this chapter is in the quantification of the margin of error for the bounds on PIBT given by the Makarov bounds (Makarov 1982, Frank et al. 1987, Williamson and Downs 1990, Fan and Park 2010). Three important questions may come up:

- Can we obtain better bounds for binary outcomes, such as with the Fréchet-Boole inequalities (Boole 1854, Hailperin 1986, Fréchet 1935, 1960, Mueller and Pearl 2019)?
- Can we bound the proportion who are harmed by an intervention?

We will answer these in the following subsections.

4.4.1 When the potential outcomes are binary, the Makarov bounds on PIBT are the same as the Boole-Fréchet bounds

Recall that the probability of necessity and sufficiency (PNS) for binary potential outcomes (Pearl 1999, Tian and Pearl 2000, Pearl 2009) is given by the joint probability,

$$\Pr\left(Y_i(1) = 1, Y_i(0) = 0\right).$$

For two measurable events A and B, the Boole-Fréchet bounds on their joint probability are:

$$\max\left(\Pr\left(A\right) + \Pr\left(B\right) - 1, 0\right) \;\leq\; \Pr\left(A \cap B\right) \;\leq\; \min\left(\Pr\left(A\right), \Pr\left(B\right)\right).$$

These bounds are sharp (Rüschendorf 1981). For example, if $A \cap B = \emptyset$, then $\Pr\left(A \cap B\right) = \max\left(\Pr\left(A\right) + \Pr\left(B\right) - 1, 0\right)$ holds. Moreover, if $A \subseteq B$, then $\Pr\left(A \cap B\right) = \min\left(\Pr\left(A\right), \Pr\left(B\right)\right)$ is the case.

We claim in Proposition 4.4.1 that the Makarov bounds on PNS are the same as the Boole-Fréchet bounds. From this proposition, it follows that all the estimation results we have shown are as applicable to the binary potential outcome case as any estimation approach involving the Boole-Fréchet bounds.

Proposition 4.4.1 (Boole-Fréchet bounds vs. Makarov bounds)**.**

When the potential outcomes are binary, we have that the tightest Makarov bounds on PNS and conditional PNS are:

- *Marginal Case:*

$$\sup_{\delta\in[0,1)} \theta^L(\delta) \;\leq\; Pr(Y_i(1)=1, Y_i(0)=0) \;\leq\; \inf_{\delta\in[0,1)} \theta^U(\delta).$$

- *Conditional Case:*

$$\sup_{\delta\in[0,1)} \theta^L(\delta,x) \;\leq\; Pr(Y_i(1)=1, Y_i(0)=0|X_i=x) \;\leq\; \inf_{\delta\in[0,1)} \theta^U(\delta,x).$$

Moreover, these tightest Makarov bounds are the same as the Boole-Fréchet bounds:

- *Marginal Case:*
  - $\sup_{\delta\in[0,1)} \theta^L(\delta) = \max\left(Pr(Y_i(1)=1) + Pr(Y_i(0)=0) - 1, 0\right)$*;*
  - $\inf_{\delta\in[0,1)} \theta^U(\delta) = \min\left(Pr(Y_i(1)=1), Pr(Y_i(0)=0)\right)$.
- *Conditional Case:*
  - $\sup_{\delta\in[0,1)} \theta^L(\delta,x) = \max\left(Pr(Y_i(1)=1|X_i=x) + Pr(Y_i(0)=0|X_i=x) - 1, 0\right)$*;*
  - $\inf_{\delta\in[0,1)} \theta^U(\delta,x) = \min\left(Pr(Y_i(1)=1|X_i=x), Pr(Y_i(0)=0|X_i=x)\right)$.

The proof of Proposition 4.4.1 can be found in Appendix 4.B.8. The general idea is to use that the CDFs of binary potential outcomes have only three values, 0, $\Pr(Y_i(w)=0)$, and 1, across evaluation points $y \in \mathbb{R}$. We also use the fact that PIBT is the same as PNS when the threshold δ is in $[0,1)$.

4.4.2 Reasoning about the proportion harmed by an intervention

Two days following the posting of our first manuscript on `arXiv`, a similar manuscript by Kallus (2022) was also posted on `arXiv`. This work studies the probability an individual

is harmed by an intervention (PIHI) in the case of binary potential outcomes. Supposing instead that $Y_i = 1$ is bad for an individual, while $Y_i = 0$ is good for an individual, we have that the PIHI is given by PNS:

$$\Pr\left(Y_i(1) = 1, Y_i(0) = 0\right).$$

This is simply, but importantly, due to notational and vocabular semantics. Moreover, if the potential outcomes are real-valued, then one can refer to the quantity

$$\Pr\left(Y_i(1) - Y_i(0) > \delta\right),$$

as PIHI instead of calling this quantity PIBT. This holds analogously when we condition on $X_i = x$.

Given the discussion surrounding Proposition 4.4.1, our model-free results extend to both real-valued and binary potential outcome cases for PIHI as well. Given that the Boole-Fréchet inequalities underlay the theoretical estimation results for binary potential outcomes in Kallus (2022), we believe that the differing contributions in their work compared to ours are:

- Their results are for binary potential outcomes, while our results hold for both binary and real-valued potential outcomes.

- Their results include a doubly robust estimation method which uses the estimated propensity score to adjust for covariates, possibly in an observational setting, when bounding marginal PIHI. We believe this warrants further investigation for the case of real-value potential outcomes. Moreover, our confidence bands for marginal PIBT are only for the randomized experiment case.

- Their presentation of estimation theorems are in terms of big-O probability notation, i.e. statistical rates of convergence. We present results in terms of non-asymptotic concentration statements. The subtle difference is that our presentation helps provide nominal coverage guarantees for the confidence bands on PIBT (or PIHI).

- Moreover, confidence bands for PIHI presented in their work make use of a standard error (the standard deviation of a random variable being averaged to obtain the bound estimators divided by $\sqrt{n}$). Likely due to an implicit Central Limit Theorem, these standard-error based confidence bands do not achieve nominal coverage until a fairly large sample size as demonstrated in their empirical results. See their *Algorithm 1* and *Figure 3* for details.

- The non-asymptotic presentation of our results, having kept track of all constants, can also help with a statistical power analysis, as discussed in Sections 4.2.3 and 4.3.2.1. However, we do note that keeping track of constants can prove difficult for some applications of Theorems 4.3.2 and 4.3.4; more generally, plausible constants can be specified in such regression settings.

4.5 Application to Criteo's uplift prediction benchmark dataset

We now present an application to Criteo AI Lab's uplift prediction benchmark dataset (Diemert Eustache, Betlei Artem et al. 2018). According to the webpage[3] that hosts the data,

> This dataset is constructed by assembling data resulting from several incrementality tests, a particular randomized trial procedure where a random part of the population is prevented from being targeted by advertising. It consists of 25M rows, each one representing a user with [12] features, a treatment indicator and 2 labels (visits and conversions).

For this application, the proportion we will estimate bounds for should be understood in plain language as the proportion of the time that the advertiser benefits from presenting an advertisement on the website, rather than the probability an individual benefits from treatment. Equation (4.9) below is the formal statement of this proportion.

[3] `https://ailab.criteo.com/criteo-uplift-prediction-dataset/`

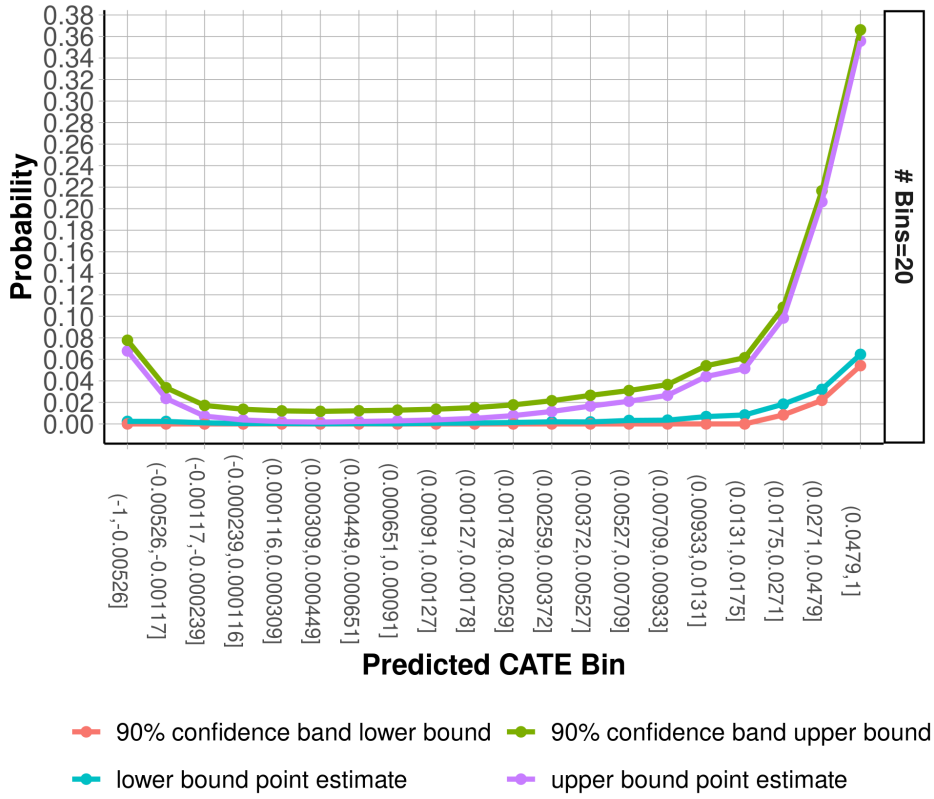

Figure 4.5.1: The 90% Bonferroni corrected lower confidence band on PIBT across bins of CATE predictions on the Criteo uplift dataset.

The available down sampled data consists of 13,979,592 observations. We focus on the effect treatment assignment (rather than treatment receipt) has on visits, making this an intent-to-treat analysis. The outcome of interest in our analysis is the visit indicator for whether a user visited the advertiser website during the test period (2 weeks).

Using the 12 pre-treatment covariates, X_i, we will study heterogeneity as follows.

1. **Obtain CATE Estimate**: With 50,000 randomly sampled rows, we will use the `grf::causal_forest` (default options) in `R` (Athey et al. 2019) to learn the conditional average treatment effect (CATE) function,

$$\tau(x) := \mathbb{E}\left[Y_i(1) - Y_i(0) | X_i = x\right].$$

2. **Partition based on similar CATE predictions:** Denote the quantiles of $\hat{\tau}(X_i)$ as

$$q_\alpha := \inf \{q : \Pr(\hat{\tau}(X_i) \leq q) \geq \alpha\}$$

for $\alpha \in [0, 1]$. We will obtain estimators, $\hat{q}_\alpha$ using the second fold of 8,987,000 randomly sampled rows. Now, for $m = 20$, define the discrete mapping $s_m : \mathcal{X} \to \{1, \ldots, m\}$ as:

$$s_m(x) = k \text{ if } \hat{\tau}(X_i) \in (\hat{q}_{(k-1)/m}, \hat{q}_{k/m}]; k \in \{1, \ldots, m\}.$$

The value m corresponds to the number of sub-groups that are created in the partitioning.

3. **Estimate PIBT conditional on partition**: We next conduct inference via the bound estimators for

$$\theta(\delta, s_m(x)) := \Pr(Y_i(1) - Y_i(0) > \delta | s_m(X_i) = s_m(x)), \tag{4.9}$$

the probability that treatment is beneficial, for the business in this example, in CATE stratum $s_m(x)$.

   - We note that, because of the randomization of treatment, we have the ignorability statement $(Y_i(1), Y_i(0)) \perp\!\!\!\perp W_i | s_m(X_i) = s_m(x)$. This allows us to identify the Makarov lower and upper bounds on $\theta(\delta, s_m(x))$, which we will denote as $\theta^L(\delta, s_m(x))$ and $\theta^U(\delta, s_m(x))$, respectively.

The mappings $s_m(\cdot)$ allow us to stratify on an interpretable univariate score, which is the estimated CATE function in this case. With the mapping $s_m(\cdot)$, if the learned function $\hat{\tau}(\cdot)$ is indicative of benefiting from treatment, we would like to see a monotone increasing relation between $k = 1, \ldots, m$ and the bound estimators $(\hat{\theta}^L(\delta, k), \hat{\theta}^U(\delta, k))$ as supporting evidence.

The stratification with the quantity we denote as $\theta(\delta, s_m(x))$ in this subsection is similar to what has been done with prognostic and propensity scores (Abadie et al. 2018, Padilla et al. 2021, Ye et al. 2021b). Similarly, there is recent work for the estimation of the

quantity $\mathbb{E}\left[Y_i(1) - Y_i(0) | h(X_i) > \gamma\right]$ (Yadlowsky et al. 2021). Here, $h(X_i)$ is a univariate score that is predictive of individual i's treatment effect, such as a prognostic score or $\hat{\tau}(X_i)$, and γ is tunable threshold that allows us to determine what individuals ought be given priority to treatment (Yadlowsky et al. 2021). Moreover, estimation strategies for $\mathbb{E}\left[Y_i(1) - Y_i(0) | \hat{\tau}(X_i) = s\right]$ have also be studied in order to determine whether $\hat{\tau}(X_i)$ is well calibrated (Xu and Yadlowsky 2022). To complement these advances in univariate score stratification, bounds on $\theta(\delta, s_m(x))$ allow us to determine the implications with respect to treatment benefit a given CATE estimate provides.

Figure 4.5.1 provides a 90% Bonferroni corrected confidence band on

$$\theta(\delta, k) = \Pr\left(Y_i(1) = 1, Y_i(0) = 0 | s_m(X_i) = k\right); \delta \in [0, 1)$$

that allows for simultaneous inference across CATE prediction bins $k = 1, 2, \ldots, 20$. The practical insight we have is that for an individual such that $\hat{\tau}(X_i) \in (0.0479, 1]$, the joint probability of interest–the probability an individual will visit the advertiser's webpage when assigned treatment and otherwise not visit the website if untreated–is between 5.42% and 36.62% with 90% confidence. Given the Bonferroni correction and that all bins have equal amounts of subjects, we may take a simple average of the upper confidence bound for every bin corresponding to a CATE prediction of 0.0271 or less. This tells than an individual with a CATE prediction of 0.0271 or less has a joint probability of interest of 4.18% or less with 90% confidence.

Of interest, the sample size to estimate $\theta^L(\delta = 0, k)$ across $k = 1, 2, \ldots, 20$ is determined according to the discussion in Section 4.2.3. We specified the margin of error between $\theta^L(\delta = 0, k)$ and $\hat{\theta}^L(\delta = 0, k)$ to be $\varepsilon = 0.01$. Because we desired 90% confidence jointly for each k, the task was to find the sample size n_k such that $\alpha_\varepsilon = 0.1/20 = 0.005$ (recall the Bonferroni correction). Because the treatment frequency in the dataset was 85%, we used the constraint that the number of untreated and treated units per bin is $0.15n_k$ and $0.85n_k$,

respectively. This leads us to create groups with $n_k = 449,350$ subjects each[4].

4.6 Discussion

For the bounds on the marginal PIBT estimated with data from a randomized experiment, we derive a closed-form concentration inequality depending on only the sample size and the desired frequentist confidence level. We discussed how this margin of error can be used for a formal statistical power analysis in Section 4.2.3.

Making strategic use of regression residuals, we also discussed how to estimate, possibly in an observational setting, the PIBT conditional on strata of an individual's pre-treatment co-variates. For this approach with regression residuals, we provide novel, tailored versions of a general statement that allow for a frequentist confidence interpretation simultaneously at all pre-treatment covariate strata. To provide an example application for this result, we demonstrated in Section 4.3.2.1 how one may use it in a low dimensional linear regression setting with Gaussian noise.

We included in Section 4.4 an extended discussion on the scope of our results. We showed in Proposition 4.4.1 that the Makarov bound approach we take to bound PIBT is equivalent to using the sharp Boole-Fréchet bounds. We discussed how to estimate bounds on PIBT when we define benefiting from treatment in terms of the ratio of an individual's two positive potential outcomes as discussed in Section 4.2.4. Moreover, this section also discusses how our results can easily extend to reasoning about the proportion of individuals that are harmed by an intervention (Kallus 2022). We also included in Section 4.5 an example application to a randomized experiment dataset, Criteo AI Lab's benchmark data for uplift prediction (Diemert Eustache, Betlei Artem et al. 2018). In particular, this section points toward a useful combination of conditional average treatment effect (CATE) estimation and inferring

[4]The constraint of $0.15n_k$ and $0.85n_k$ untreated and treated units ended up being approximate due to the random sampling of rows.

PIBT (or related quantity) to better understand the implication of the CATE estimate.

Interesting extensions of the work presented here include applying the inequality in Theorem 4.3.2 to an estimation approach that is more general than that provided in Theorem 4.3.4. Alternatively, we may like to tailor a version of Theorem 4.3.4 to certain modeling assumptions that are sufficient for interesting applications, such as involving generalized linear models (McCullagh and Nelder 2019, Sur and Candès 2019). With respect to Theorem 4.3.4 itself, we think it is also worthwhile in practice to apply it to regression scenarios beyond the unregularized linear Gaussian model. For settings where comparing an individual's two potential outcomes in a ratio (rather than a difference) can provide interesting insight, such as studies where time to an event is of interest (Cox 1972, Stitelman and van der Laan 2010, Austin 2014, Schober and Vetter 2018, Cai and van der Laan 2020), it seems worthwhile to extend the discussion in Section 4.2.4.

Our probability bound formulation to reason about treatment effects, rather than the more common average formulation, is similar to recent model-free work that moves beyond an average in favor of controlling the type I error (false discovery) violation probability (Tong et al. 2018, Li et al. 2021). Tong et al. (2018) propose a Neyman-Pearson classification paradigm that that helps prevent classification algorithms from incorrectly classifying individuals with $Y_i = 0$ labels in sensitive scenarios where doing so can have negative consequences. Meanwhile, Li et al. (2021) provide a method for marginal ranking of features for binary classification using a classical criterion and the Neyman-Pearson criterion. Future work relating our work in the present chapter to these two works could be of interest. It can involve more rigorously choosing the appropriate threshold δ used to define PIBT in Definition 4.1.1 in accordance with an appropriately formulated Neyman-Pearson objective. Currently, the treshold δ is anything (or everything; recall the uniform control across this threshold) a practitioner finds reasonable.

APPENDIX

4.A The Makarov bounds

Lemma 4.A.1 (The Makarov Bounds as stated in Williamson and Downs (1990)'s *Theorem 2*)**.**

Uniformly across all possible, unknown joint distributions

$$(V_1, V_2) \sim Pr(V_1 \leq v_1, V_2 \leq v_2)$$

having fixed marginal CDFs $F_1(v_1) = Pr(V_1 \leq v_1), F_2(v_2) = Pr(V_2 \leq v_1)$*, the CDF of* $V_1 - V_2$ *evaluated at* $\delta \in \mathbb{R}$ *satisfies:*

$$F^L(\delta) \leq \quad Pr(V_1 - V_2 \leq \delta) \quad \leq F^U(\delta), \tag{4.10}$$

where

$$F^L(\delta) = \quad \max\left(\sup_{a,b:\ a+b=\delta} \{F_1(a) - F_2(-b)\}, 0\right)$$

and

$$F^U(\delta) = \quad 1 + \min\left(\inf_{a,b:\ a+b=\delta} \{F_1(a) - F_2(-b)\}, 0\right).$$

Lemma 4.A.1 was first proved in (Makarov 1982) to bound the distribution of a sum of two random variables. We present this result for subtraction, which is a simple extension, as $V_1 - V_2$ is technically the sum of two random variables V_1 and $(-V_2)$. Lemma 4.A.1's proof was later rigourized in (Frank et al. 1987) and (Williamson and Downs 1990), who also seek bounds on the distribution of other binary operations on V_1 and V_2, like their difference, product, and their ratio, under minimal distributional assumptions.

4.A.1 Equivalent forms of the bounds

We note that $F^L(\delta)$ and $F^U(\delta)$ in Lemma 4.A.1 can be rewritten.

1. Consider a one-to-one change of variables $(a, b) \mapsto (u + \delta/2, -u + \delta/2)$. With it, we have equivalently:

$$F^L(\delta) \;=\; \max\left(\sup_u\{F_1(u+\delta/2) - F_2(u-\delta/2)\}, 0\right)$$

along with

$$F^U(\delta) \;=\; 1 + \min\left(\inf_u\{F_1(u+\delta/2) - F_2(u-\delta/2)\}, 0\right).$$

This is in line with what we have written in Section 4.2 and 4.3 of the main text.

2. Consider instead the one-to-one change of variables $(a, b) \mapsto (u, -u + \delta)$. With it, we have equivalently:

$$F^L(\delta) \;=\; \max\left(\sup_u\{F_1(u) - F_2(u-\delta)\}, 0\right)$$

along with

$$F^U(\delta) \;=\; 1 + \min\left(\inf_u\{F_1(u) - F_2(u-\delta)\}, 0\right).$$

This is in line with Fan and Park (2010)'s *Lemma 2.1* and *Equation (2) and (3)*, and it also agrees with the alternative form given in *Equations (21) and (22)* of Williamson and Downs (1990).

Moreover, it is straightforward to see that:

$$1 - F^U(\delta) \;\leq\; \Pr\left(U_1 - U_0 > \delta\right) \;\leq\; 1 - F^L(\delta).$$

We make use of this in the main text when bounding PIBT.

4.B Proofs for the main theoretical results

4.B.1 The key lemma

Lemma 4.B.1 (Plug-in estimation of Makarov (1982)'s conditional bounds)**.**
Consider jointly distributed random variables (U_0, U_1, V)*. Denote:*

$$\gamma^L(\delta, v) := -\min\left(\inf_u\left\{Pr\left(U_1 \leq u + \delta/2 | V = v\right) - Pr\left(U_0 \leq u - \delta/2 | V = v\right)\right\}, 0\right)$$

and

$$\gamma^U(\delta, v) := 1 - \max\left(\sup_u \left\{Pr\left(U_1 \leq u + \delta/2 | V = v\right) - Pr\left(U_0 \leq u - \delta/2 | V = v\right)\right\}, 0\right),$$

the Makarov (1982), Williamson and Downs (1990) lower and upper bounds for $Pr\left(U_1 - U_0 > \delta | V = v\right)$. *Denote*

$$H(u, \delta, v) := Pr\left(U_1 \leq u + \delta/2 | V = v\right) - Pr\left(U_0 \leq u - \delta/2 | V = v\right).$$

Consider any estimator $\hat{H}(u, \delta, v)$ *of* $H(u, \delta, v)$ *based on a sample*

$$\{(U_{i0}, V_i)\}_{i=1}^{n_0} \cup \{(U_{i1}, V_i)\}_{i=1}^{n_1}$$

such that (U_{i0}, V_i) *are i.i.d. copies of* (U_0, V) *for* $i = 1, \ldots, n_0$ *and* (U_{i1}, V_i) *are i.i.d. copies of* (U_1, V_1) *for* $i = 1, \ldots, n_1$. *Now let*

$$\hat{\gamma}^L(\delta, v) := -\min\left(\inf_u \left\{\hat{H}(u, \delta, v)\right\}, 0\right)$$

and

$$\hat{\gamma}^U(\delta, v) := 1 - \max\left(\sup_u \left\{\hat{H}(u, \delta, v)\right\}, 0\right).$$

We claim for every $\delta \in \mathbb{R}$ *and every* $v \in$ *support*(V) *that*

$$\left|\hat{\gamma}^L(\delta, v) - \gamma^L(\delta, v)\right| \vee \left|\hat{\gamma}^U(\delta, v) - \gamma^U(\delta, v)\right| \;\leq\; \sup_u \left|\hat{H}(u, \delta, v) - H(u, \delta, v)\right|.$$

Proof.

For any real-valued function $g(t)$, denote:

$$g^+(t) = \max\left(g(t), 0\right) \text{ and } g^-(t) = -\min\left(g(t), 0\right),$$

the positive and negative parts of $g(t)$, respectively. We have the following properties we will make use of

- $g(t) = g^+(t) - g^-(t)$.
- $|g(t)| = g^+(t) + g^-(t)$.

- $g^+(t) = \frac{|g(t)|+g(t)}{2}$.
- $g^-(t) = \frac{|g(t)|-g(t)}{2}$.

Below, we will use the positive and negative parts of

$$g_{\inf}(\delta, v) := \inf_u H(u, \delta, v) \text{ and } g_{\sup}(\delta, v) := \sup_u H(u, \delta, v)$$

along with

$$\hat{g}_{\inf}(\delta, v) := \inf_u \hat{H}(u, \delta, v) \text{ and } \hat{g}_{\sup}(\delta, v) := \sup_u \hat{H}(u, \delta, v).$$

Consider:

- **Lower bound on Pr** $(U_1 - U_0 > \delta | V = v)$

  We are bounding the difference

$$\begin{aligned}
& \left|\hat{\gamma}^L(\delta, v) - \gamma^L(\delta, v)\right| \\
= & \left|g^-_{\inf}(\delta, v) - \hat{g}^-_{\inf}(\delta, v)\right| \\
\overset{(i)}{=} & \frac{1}{2}\left|\hat{g}_{\inf}(\delta, v) - g_{\inf}(\delta, v) + |g_{\inf}(\delta, v)| - |\hat{g}_{\inf}(\delta, v)|\right| \\
\overset{(ii)}{\leq} & \left|\hat{g}_{\inf}(\delta, v) - g_{\inf}(\delta, v)\right| \\
= & \left|\inf_u \hat{H}(u, \delta, v) - \inf_u H(u, \delta, v)\right|.
\end{aligned}$$

  In (i), we used the properties of the negative part of a function introduced above, while in (ii) we used triangle inequality followed by reverse triangle inequality. Consider that

$$\begin{aligned}
& \inf_u \hat{H}(u, \delta, v) - \inf_u H(u, \delta, v) \\
= & \inf_u \left[\hat{H}(u, \delta, v) - H(u, \delta, v) + H(u, \delta, v)\right] - \inf_u H(u, \delta, v) \\
\overset{(i)}{=} & \inf_u \left[\hat{H}(u, \delta, v) - H(u, \delta, v)\right] \\
\leq & \sup_u \left|\hat{H}(u, \delta, v) - H(u, \delta, v)\right|.
\end{aligned}$$

Equality (i) follows from the fact that $\inf\{a+b:\ a\in A, b\in B\} = \inf(A)+\inf(B)$. Similarly,

$$\begin{aligned}
&\inf_u H(u,\delta,v) - \inf_u \hat{H}(u,\delta,v) \\
=\ &\inf_u \left[\hat{H}(u,\delta,v) - H(u,\delta,v) + H(u,\delta,v)\right] - \inf_u H(u,\delta,v) \\
\leq\ &\sup_u \left|\hat{H}(u,\delta,v) - H(u,\delta,v)\right|.
\end{aligned}$$

The previous two equations imply that

$$\left|\inf_u H(u,\delta,v) - \inf_u \hat{H}(u,\delta,v)\right| \leq \sup_u \left|\hat{H}(u,\delta,v) - H(u,\delta,v)\right|.$$

so that overall we have that

$$\left|\hat{\gamma}^L(\delta,v) - \gamma^L(\delta,v)\right| \leq \sup_u \left|\hat{H}(u,\delta,v) - H(u,\delta,v)\right|. \tag{4.11}$$

- **Upper bound on $\Pr(U_1 - U_0 > \delta | V = v)$**

We are bounding the difference

$$\begin{aligned}
&\left|\hat{\gamma}^U(\delta,v) - \gamma^U(\delta,v)\right| \\
=\ &\left|g^+_{\sup}(\delta,v) - \hat{g}^+_{\sup}(\delta,v)\right| \\
\overset{(i)}{=}\ &\frac{1}{2}\left|\hat{g}_{\sup}(\delta,v) - g_{\sup}(\delta,v) + |g_{\sup}(\delta,v)| - |\hat{g}_{\sup}(\delta,v)|\right| \\
\overset{(ii)}{\leq}\ &\left|\hat{g}_{\sup}(\delta,v) - g_{\sup}(\delta,v)\right| \\
=\ &\left|\sup_u \hat{H}(u,\delta,v) - \sup_u H(u,\delta,v)\right|.
\end{aligned}$$

In (i), we used the properties of the positive part of a function introduced above, while in (ii) we used triangle inequality followed by reverse triangle inequality. Noting that $\sup\{a+b:\ a\in A, b\in B\} = \sup(A)+\sup(B)$, we can arrive at the below inequality based on similar steps to the case with the lower bound:

$$\left|\sup_u H(u,\delta,v) - \sup_u \hat{H}(u,\delta,v)\right| \leq \sup_u \left|\hat{H}(u,\delta,v) - H(u,\delta,v)\right|.$$

so that overall we have that

$$\left|\hat{\gamma}^U(\delta, v) - \gamma^U(\delta, v)\right| \leq \sup_u \left|\hat{H}(u, \delta, v) - H(u, \delta, v)\right|, \tag{4.12}$$

as with the lower bound estimate.

The inequalities in (4.11) and (4.12) gives us the desired conclusion:

$$\left\{\left|\hat{\gamma}^L(\delta, v) - \gamma^L(\delta, v)\right| \vee \left|\hat{\gamma}^U(\delta, v) - \gamma^U(\delta, v)\right|\right\} \;\leq\; \sup_u \left|\hat{H}(u, \delta, v) - H(u, \delta, v)\right|.$$

□

Lemma 4.B.2 (Bounding a probability statement with respect to a sum of random variables)**.**

Let U and V be arbitrary real-valued random variables, and let $u, v \in \mathbb{R}$ be non-random scalars. We have that:

$$Pr\left(U + V > u + v\right) \leq Pr\left(U > u\right) + Pr\left(V > v\right).$$

Proof.
Consider that we have the following:

$$\{U + V \leq u + v\} \supseteq \{U \leq u\} \cap \{V \leq v\} \iff \{U + V > u + v\} \subseteq \{U > u\} \cup \{V > v\}.$$

This containment of events holds because $U \leq u$ and $V \leq v$ implies that $U + V \leq u + v$. It follows that

$$\Pr\left(U + V > u + v\right) \leq \Pr\left(\{U > u\} \cup \{V > v\}\right) \leq \Pr\left(U > u\right) + \Pr\left(V > v\right),$$

where the second inequality is due to union bound. □

4.B.2 The proof of Theorem 4.2.1

Proof of Theorem 4.2.1.

We apply Lemma 4.B.1 with $U_0 := Y_i(0)$, $U_1 := Y_i(1)$, and V any arbitrary random variable such that $V \perp\!\!\!\perp (Y_i(0), Y_i(1))$. We have that:

$$\begin{aligned}
&\sup_{\delta} \left\{ \left| \hat{\theta}^L(\delta) - \theta^L(\delta) \right| \vee \left| \hat{\theta}^U(\delta) - \theta^U(\delta) \right| \right\} \\
\leq\ &\sup_{\delta, y} \left| \left\{ \hat{F}_{1n}(y + \delta/2) - \hat{F}_{0n}(y - \delta/2) \right\} - \{ F_1(y + \delta/2) - F_0(y - \delta/2) \} \right| \\
\leq\ &\sup_{\delta, y} \left| \hat{F}_{1n}(y + \delta/2) - F_1(y + \delta/2) \right| + \sup_{\delta, y} \left| \hat{F}_{0n}(y - \delta/2) - F_0(y - \delta/2) \right| \qquad (4.13) \\
=\ &\sup_{y} \left| \hat{F}_{1n}(y) - F_1(y) \right| + \sup_{y} \left| \hat{F}_{0n}(y) - F_0(y) \right| \\
\overset{(i)}{\leq}\ &t_0 + t_1.
\end{aligned}$$

Here, $t_w \geq 0$ for $w = 0, 1$. Inequality (i) holds with probability at least

$$1 - 2 \sum_{w=0,1} \exp(-2 n_w t_w^2)$$

due to:

1. Lemma 4.B.2, which implies:

$$\begin{aligned}
&\Pr \left(\sup_{y} \left| \hat{F}_{1n}(y) - F_1(y) \right| + \sup_{y} \left| \hat{F}_{0n}(y) - F_0(y) \right| \geq t_0 + t_1 \right) \\
\leq\ &\sum_{w=0,1} \Pr \left(\sup_{y} \left| \hat{F}_{wn}(y) - F_w(y) \right| \geq t_w \right).
\end{aligned}$$

2. The Dvoretzky-Kiefer-Wolfowitz (DKW) inequality (Dvoretzky et al. 1956, Massart 1990, Naaman 2021) which tells us:

$$\Pr \left(\sup_{y} \left| \hat{F}_{wn}(y) - F_w(y) \right| \geq t_w \right) \leq 2 \exp(-2 n_w t_w^2)$$

   based on the i.i.d. assumption.

Given (4.13), we set

$$t_w \doteq [2^{-1} n_w^{-1} \log(4/\alpha)]^{\frac{1}{2}}$$

for $w = 0, 1$ to arrive at the desired conclusion.

□

4.B.3 The proof of Theorem 4.3.2

Proof of Theorem 4.3.2.

This result is a direct consequence of Lemma 4.B.1 with $U_{i0} := Y_i(0)$, $U_{i1} := Y_i(1)$, and $V_i := X_i$. □

4.B.4 The proof of Corollary 4.3.3

Proof of Corollary 4.3.3.

Specify

$$\hat{G}(y, \delta, x) = \hat{F}_{1n}(y + \delta/2|x) - \hat{F}_{0n}(y - \delta/2|x),$$

which simply plugs in the estimate of $F_1(y + \delta/2|x)$ and $F_0(y - \delta/2|x)$. Consider that:

$$\sup_y \left| \hat{F}_{wn}(y|x) - F_w(y|x) \right| = \sup_{a,y} \left| \hat{F}_{wn}(y + a|x) - F_w(y + a|x) \right|,$$

where the supremum is with respect to y is across the real line, while the supremum across (y, a) is across the euclidean plane. Therefore, when the inequality

$$\sup_y \left| \hat{F}_{wn}(y|x) - F_w(y|x) \right| \leq t_{w,\alpha}(x) \tag{4.14}$$

holds, triangle inequality and the definition of $G(y, \delta, x)$ tell us that

$$\begin{aligned}\sup_{\delta,y}\left|\hat{G}_n(y,\delta,x) - G(y,\delta,x)\right| &\leq \sup_{\delta,y}\left|\hat{F}_{1n}(y+\delta/2|x) - F_w(y+\delta/2|x)\right| \\ &\quad + \sup_{\delta,y}\left|\hat{F}_{0n}(y-\delta/2|x) - F_0(y-\delta/2|x)\right| \\ &= \sup_{y}\left|\hat{F}_{1n}(y|x) - F_w(y|x)\right| + \sup_{y}\left|\hat{F}_{0n}(y|x) - F_0(y|x)\right| \\ &\leq \sum_{w=0,1} t_{w,\alpha}(x).\end{aligned} \tag{4.15}$$

Combining inequality (4.15) and inequality (4.7) in Theorem 4.3.2 (after taking the supremum on both sides with respect δ), we get:

$$\sup_{\delta}\left(\left|\hat{\theta}^L(\delta,x) - \theta^L(\delta,x)\right| \vee \left|\hat{\theta}^U(\delta,x) - \theta^U(\delta,x)\right|\right) \leq \sum_{w=0,1} t_{w,\alpha}(x), \tag{4.16}$$

also. Now, if the inequality in (4.14) holds with high probability (the premise of this corollary), then so must (4.16). □

4.B.5 The proof of Theorem 4.3.4

Proof of Theorem 4.3.4.

Given inequality (4.7) in Theorem 4.3.2, our specification for

$$\hat{G}(y,\delta,x) := \hat{F}_{1n}(y|x) - \hat{F}_{0n}(y|x),$$

and a similar argument to the proof of Corollary 4.3.3 in Appendix 4.B.4, we have that:

$$\sup_{x}\left(\left|\hat{\theta}^L(\delta,x) - \theta^L(\delta,x)\right| \vee \left|\hat{\theta}^U(\delta,x) - \theta^U(\delta,x)\right|\right) \leq \sum_{w=0,1}\sup_{y,x}\left|\hat{F}_{wn}(y|x) - F_w(y|x)\right|. \tag{4.17}$$

Consider $s_0, s_1 \geq 0$. Due to Lemma 4.B.2 and (4.17), we have that:

$$\begin{aligned}
&\Pr\left(\sup_x \left(\left|\hat{\theta}^L(\delta,x) - \theta^L(\delta,x)\right| \vee \left|\hat{\theta}^U(\delta,x) - \theta^U(\delta,x)\right|\right) > s_0 + s_1 \middle| \mathbf{X}\right) \\
&\leq \Pr\left(\sum_{w=0,1} \sup_{y,x} \left|\hat{F}_{wn}(y|x) - F_w(y|x)\right| > s_0 + s_1 \middle| \mathbf{X}\right) \\
&\leq \sum_{w=0,1} \Pr\left(\sup_{y,x} \left|\hat{F}_{wn}(y|x) - F_w(y|x)\right| > s_w \middle| \mathbf{X}\right).
\end{aligned} \tag{4.18}$$

Let $t_w \geq 0$. Moreover, Lemma 4.B.3 tells us that for

$$s_w \doteq \sup_r \{\Pr(r < R_i \leq r + 2t|\mathbf{V}) \vee \Pr(r - 2t < R_i \leq r|\mathbf{V})\} + \left(\frac{\log(4/\alpha)}{2}\right)^{\frac{1}{2}} n_w^{-\frac{1}{2}},$$

we get:

$$\Pr\left(\sup_{y,x} \left|\hat{F}_{wn}(y|x) - F_w(y|x)\right| > s_w \middle| \mathbf{X}\right) \leq \alpha/2 + \Pr\left(\sup_x |\hat{\mu}_w(x) - \mu_w(x)| > t_w \middle| \mathbf{X}\right) \tag{4.19}$$

Now, combining (4.18) and (4.19), we have conditional on $\mathbf{X}$:

$$\begin{aligned}
&\Pr\left(\sup_x \left(\left|\hat{\theta}^L(\delta,x) - \theta^L(\delta,x)\right| \vee \left|\hat{\theta}^U(\delta,x) - \theta^U(\delta,x)\right|\right) > s_0 + s_1 \middle| \mathbf{X}\right) \\
&\leq \alpha + \sum_{w=0,1} \Pr\left(\sup_x |\hat{\mu}_w(x) - \mu_w(x)| > t_w \middle| \mathbf{X}\right).
\end{aligned}$$

This implies the desired conclusion.

□

4.B.6 A lemma for Theorem 4.3.4: conditional CDF estimation with regression residuals

Lemma 4.B.3 (Conditional CDF estimator with regression residuals)**.**

Consider the setup:

- $(U_1, V_1), \ldots, (U_n, V_n)$ *are i.i.d. jointly distributed random variables.*
- *Partition the training indices* $\{1, \ldots, n\}$ *into two non-intersecting splits,* $\mathcal{I}_1$ *and* $\mathcal{I}_2$*, respectively. Let* $\mathcal{T}_j := \{(U_i, V_i)\}_{i \in \mathcal{I}_j}$ *for* $j = 1, 2$*.*

- *Denote* $\mu(v) := \mathbb{E}[U_i|V_i = v]$*, and let* $\hat{\mu}(v)$ *denote its estimator based on* $\mathcal{T}_1$*.*
- *Denote* $R_i := U_i - \mu(V_i)$ *along with its approximation given by* $\hat{R}_i := U_i - \hat{\mu}(V_i)$ *across* $i = 1, 2, \ldots, n$*.*
- *For* $u \in$ *support*(U_i) *and* $v \in$ *support*(V_i)*, denote* $F(u|v) := Pr(U \leq u|V = v)$*. Consider its estimator given by:*

$$\hat{F}_n(u|v) := \frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \mathbf{1}\left\{\hat{\mu}(v) + \hat{R}_i \leq u\right\},$$

where $|\mathcal{I}_2|$ *is the number of indices in* $\mathcal{I}_2$*.*

If $R_i \perp\!\!\!\perp V_i$*, then conditional on* $\mathbf{V} = (V_i; i \in \mathcal{I}_1)$*, we have that for any* $t \geq 0$ *(possibly dependent on* $\mathbf{V}$*):*

$$\sup_{u,v}\left|\hat{F}(u|v) - F(u|v)\right| \;\leq\; \sup_{r}\left\{Pr(r < R_i \leq r + 2t|\mathbf{V}) \vee Pr(r - 2t < R_i \leq r|\mathbf{V})\right\} + \left(\frac{\log(4/\alpha)}{2}\right)^{\frac{1}{2}} |\mathcal{I}_2|^{-\frac{1}{2}}$$

with probability at least

$$1 - \alpha/2 - Pr\left(\sup_{v} |\hat{\mu}(v) - \mu(v)| \geq t \middle| \mathbf{V}\right).$$

Proof.

Below, let the probability statements be with respect to $i \notin \mathcal{I}_1$, where $\mathcal{I}_1$ is the set of indices used to train $\hat{\mu}$. This is important to note as we will use that $R_i \perp\!\!\!\perp (\mathcal{T}_1, \mathbf{V})$ later. Consider that:

$$\begin{aligned}
&\sup_{u,v}\left|\hat{F}_n(u|v) - F(u|v)\right| \\
=\; &\sup_{u,v}\left|\hat{F}_n(u|v) - \Pr\left(\mu(v) + R_i \leq u\right)\right| \\
\leq\; &\sup_{u,v}\left|\hat{F}_n(u|v) - \Pr\left(\hat{\mu}(v) + \hat{R}_i \leq u \middle| \mathcal{T}_1\right)\right| \\
&+ \sup_{u,v}\left|\Pr\left(\hat{\mu}(v) + \hat{R}_i \leq u \middle| \mathcal{T}_1\right) - \Pr\left(\mu(v) + R_i \leq u\right)\right| \\
\overset{(i)}{=}\; &\sup_{u,v}\left|\frac{1}{|\mathcal{I}_2|}\sum_{i \in \mathcal{I}_2} \mathbf{1}\left\{U_i - \hat{\mu}(V_i) \leq u - \hat{\mu}(v)\right\} - \Pr\left(U_i - \hat{\mu}(V_i) \leq u - \hat{\mu}(v)|\mathcal{T}_1\right)\right| \\
&+ \sup_{u,v}\left|\Pr\left(R_i \leq u - \mu(v) + [\hat{b}(V_i) - \hat{b}(v)] \middle| \mathcal{T}_1\right) - \Pr\left(R_i \leq u - \mu(v)\right)\right|.
\end{aligned} \tag{4.20}$$

For any v, define:

$$\hat{b}(v) := \hat{\mu}(v) - \mu(v), \tag{4.21}$$

a term that characterizes the bias in $\hat{\mu}(v)$ for any $v \in \text{support}(V_i)$. In equality (i), we used the identities $\hat{\mu}(v) = \mu(v) + \hat{b}(v)$ and $\hat{R}_i = R_i - \hat{b}(V_i)$, which are a consequence of the definition of $\hat{b}(\cdot)$.

Now, denote:

$$A := \sup_{u,v} \left| \frac{1}{|\mathcal{I}_w|} \sum_{i \in \mathcal{I}_2} \mathbf{1}\left\{U_i - \hat{\mu}(V_i) \leq u - \hat{\mu}(v)\right\} - \Pr\left(U_i - \hat{\mu}(V_i) \leq u - \hat{\mu}(v) | \mathcal{T}_1\right) \right|$$

and

$$B := \sup_{u,v} \left| \Pr\left(R_i \leq u - \mu(v) + [\hat{b}(V_i) - \hat{b}(v)] \middle| \mathcal{T}_1\right) - \Pr\left(R_i \leq u - \mu(v)\right) \right|.$$

Due to Equation (4.20) and Lemma 4.B.2:

$$\begin{aligned} &\Pr\left(\sup_{u,v} \left| \hat{F}_n(u|v) - F(u|v) \right| > a + b \middle| \mathbf{V}\right) \\ &\leq \Pr\left(A + B > a + b | \mathbf{V}\right) \\ &\leq \Pr\left(A > a | \mathbf{V}\right) + \Pr\left(B > b | \mathbf{V}\right). \end{aligned} \tag{4.22}$$

We now control the two terms in the latter inequality separately in § 4.B.6.1 and § 4.B.6.2, respectively. We also explain the choices

$$a \doteq \sqrt{2^{-1}(|\mathcal{I}_2|)^{-1} \log(4/\alpha)}$$

and

$$b \doteq \sup_r \left\{ \Pr\left(r < R_i \leq r + 2t | \mathbf{V}\right) \vee \Pr\left(r - 2t < R_i \leq r | \mathbf{V}\right) \right\}.$$

From these choices, we get the desired conclusion:

$$\sup_{u,v} \left| \hat{F}_n(u|v) - F(u|v) \right| \leq a + b.$$

with probability at least $1 - \alpha/2 - \Pr\left(\sup_v |\hat{\mu}(v) - \mu(v)| | \mathbf{V}\right)$.

4.B.6.1 The term $\Pr(A > a|\mathbf{V})$ in (4.22)

First, notice that:

$$A = \sup_r \left| \frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \mathbf{1}\left\{U_i - \hat{\mu}(V_i) \leq r\right\} - \Pr\left(U_i - \hat{\mu}(V_i) \leq r | \mathcal{T}_1\right) \right|,$$

where the supremum with respect to $r = u - \hat{\mu}(v)$ is taken in the support of $u - \hat{\mu}(v)$.

Also note that $\hat{\mu}(v)$ is random, even if v is not, because it is an estimator based on $\mathcal{T}_1$. However, conditional on $(\mathcal{T}_1, \mathbf{V})$, $r = u - \hat{\mu}(v)$ is a constant. Importantly, conditional on $(\mathcal{T}_1, \mathbf{V})$, we have that $\hat{R}_i = U_i - \hat{\mu}(V_i)$ is an independent and identically distributed random variable across $i \notin \mathcal{I}_1$. This means that the estimator

$$\frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \mathbf{1}\left\{U_i - \hat{\mu}(V_i) \leq r\right\}$$

for

$$\Pr\left(U_i - \hat{\mu}(V_i) \leq r | \mathcal{T}_1, \mathbf{V}\right)$$

satisfies the i.i.d. sample condition for the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality (Dvoretzky et al. 1956, Massart 1990, Naaman 2021). The DKW inequality gives:

$$\begin{aligned}
&\Pr\left(A > a | \mathcal{T}_1, \mathbf{V}\right) \\
&= \Pr\left(\sup_r \left| \frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \mathbf{1}\left\{U_i - \hat{\mu}(V_i) \leq r\right\} - \Pr\left(U_i - \hat{\mu}(V_i) \leq r | \mathcal{T}_1\right) \right| > a \middle| \mathcal{T}_1, \mathbf{V} \right) \\
&\leq 2e^{-2|\mathcal{I}_2|a^2}.
\end{aligned}$$

Noting that the upper bound does not depend on $(\mathcal{T}_1, \mathbf{V})$, we get due to law of total expectation that:

$$\Pr\left(A > a | \mathbf{V}\right) \leq 2e^{-2|\mathcal{I}_2|a^2}$$

and

$$\Pr\left(A > a\right) \leq 2e^{-2|\mathcal{I}_2|a^2}.$$

We will take

$$a \doteq \sqrt{2^{-1}(|\mathcal{I}_2|)^{-1}\log(4/\alpha)},$$

so that we are guaranteed $\Pr(A > a) \leq \alpha/2$ and $\Pr(A > a|\mathbf{V}) \leq \alpha/2$.

4.B.6.2 The term $\Pr(B > b|\mathbf{V})$ in (4.22)

We note that

$$B := \sup_{u,v} \left| \Pr\left(R_i \leq u - \mu(v) + [\hat{b}(V_i) - \hat{b}(v)] \middle| \mathcal{T}_1\right) - \Pr(R_i \leq u - \mu(v)) \right|$$

is a random variable with respect to $\mathcal{T}_1$. For t, possibly a function of $\mathbf{V}$, consider the following event with respect to the random data in $\mathcal{T}_1$:

$$E(t) := \left\{ \sup_v |\hat{\mu}(v) - \mu(v)| \leq t \right\}.$$

The event $E(t)$ concerns the deviation between $\hat{\mu}$ and μ uniformly across the support of V_i. Conditional on the event $E(t)$ and $\mathbf{V}$, we have that the regression bias term defined in Equation (4.21) satisfies $\hat{b}(V_i) \in [-t, t]$, in spite of V_i being random, along with $\hat{b}(v) \in [-t, t]$ for any v. This means that conditionally on $(E(t), \mathbf{V})$:

$$\hat{b}(V_i) - \hat{b}(v) \in [-2t, 2t] \tag{4.23}$$

almost surely.

Now let $E(t)^C$ denote the complement of $E(t)$. Consider that:

$$\begin{aligned} \Pr(B > b|\mathbf{V}) = &\quad \Pr(B > b|E(t), \mathbf{V}) \Pr(E(t)|\mathbf{V}) + \Pr\left(B > b \middle| E(t)^C, \mathbf{V}\right) \Pr\left(E(t)^C \middle| \mathbf{V}\right) \\ \leq &\quad \Pr(B > b|E(t), \mathbf{V}) + \Pr\left(\sup_v |\hat{\mu}(v) - \mu(v)| > t \middle| \mathbf{V}\right). \end{aligned} \tag{4.24}$$

This is by the law of total probability, the fact that $\Pr(\cdot|\mathbf{V}), \Pr(\cdot|E(t), \mathbf{V}) \in [0, 1]$, along with the definition of $E(t)^C$.

For a random event $D \perp\!\!\!\perp R_i$ (recall, $i \notin \mathcal{I}_1$), consider the function:

$$\nu(r, s, D, \mathbf{V}) := |\Pr(R_i \leq r + s|D, \mathbf{V}) - \Pr(R_i \leq r)|.$$

For $D = \emptyset$, we write:

$$\nu(r, s, \emptyset, \mathbf{V}) := \left| \Pr\left(R_i \leq r + s | \mathbf{V}\right) - \Pr\left(R_i \leq r\right) \right|.$$

Fixing $(r, D, \mathbf{V})$, we can see that $\nu(r, s, D, \mathbf{V})$ is an increasing function with respect to s in the domain $[0, 2t]$, and it is a decreasing function with respect to s in the domain $[-2t, 0)$. From this, it follows that for all $s \in [-2t, 2t]$:

$$\nu(r, s, D, \mathbf{V}) \ \leq \ \nu(r, -2t, D, \mathbf{V}) \vee \nu(r, 2t, D, \mathbf{V}).$$

Based on Equation (4.23) and this property of $\nu(r, s, D, \mathbf{V})$, it follows that conditional on $E(t)$ and $\mathbf{V}$ we have:

$$B \in \left[0, \sup_{u,v} \left\{\nu(u - \mu(v), -2t, E(t), \mathbf{V}) \vee \nu(u - \mu(v), 2t, E(t), \mathbf{V})\right\}\right]$$

almost surely. We have further that

$$\begin{aligned} &\sup_{r} \left\{\nu(r, -2t, \emptyset, \mathbf{V}) \vee \nu(r, 2t, \emptyset, \mathbf{V})\right\} \\ = \ &\sup_{u,v} \left\{\nu(u - \mu(v), -2t, E(t), \mathbf{V}) \vee \nu(u - \mu(v), 2t, E(t), \mathbf{V})\right\}. \end{aligned}$$

This is based on the properties of the supremum, the fact that t is fixed conditional on $\mathbf{V}$, and because $R_i \perp\!\!\!\perp E(t)$. So we can re-write that conditional on $E(t)$ and $\mathbf{V}$:

$$B \in \left[0, \sup_{r} \left\{\nu(r, -2t, \emptyset, \mathbf{V}) \vee \nu(r, 2t, \emptyset, \mathbf{V})\right\}\right]$$

almost surely. We will strategically set:

$$b \doteq \sup_{r} \left\{\nu(r, -2t, \emptyset, \mathbf{V}) \vee \nu(r, 2t, \emptyset, \mathbf{V})\right\}.$$

Moreover, $R_i \perp\!\!\!\perp \mathbf{V}$ means that $\Pr\left(R_i \leq r\right) = \Pr\left(R_i \leq r | \mathbf{V}\right)$, so we have:

$$b \ = \ \sup_{r} \left\{\Pr\left(r < R_i \leq r + 2t | \mathbf{V}\right) \vee \Pr\left(r - 2t < R_i \leq r | \mathbf{V}\right)\right\}.$$

With this choice of b, we have that $\Pr\left(B > b | E(t), \mathbf{V}\right) = 0$. Using this in (4.24), we get that:

$$\Pr\left(B > b | \mathbf{V}\right) \leq \Pr\left(\sup_{v} |\hat{\mu}(v) - \mu(v)| > t \middle| \mathbf{V}\right),$$

as we wanted.

□

4.B.7 The proof of Proposition 4.3.7

Proof of Proposition 4.3.7.

Let $\mathbf{\Lambda}_w \in \mathbb{R}^{|\mathcal{I}_1 \cap S_w| \times d}$ be such that its rows are formed by stacking $\Psi_w(X_i)^T$ for each $i \in \mathcal{I}_1 \cap S_w$. Further, let $\mathbf{Y}_w \in \mathbb{R}^{|\mathcal{I}_1 \cap S_w| \times 1}$ contain the corresponding observed outcome Y_i for each $i \in \mathcal{I}_1 \cap S_w$. The ordinary least squares estimator for the coefficient vector β_w is given by

$$\hat{\beta}_w = (\mathbf{\Lambda}_w^T \mathbf{\Lambda}_w)^{-1} \mathbf{\Lambda}_w^T \mathbf{Y}_w.$$

We have that

$$\begin{aligned}
\sup_x |\hat{\mu}_w(x) - \mu_w(x)| &= \sup_x \left| \Psi_w^T(x)(\hat{\beta}_w - \beta_w) \right| \\
&\leq \left\| \hat{\beta}_w - \beta_w \right\|_2 \\
&\leq \left\| \left| \sigma_w (\mathbf{\Lambda}_w^T \mathbf{\Lambda}_w)^{-1/2} \right| \right\|_{op} \left\| \frac{1}{\sigma_w} (\mathbf{\Lambda}_w^T \mathbf{\Lambda}_w)^{1/2} (\hat{\beta}_w - \beta_w) \right\|_2 .
\end{aligned}$$

The first inequality is based on Hölder's inequality, while the second inequality is again due to Hölder's inequality and the definition of the operator norm. Conditional on $\mathbf{\Lambda}_w$, the d-dimensional sampling distribution for $\hat{\beta}$ is:

$$\hat{\beta}_w | \mathbf{\Lambda}_w \sim \mathcal{N}(\beta_w, \sigma_w^2 (\mathbf{\Lambda}_w^T \mathbf{\Lambda}_w)^{-1}).$$

This is based on a standard argument in low dimensional linear regression given the residual distribution assumption. This further implies that conditionally on $\mathbf{\Lambda}_w$,

$$\left\| \left| \sigma_w (\mathbf{\Lambda}_w^T \mathbf{\Lambda}_w)^{-1/2} \right| \right\|_{op}^2 \left\| \frac{1}{\sigma_w} (\mathbf{\Lambda}_w^T \mathbf{\Lambda}_w)^{1/2} (\hat{\beta}_w - \beta_w) \right\|_2^2 \Bigg| \mathbf{\Lambda}_w \sim \sigma_w^2 \left\| \left| (\mathbf{\Lambda}_w^T \mathbf{\Lambda}_w)^{-1/2} \right| \right\|_{op}^2 \times \chi_d^2,$$

a distribution generated by taking a chi-squared distributed random variable (d degrees of freedom) and multiplying it by a factor of $\sigma_w^2 \left\| \left| (\mathbf{\Lambda}_w^T \mathbf{\Lambda}_w)^{-1/2} \right| \right\|_{op}^2$. This is because

$$\frac{1}{\sigma_w} (\mathbf{\Lambda}_w^T \mathbf{\Lambda}_w)^{1/2} (\hat{\beta}_w - \beta_w) \Big| \mathbf{\Lambda}_w \sim \mathcal{N}_d(\mathbf{0}, \mathbb{I}) \implies \left\| \frac{1}{\sigma_w} (\mathbf{\Lambda}_w^T \mathbf{\Lambda}_w)^{1/2} (\hat{\beta}_w - \beta_w) \right\|_2^2 \Bigg| \mathbf{\Lambda}_w \sim \chi_d^2.$$

Let V have a χ^2_d distribution conditional on $\mathbf{\Lambda}_w$, and denote $v_{d,\alpha}$ as the $(1-\alpha/2)^{th}$ quantile of V's distribution. It follows that setting

$$t_{w,\alpha} \doteq \sqrt{v_{d,\alpha}}\sigma_w \left|\left|\left|(\mathbf{\Lambda}_w^T\mathbf{\Lambda}_w)^{-1/2}\right|\right|\right|_{op}$$

implies that

$$\begin{aligned} 2\max_{w=0,1}\Pr\left(\sup_x|\hat{\mu}_w(x)-\mu_w(x)|>t_{w,\alpha}\middle|\mathbf{X}\right) &\leq 2\max_{w=0,1}\Pr\left(V>\frac{t^2_{w,\alpha}}{\sigma^2_w\left|\left|\left|(\mathbf{\Lambda}_w^T\mathbf{\Lambda}_w)^{-1/2}\right|\right|\right|^2_{op}}\middle|\mathbf{\Lambda}_w\right) \\ &= \alpha. \end{aligned} \tag{4.25}$$

Now, consider that the distributional assumption on $R_i(w)$ implies the following. We have:

$$\begin{aligned} & 2\max_{w=0,1}\sup_r\Pr\left(r<R_i(w)\leq r+2t_{w,\alpha}|\mathbf{\Lambda}_w\right)\vee\Pr\left(r-2t_{w,\alpha}<R_i(w)\leq r|\mathbf{\Lambda}_w\right) \\ \overset{(i)}{=}\ & 2\max_{w=0,1}\sup_r\Pr\left(r<R_i(w)\leq r+2t_{w,\alpha}|\mathbf{\Lambda}_w\right) \\ \overset{(ii)}{=}\ & 2\max_{w=0,1}\Pr\left(-t_{w,\alpha}<R_i(w)\leq t_{w,\alpha}|\mathbf{\Lambda}_w\right) \\ \overset{(iii)}{=}\ & 2\max_{w=0,1}\Pr\left(-\sqrt{v_{d,\alpha}}\left|\left|\left|(\mathbf{\Lambda}_w^T\mathbf{\Lambda}_w)^{-1/2}\right|\right|\right|_{op}<R_i(w)/\sigma_w\leq\sqrt{v_{d,\alpha}}\left|\left|\left|(\mathbf{\Lambda}_w^T\mathbf{\Lambda}_w)^{-1/2}\right|\right|\right|_{op}\middle|\mathbf{\Lambda}_w\right) \\ \overset{(iv)}{=}\ & 2\max_{w=0,1}\left\{\Phi\left(\sqrt{v_{d,\alpha}}\left|\left|\left|(\mathbf{\Lambda}_w^T\mathbf{\Lambda}_w)^{-1/2}\right|\right|\right|_{op}\right)-\Phi\left(-\sqrt{v_{d,\alpha}}\left|\left|\left|(\mathbf{\Lambda}_w^T\mathbf{\Lambda}_w)^{-1/2}\right|\right|\right|_{op}\right)\right\}. \end{aligned} \tag{4.26}$$

Here, Φ denotes the CDF for the standard normal distribution. Equality (i) holds due to the symmetry of the Gaussian density around its mean, which is zero in the case of $R_i(w)$ conditional on Λ_w. Moreover, (ii) holds because the biggest slice of area under the normal density of width $2t_{w,\alpha}$ is the one centered at its mean. Next, (iii) holds due to our choice of $t_{w,\alpha}$, while (iv) holds since $R_i(w)/\sigma_w$ follows a standard normal distribution.

The final conclusion in Proposition 4.3.7 follows by applying Theorem 4.3.4 with (4.25) and (4.26).

□

4.B.8 Proof of Proposition 4.4.1

Proof of Proposition 4.4.1.

The first claim is immediate from Lemma 4.A.1 and our definition of $\theta^L(\delta)$ and $\theta^U(\delta)$ in Section 4.2. We also use the fact that PNS is the same as PIBT when $\delta \in [0,1)$.

We now show why the second part of our claim holds. We will use that CDFs for binary potential outcomes satisfy:

$$F_w(y) = \begin{cases} 0 & \text{if } y < 0 \\ \Pr\left(Y_i(w) = 0\right) & \text{if } 0 \leq y < 1 \\ 1 & \text{if } y \geq 1 \end{cases}; w = 0, 1.$$

- For $\delta \in [0,1)$, the Makarov lower bound on PIBT is:

$$\begin{aligned} \theta^L(\delta) &= -\min\left(\inf_{y=0,1}\left\{F_1(y+\delta/2) - F_0(y-\delta/2)\right\}, 0\right) \\ &= -\min\left(F_1(\delta/2) - F_0(-\delta/2), F_1(1+\delta/2) - F_0(1-\delta/2), 0\right) \\ &= \begin{cases} -\min\left(\Pr\left(Y_i(1)=0\right) - \Pr\left(Y_i(0)=0\right), 0\right) & \text{if } \delta = 0 \\ -\min\left(\Pr\left(Y_i(1)=0\right), 1 - \Pr\left(Y_i(0)=0\right), 0\right) & \text{if } 0 < \delta < 1 \end{cases} \\ &= \begin{cases} -\min\left(1 - \Pr\left(Y_i(1)=1\right) - \Pr\left(Y_i(0)=0\right), 0\right) & \text{if } \delta = 0 \\ 0 & \text{if } 0 < \delta < 1 \end{cases} \\ &= \begin{cases} \max\left(\Pr\left(Y_i(1)=1\right) + \Pr\left(Y_i(0)=0\right) - 1, 0\right) & \text{if } \delta = 0 \\ 0 & \text{if } 0 < \delta < 1 \end{cases}. \end{aligned}$$

  It follows that

$$\sup_{\delta \in [0,1)} \theta^L(\delta) = \max\left(\Pr\left(Y_i(1)=1\right) + \Pr\left(Y_i(0)=0\right) - 1, 0\right),$$

  as we wanted.

- Similarly, for $\delta \in [0,1)$, the Makarov upper bound on PIBT is:

$$\begin{aligned}
\theta^U(\delta) &= 1 - \max\left(\sup_{y=0,1}\left\{F_1(y+\delta/2) - F_0(y-\delta/2)\right\}, 0\right) \\
&= 1 - \max\left(F_1(\delta/2) - F_0(-\delta/2), F_1(1+\delta/2) - F_0(1-\delta/2), 0\right) \\
&= \begin{cases} 1 - \max\left(\Pr\left(Y_i(1)=0\right) - \Pr\left(Y_i(0)=0\right), 0\right) & \text{if } \delta = 0 \\ 1 - \max\left(\Pr\left(Y_i(1)=0\right), 1 - \Pr\left(Y_i(0)=0\right)\right) & \text{if } 0 < \delta < 1 \end{cases} \\
&= \begin{cases} 1 - \max\left(1 - \Pr\left(Y_i(1)=1\right) - \Pr\left(Y_i(0)=0\right), 0\right) & \text{if } \delta = 0 \\ 1 - \max\left(\Pr\left(Y_i(1)=0\right), \Pr\left(Y_i(0)=1\right)\right) & \text{if } 0 < \delta < 1 \end{cases} \\
&= \begin{cases} \min\left(\Pr\left(Y_i(1)=1\right) + \Pr\left(Y_i(0)=0\right), 1\right) & \text{if } \delta = 0 \\ \min\left(\Pr\left(Y_i(1)=1\right), \Pr\left(Y_i(0)=0\right)\right) & \text{if } 0 < \delta < 1 \end{cases}
\end{aligned}$$

  It follows that

$$\inf_{\delta \in [0,1)} \theta^U(\delta) = \min\left(\Pr\left(Y_i(1)=1\right), \Pr\left(Y_i(0)=0\right)\right),$$

  as we wanted.

With essentially the same reasoning, we have the analogous claims for

$$\Pr\left(Y_i(1)=1, Y_i(0)=0 | X_i = x\right).$$

□

CHAPTER 5

Summary and Possible Extensions

In Chapter 2, we developed a novel structure learning method that allows us to estimate the topological ordering of a large number of variables in an unknown causal graph. A key strength is the computational scalability to a large number of variables and the identifiability guarantee. The key limitation is the assumed structure: the linear relation between variables. Further study is needed to understand whether incorporating a larger amount of variables can make it so that, generally, any given variable is well explained. For example, this can be understood by calculating coefficients of determination as in the real data application of Section 2.3.3.2.

Along the lines of calculating coefficients of determination, it could be interesting to study further the relative goodness of fit to our data between two causal discovery models (see for example Ramsey et al. (2020)). Concretely, consider that for any arbitrary random vector $X \sim f(x)$, we may define the cross entropy (Cover and Thomas 2005) of a specified probability density function $g(x)$ as the negative mean log-likelihood:

$$\mathbf{H}(f, g) := -\int_x \log g(x) f(x) dx.$$

The minimizer of $\mathbf{H}(f, g)$ with respect to g is f (Cover and Thomas 2005). Given that in practice specifying f correctly is difficult, if not impossible, consider two candidate densities $g_1(x)$ and $g_2(x)$ corresponding to two structural equation models (SEMs) output by differing causal discovery models. At the limit of sample size, we may prefer to specify density g_1 for X over g_2 if

$$\mathbf{H}(f, g_1) \leq \mathbf{H}(f, g_2).$$

This inequality says that g_1 is a closer fit to the distribution of X compared to g_2. As an example of how one may reason about which model provides a better cross-entropy in practice, consider specifying Gaussian noise for both a nonlinear SEM (e.g. as estimable by Peters et al. (2014) or Gao et al. (2020)) and for a linear SEM (e.g. as estimable by Aragam and Zhou (2015) or Ye et al. (2021a)). We may be more willing to live with the nonlinear specification if the negative sample mean of the log-likelihood specification is better with its specification vs. under the linear specification that does not allow us a unique SEM (Ye et al. 2021a). The negative sample mean of the log-likelihood, an approximation to the cross entropy $\mathbf{H}(f, g_j)$ $(j = 1, 2)$, can be calculated on a validation fold as in the model comparison of Section 2.3.3.1. More formally, considering that parameter degrees of freedom should also be taken into account, it can be of interest to build from the literature on Akaike Information Criterion (AIC) and other information criterion rules for model selection (Stoica and Selen 2004). It could also be of interest to look into formal statistical tests for comparing non-nested models, such as based on the Vuong statistic (Vuong 1989). Moreover, it can be of interest to incorporate recent work that cautions about the purported benefit of uniquely identifiable graphical models and instead vouches for specifying a general likelihood that encodes an equivalence class of Bayesian networks (Shpitser 2022).

Moreover, Chapter 3 provides finite sample and asymptotic guarantees for the learning of the linear structural equation model discussed in Chapter 2. Given the importance of applying theoretically justified methods in practice, the results of this chapter help provide a reasonable expectation for a causal discovery method. Extensions of the work here include understanding further the sensitivity of the accuracy results to the density specification (e.g. Laplace) vs. the actual density which can be power law decaying, exponentially decaying, or something else. Moreover, it would be interesting to see whether the argument of the estimation theory in this chapter can be generalized to more sequential approaches to estimate a permutation of variables based on some other heuristic, such as tailored to a non-linear SEM specification.

In Chapter 4, we developed novel statistical estimation theory that allows us to reason in a frequentist sense about the probability an individual benefits from treatment–an inestimable parameter. A key component of our approach is the use distribution-free bounds on the parameter of interest along with non-asymptotic concentration inequalities. It could be interesting to study how conservative this approach is, if at all. For example, in the case that the potential outcomes' residuals are jointly Gaussian, appropriate bounds on PIBT might be obtained by simply tuning the correlation between the potential outcomes' residuals (recall the example in Equation (1.1) of Chapter 1). As with the case of jointly Gaussian potential outcomes' residuals, we may like to study other structural assumptions to obtain estimators for bounds on PIBT. For example, we may like to develop results for generalized linear models (McCullagh and Nelder 2019) in analogy to Proposition 4.3.7, including involving logistic or probit regression for binary potential outcomes. Additionally, extensions of the theory in this chapter may like to incorporate the assumption of positively correlated potential outcomes–an assumption which may be realistic in practice (Frandsen and Lefgren 2021). Moreover, Section 4.5 of this chapter, in an application to a large randomized experiment dataset, demonstrates how our proposed methodology can help augment existing approaches for average causal effect modeling. Along these lines, and as alluded to in Chapter 1, it seems interesting to study further the practical implications of seemingly significant, possibly heterogeneous, average treatment effects. We may also like to study whether a connection exists between the counterfactuals involved in reasoning about unconfounded (identifiable), possibly path specific, effects (Malinsky and Spirtes 2017, Malinsky et al. 2019) and a parameter similar to Definition 4.1.1 (PIBT) in Chapter 4.

Bibliography

Abadie, A., Chingos, M. M., and West, M. R. (2018). Endogenous Stratification in Randomized Experiments. *The Review of Economics and Statistics*, 100(4):567–580.

Aragam, B., Gu, J., and Zhou, Q. (2019). Learning large-scale bayesian networks with the sparsebn package. *Journal of Statistical Software*, 91(11):1–38.

Aragam, B. and Zhou, Q. (2015). Concave penalized estimation of sparse gaussian bayesian networks. *Journal of Machine Learning Research*, 16(69):2273–2328.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148 – 1178.

Austin, P. C. (2014). The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Stat. Med.*, 33(7):1242–1258.

Bickel, P. J., Götze, F., and van Zwet, W. R. (1997). Resampling fewer than n observations: Gains, losses, and remedies for losses. *Statistica Sinica*, 7(1):1–31.

Boole, G. (1854). *An investigation of the laws of thought : on which are founded the mathematical theories of logic and probabilities / By George Boole*. Walton and Maberly, London.

Bühlmann, P., Peters, J., and Ernest, J. (2014). Cam: Causal additive models, high-dimensional order search and penalized regression. *Ann. Statist.*, 42(6):2526–2556.

Burkhart, M. C. and Ruiz, G. (2022). Neuroevolutionary feature representations for causal inference. In Groen, D., de Mulatier, C., Paszynski, M., Krzhizhanovskaya, V. V., Dongarra, J. J., and Sloot, P. M. A., editors, *Computational Science – ICCS 2022*, pages 3–10, Cham. Springer International Publishing.

Cai, W. and van der Laan, M. J. (2020). One-step targeted maximum likelihood estimation for time-to-event outcomes. *Biometrics*, 76(3):722–733.

Caughey, D., Dafoe, A., Li, X., and Miratrix, L. (2021). Randomization inference beyond the sharp null: Bounded null hypotheses and quantiles of individual treatment effects.

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

Chernozhukov, V., Fernández-Val, I., and Melly, B. (2013). Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268.

Chickering, D. (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554.

Chickering, D. M. (1996). *Learning Bayesian Networks is NP-Complete*, pages 121–130. Springer New York, New York, NY.

Cinelli, C. and Pearl, J. (2021). Generalizing experimental results by leveraging knowledge of mechanisms. *Eur. J. Epidemiol.*, 36(2):149–164.

Cover, T. M. and Thomas, J. A. (2005). *Inequalities in Information Theory*, chapter 17, pages 657–687. John Wiley & Sons, Ltd.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.

Diemert Eustache, Betlei Artem, Renaudin, C., and Massih-Reza, A. (2018). A large scale benchmark for uplift modeling. In *Proceedings of the AdKDD and TargetAd Workshop, KDD, London,United Kingdom, August, 20, 2018*. ACM.

Ding, P., Feller, A., and Miratrix, L. (2019). Decomposing treatment effect variation. *Journal of the American Statistical Association*, 114(525):304–317.

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956). Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator. *The Annals of Mathematical Statistics*, 27(3):642 – 669.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Chapman and Hall/CRC.

Fan, Y. and Park, S. S. (2010). Sharp bounds on the distribution of treatment effects and their statistical inference. *Econometric Theory*, 26(3):931–951.

Fan, Y. and Park, S. S. (2012). Confidence intervals for the quantile of treatment effects in randomized experiments. *Journal of Econometrics*, 167(2):330–344. Fourth Symposium on Econometric Theory and Applications (SETA).

Fay, M. P., Brittain, E. H., Shih, J. H., Follmann, D. A., and Gabriel, E. E. (2018). Causal estimands and confidence intervals associated with wilcoxon-mann-whitney tests in randomized experiments. *Statistics in Medicine*, 37(20):2923–2937.

Fay, M. P. and Proschan, M. A. (2010). Wilcoxon-Mann-Whitney or t-test? on assumptions for hypothesis tests and multiple interpretations of decision rules. *Stat. Surv.*, 4:1–39.

Firpo, S. and Ridder, G. (2019). Partial identification of the treatment effect distribution and its functionals. *Journal of Econometrics*, 213(1):210–234. Annals: In Honor of Roger Koenker.

Firpo, S. P. and Ridder, G. (2010). Bounds on functionals of the distribution treatment effects. Textos para discussão 201, FGV EESP - Escola de Economia de São Paulo, Fundação Getulio Vargas (Brazil).

Fisher, R. A. (1935). *The design of experiments*. Oliver and Boyd, Edinburgh.

Forbes, C., Evans, M., Hastings, N., and Peacock, B. (2010). *Statistical Distributions*. Wiley-Blackwell, Hoboken, NJ, 4 edition.

Frandsen, B. R. and Lefgren, L. J. (2021). Partial identification of the distribution of treatment effects with an application to the knowledge is power program (kipp). *Quantitative Economics*, 12(1):143–171.

Frank, M. J., Nelsen, R. B., and Schweizer, B. (1987). Best-possible bounds for the distribution of a sum — a problem of kolmogorov. *Probability Theory and Related Fields*, 74:199–211.

Fréchet, M. (1935). Généralisation du théorème des probabilités totales. *Fundamenta mathematicae*, 25:379–387.

Fréchet, M. (1960). Sur les tableaux dont les marges et des bornes sont données. *Revue de l'Institut international de statistique*, 28(1/2):10–32.

Frydenberg, M. (1990). The chain graph markov property. *Scandinavian Journal of Statistics*, 17:333–353.

Gao, M., Ding, Y., and Aragam, B. (2020). A polynomial-time algorithm for learning nonparametric causal graphs. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11599–11611. Curran Associates, Inc.

Ghoshal, A. and Honorio, J. (2017). Learning identifiable gaussian bayesian networks in polynomial time and sample complexity. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6460–6469, Red Hook, NY, USA. Curran Associates Inc.

Glymour, C., Zhang, K., and Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10:524.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Greenland, S., Fay, M. P., Brittain, E. H., Shih, J. H., Follmann, D. A., Gabriel, E. E., and Robins, J. M. (2020). On causal inferences for personalized medicine: How hidden causal assumptions led to erroneous causal claims about the d-value. *Am. Stat.*, 74(3):243–248.

Hailperin, T. (1986). *Boole's logic and probability : a critical exposition from the standpoint of contemporary algebra, logic, and probability theory / Theodore Hailperin.* Studies in logic and the foundations of mathematics ; v. 85. North-Holland Pub. Co., Amsterdam, Netherlands ;, 2nd ed., rev. and enl. edition.

Hand, D. J. (1992). On comparing two treatments. *The American Statistician*, 46(3):190–192.

Henckel, L., Perkovi'c, E., and Maathuis, M. (2019). Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *arXiv: Statistics Theory*.

Hernán, M. A. and Robins, J. (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.

Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Networks*, 13(4):411–430.

Hyvärinen, A. and Smith, S. M. (2013). Pairwise likelihood ratios for estimation of non-gaussian structural equation models. *J. Mach. Learn. Res.*, 14(1):111–152.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Jin, Y., Ren, Z., and Candès, E. J. (2021). Sensitivity analysis of individual treatment effects: A robust conformal inference approach.

Kallus, N. (2022). What's the harm? sharp bounds on the fraction negatively affected by treatment.

Kennedy, E. H. (2020). Towards optimal doubly robust estimation of heterogeneous causal effects.

Kneib, T., Silbersdorff, A., and Säfken, B. (2021). Rage against the mean – a review of distributional regression approaches. *Econometrics and Statistics*.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.

Koenker, R., Leorato, S., and Peracchi, F. (2013). Distributional vs. quantile regression.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165.

Lehmann, E. L. (1966). Some concepts of dependence. *The Annals of Mathematical Statistics*, 37(5):1137–1153.

Lei, L. and Candès, E. J. (2021). Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B*, 83(5):911–938.

Li, J. J., Chen, Y. E., and Tong, X. (2021). A flexible model-free prediction-based framework for feature ranking. *Journal of Machine Learning Research*, 22(124):1–54.

Lu, J., Ding, P., and Dasgupta, T. (2015). Treatment effects on ordinal outcomes: Causal estimands and sharp bounds. *Journal of Educational and Behavioral Statistics*, 43:540 – 567.

Maathuis, M., Colombo, D., Kalisch, M., and Bühlmann, P. (2010). Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7:247–248.

Maathuis, M. H., Kalisch, M., and Bühlmann, P. (2009). Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133 – 3164.

Makarov, G. D. (1982). Estimates for the distribution function of a sum of two random variables when the marginal distributions are fixed. *Theory of Probability & Its Applications*, 26(4):803–806.

Malinsky, D., Shpitser, I., and Richardson, T. (2019). A potential outcomes calculus for identifying conditional path-specific effects. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of*

the Twenty-Second International Conference on Artificial Intelligence and Statistics, volume 89 of *Proceedings of Machine Learning Research*, pages 3080–3088. PMLR.

Malinsky, D. and Spirtes, P. (2017). Estimating bounds on causal effects in high-dimensional and possibly confounded systems. *International Journal of Approximate Reasoning*, 88:371–384.

Mann, H. B. and Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50 – 60.

Massart, P. (1990). The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality. *The Annals of Probability*, 18(3):1269 – 1283.

McCullagh, P. and Nelder, J. A. (2019). *Generalized Linear Models*. Routledge.

McDavid, A., Gottardo, R., Simon, N., and Drton, M. (2019). Graphical models for zero-inflated single cell gene expression. *The Annals of Applied Statistics*, 13(2):848 – 873.

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

Mueller, S. and Pearl, J. (2019). Fréchet inequalities – visualization, applications, and history.

Naaman, M. (2021). On the tight constant in the multivariate dvoretzky–kiefer–wolfowitz inequality. *Statistics & Probability Letters*, 173:109088.

Nandy, P., Maathuis, M. H., and Richardson, T. S. (2017). Estimating the effect of joint interventions from observational data in sparse high-dimensional settings. *Ann. Statist.*, 45(2):647–674.

Nie, X. and Wager, S. (2020). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319.

Padilla, O. H. M., Ding, P., Chen, Y., and Ruiz, G. (2021). A causal fused lasso for interpretable heterogeneous treatment effects estimation.

Park, G. (2020). Identifiability of additive noise models using conditional variances. *Journal of Machine Learning Research*, 21(75):1–34.

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

Pearl, J. (1999). Probabilities of causation: Three counterfactual interpretations and their identification. *Synthese*, 121(1):93–149.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition.

Pearl, J., Glymour, M., and Jewell, N. P. (2016). *Causal inference in statistics: a primer*. Chichester, West Sussex, UK : John Wiley & Sons Ltd.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. (2014). Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(58):2009–2053.

Ramsey, J. D., Malinsky, D., and Bui, K. V. (2020). Algcomparison: Comparing the performance of graphical structure learning algorithms with tetrad. *J. Mach. Learn. Res.*, 21(1).

Raskutti, G. and Uhler, C. (2018). Learning directed acyclic graph models based on sparsest permutations. *Stat*, 7(1):e183. e183 sta4.183.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica*, 56:931–954.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701.

Ruiz, G. and Padilla, O. H. M. (2022). Non-asymptotic confidence bands on the probability an individual benefits from treatment (pibt).

Ruiz, G., Padilla, O. H. M., and Zhou, Q. (2022). Sequentially learning the topological ordering of causal directed acyclic graphs with likelihood ratio scores.

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M. D., Muñoz-Marí, J., van Nes, E. H., Peters, J., Quax, R., Reichstein, M., Scheffer, M., Schölkopf, B., Spirtes, P., Sugihara, G., Sun, J., Zhang, K., and Zscheischler, J. (2019). Inferring causation from time series in earth system sciences. *Nature Communications*, 10(1).

Rüschendorf, L. (1981). Sharpness of fréchet-bounds. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 57:293–302.

Saleeba, C., Dempsey, B., Le, S., Goodchild, A., and McMullan, S. (2019). A student's guide to neural circuit tracing. *Frontiers in Neuroscience*, 13:897.

Sani, N., Malinsky, D., and Shpitser, I. (2020). Explaining the behavior of black-box prediction algorithms with causal learning. *CoRR*, abs/2006.02482.

Sarkar, A. and Stephens, M. (2021). Separating measurement and expression models clarifies confusion in single-cell rna sequencing analysis. *Nature genetics*, 53(6):770—777.

Schober, P. and Vetter, T. R. (2018). Survival analysis and interpretation of time-to-event data: The tortoise and the hare. *Anesth. Analg.*, 127(3):792–798.

Shafer, G. and Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12):371–421.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. (2006). A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(72):2003–2030.

Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., Hoyer, P. O., and Bollen, K. (2011). Directlingam: A direct method for learning a linear non-gaussian structural equation model. *J. Mach. Learn. Res.*, 12(null):1225–1248.

Shpitser, I. (2022). The lauritzen-chen likelihood for graphical models.

Solus, L., Wang, Y., and Uhler, C. (2021). Consistency guarantees for greedy permutation-based causal inference algorithms. *Biometrika*, 108(4):795–814.

Spirtes, P. and Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):62–72.

Spirtes, P., Glymour, C., Scheines, R., Kauffman, S., Aimale, V., and Wimberly, F. (2000). Constructing bayesian network models of gene expression networks from microarray data.

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. P. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statist. Sci.*, 5(4):465–472.

Stekhoven, D. J., Moraes, I., Sveinbjörnsson, G., Hennig, L., Maathuis, M. H., and Bühlmann, P. (2012). Causal stability ranking. *Bioinformatics*, 28(21):2819–2823.

Stitelman, O. M. and van der Laan, M. J. (2010). Collaborative targeted maximum likelihood for time to event data. *Int. J. Biostat.*, 6(1):Article 21.

Stoica, P. and Selen, Y. (2004). Model-order selection: a review of information criterion rules. *IEEE Signal Processing Magazine*, 21(4):36–47.

Sur, P. and Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525.

Tian, J. and Pearl, J. (2000). Probabilities of causation: Bounds and identification. *Ann. Math. Artif. Intell.*, 28(1/4):287–313.

Tibshirani, R. J., Barber, R. F., Candès, E. J., and Ramdas, A. (2019). *Conformal Prediction under Covariate Shift*. Curran Associates Inc., Red Hook, NY, USA.

Tong, X., Feng, Y., and Li, J. J. (2018). Neyman-pearson classification algorithms and np receiver operating characteristics. *Science Advances*, 4(2):eaao1659.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer New York, New York, NY.

Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. *Probabilistic and Causal Inference*.

Verma, T. and Pearl, J. (2022). *Equivalence and Synthesis of Causal Models*, page 221–236. Association for Computing Machinery, New York, NY, USA, 1 edition.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57(2):307–333.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wang, Y., Solus, L., Yang, K., and Uhler, C. (2017). Permutation-based causal inference algorithms with interventions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wang, Y. S. and Drton, M. (2019). High-dimensional causal discovery under non-Gaussianity. *Biometrika*, 107(1):41–59.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Williamson, R. C. and Downs, T. (1990). Probabilistic arithmetic. i. numerical methods for calculating convolutions and dependency bounds. *International Journal of Approximate Reasoning*, 4(2):89–158.

Xu, Y. and Yadlowsky, S. (2022). Calibration error for heterogeneous treatment effects. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I., editors, *International Conference on Artificial Intelligence and Statistics, AISTATS 2022, 28-30 March 2022, Virtual Event*, volume 151 of *Proceedings of Machine Learning Research*, pages 9280–9303. PMLR.

Yadlowsky, S., Fleming, S., Shah, N., Brunskill, E., and Wager, S. (2021). Evaluating treatment prioritization rules via rank-weighted average treatment effects.

Yao, Z., van Velthoven, C. T., Nguyen, T. N., Goldy, J., Sedeno-Cortes, A. E., Baftizadeh, F., Bertagnolli, D., Casper, T., Chiang, M., Crichton, K., Ding, S.-L., Fong, O., Garren, E., Glandon, A., Gouwens, N. W., Gray, J., Graybuck, L. T., Hawrylycz, M. J., Hirschstein, D., Kroll, M., Lathia, K., Lee, C., Levi, B., McMillen, D., Mok, S., Pham, T., Ren, Q., Rimorin, C., Shapovalova, N., Sulc, J., Sunkin, S. M., Tieu, M., Torkelson, A., Tung, H., Ward, K., Dee, N., Smith, K. A., Tasic, B., and Zeng, H. (2021). A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *Cell*, 184(12):3222–3241.e26.

Ye, Q., Amini, A. A., and Zhou, Q. (2021a). Optimizing regularized cholesky score for order-based learning of bayesian networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3555–3572.

Ye, S. S., Chen, Y., and Padilla, O. H. M. (2021b). 2d score based estimation of heterogeneous treatment effects.

Yin, M., Shi, C., Wang, Y., and Blei, D. M. (2021). Conformal sensitivity analysis for individual treatment effects.

Yu, S., Drton, M., and Shojaie, A. (2020). Directed graphical models and causal discovery for zero-inflated data.

Zeng, Y., Hao, Z., Cai, R., Xie, F., Ou, L., and Huang, R. (2020). A causal discovery algorithm based on the prior selection of leaf nodes. *Neural Networks*, 124:130–145.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. (2018). Dags with no tears: Continuous optimization for structure learning. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Zheng, X., Dan, C., Aragam, B., Ravikumar, P., and Xing, E. (2020). Learning sparse nonparametric dags. In Chiappa, S. and Calandra, R., editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3414–3425. PMLR.