# Statistical Advancements in Causal Decomposition and Causal Discovery for Real-World Applications

Shreya Prakash

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2026

Reading Committee:
Elena A. Erosheva, Chair
Yen-Chi Chen
Fan Xia

Program Authorized to Offer Degree:
Statistics

University of Washington

**Abstract**

Statistical Advancements in Causal Decomposition and Causal Discovery for Real-World Applications

Shreya Prakash

Chair of the Supervisory Committee:
Elena A. Erosheva
Department of Statistics

Given recent advances in causal inference, what challenges still prevent these methods from being widely or reliably used in real-world applications? This dissertation considers two tasks in causal inference—causal decomposition and causal discovery—and develops statistical methodology in three separate projects to support their use in practice.

In the first project, we study Black–white disparities in NIH R01 grant peer review. Prior work has identified the step of selection into panel discussion as an important contributor to Black–white funding disparities at the NIH. In this project, our objective is to use a causal decomposition framework to understand what factors could causally contribute to reducing Black–white disparities in selection into discussion, and to assess how recent procedural changes may affect Black–white disparities in selection into discussion. To achieve these objectives, we extend the causal decomposition framework to accommodate the complexities of the NIH peer review process. For example, we account for interference between applications, since proposals are evaluated relative to others and interventions on one proposal's score may affect the discussion outcomes of others. We also develop a Bayesian estimation procedure for causal decomposition that incorporates group-level structure. Using our causal decomposition framework, we first explore whether hypothetical interventions eliminating Black-white disparities and/or differences in attributes would reduce Black-white disparities in selection

of proposals for panel discussion. The attributes we consider include those associated with applicants (degree, career stage, their institution's funding bin) and their submitted proposals (application type, amended status, Preliminary Overall Impact Score, and NIH assigned Administering Organization and Integrated Review Group). Under reasonable assumptions, we find that among the attributes considered, Preliminary Overall Impact Score was the only one that, after equalizing, could dissolve Black-white disparities in selection into discussion. In addition, we conduct a thought experiment to explore potential impacts of the NIH's move to binary scoring for the Investigator and Environment criteria in the new Simplified Peer Review Framework on Black-white disparities in selection into discussion. To carry out this thought experiment, we introduce a new estimand within our causal decomposition framework, the *Thresholded Disparity*, which captures the impact of discretizing evaluation criteria while accounting for the same complexities of the NIH peer review process (e.g., interference), and estimate it using a similar Bayesian procedure that incorporates group-level structure. Analyses from our thought experiment suggest that changing to binary scoring of the Investigator-and-Environment factor may not have any effect on disparities in selection into discussion. Overall, under the assumptions of the causal decomposition framework, these results suggest that efforts aimed at reducing disparities in selection into discussion should focus on identifying upstream, policy-relevant, and practically feasible interventions for eliminating racial disparities in Preliminary Overall Impact Scores. In contrast, the transition to the Simplified Peer Review Framework is unlikely to substantially reduce disparities in selection into discussion.

In the second and third projects, we develop statistical inference tools for causal discovery to address the lack of formal inference and sensitivity to assumption violations in existing methods. Existing functional causal discovery approaches use structural asymmetries to identify causal directionality but rely on strong modeling assumptions and provide limited support for uncertainty quantification. To address these limitations, in the second project, we

develop a foundational hypothesis testing-based framework that provides formal inference on causal direction and diagnostic information about potential assumption violations for bivariate functional causal discovery. We demonstrate these inferential and diagnostic capabilities through simulations with varying degrees of assumption violation and through case studies of cause-effect datasets. Building on this foundational hypothesis testing-based framework, in the third project, we introduce Causal Discovery via Statistical Power (CDSP), a method that connects causal direction estimation—a.k.a. causal discovery—with statistical power. Within CDSP, we define notions of statistical power and effect size in causal discovery and characterize, through the effect-size asymmetry assumption, when the observational data contain sufficient information to favor one causal direction over the other. We establish theoretical results linking effect-size asymmetry to direction detection, showing that when this assumption holds, the probability of correctly detecting the causal direction (i.e., the power of causal discovery) exceeds that of favoring the reverse direction. Leveraging these results, we develop a procedure for causal direction estimation via effect-size asymmetry and provide statistical inference for the estimated direction through a measure of directional support. Simulations show that CDSP direction estimation is robust to mild and moderate model misspecification. Real data analyses on 100 cause–effect benchmark pairs further demonstrate that CDSP reduces false discovery rates by approximately 18% relative to a commonly used existing method.

Together, the contributions of this dissertation support the use of statistically grounded causal inference methodologies in complex real-world settings.

# TABLE OF CONTENTS

Page

List of Figures . . . iii

List of Tables . . . v

Chapter 1: Introduction . . . 1
1.1 Organization of Dissertation . . . 4

Chapter 2: Background . . . 5
2.1 Causal Inference . . . 5
2.2 Causal Inference for Disparity Quantification . . . 6
2.3 Causal Discovery . . . 8

Chapter 3: Causal decomposition of Black-white disparities in NIH's selection of proposals for panel discussion . . . 11
3.1 Introduction . . . 11
3.2 Preliminary analysis . . . 19
3.3 Data Sources . . . 20
3.4 Causal Decomposition Analysis . . . 22
3.5 Estimation . . . 28
3.6 Results . . . 36
3.7 Discussion . . . 44

Chapter 4: Hypothesis Testing for Functional Causal Discovery . . . 48
4.1 Introduction . . . 48
4.2 Current State of Statistical Guarantees in Functional Causal Discovery . . . 49
4.3 A Hypothesis-Testing Framework for Bivariate Functional Causal Discovery . . . 55
4.4 Numerical Results . . . 63
4.5 Discussion . . . 76

Chapter 5: Causal Discovery via Statistical Power (CDSP) . . . 79
5.1 Introduction . . . 79
5.2 Causal Discovery via Statistical Power (CDSP) . . . 82
5.3 Numerical Results . . . 100
5.4 Discussion . . . 113

Chapter 6: Discussion . . . 117
6.1 Contributions . . . 117
6.2 Discussion and Future Work . . . 119

Appendix A: Appendix to Chapter 3 . . . 132
A.1 Exploratory Data Analysis . . . 132
A.2 Selection into Discussion and Preliminary Overall Impact Scores . . . 137
A.3 Discussion of Partial Interference Assumption . . . 138
A.4 Simulation Study . . . 139
A.5 Identification Proofs . . . 144
A.6 Selection . . . 149
A.7 Full Thresholded Disparity Results . . . 150

Appendix B: Appendix to Chapter 4 . . . 153
B.1 Unidentifiable Cases in Post-Nonlinear Models . . . 153
B.2 Example Behaviors of $\xi$ . . . 154
B.3 PNL Model Misspecification Simulation Results . . . 154

Appendix C: Appendix to Chapter 5 . . . 156
C.1 Asymptotic Joint Distribution of $(T_{1b,X\to Y}, T_{1c,Y\to X})$ . . . 156
C.2 Simulation Setup Details . . . 163
C.3 Severe Model Misspecification Simulation Results . . . 164

# LIST OF FIGURES


Figure Number Page

3.1 Overview of the NIH R01 peer review process. . . . 13

3.2 Relationship between two binarized criteria scores Investigator and Environment and the new combined binary Investigator and Environment Score. . . . 18

3.3 Discussion rates (top); Black-white disparities in selecting into discussion by IRG (middle); and estimated discussion rates and the associated 95% Credible Intervals for SRGs within the largest IRG (IRG 7) (bottom). . . . 21

3.4 Causal diagram of the NIH peer review process. . . . 23

3.5 Simplified DAGs representing two conceptual extremes for how criterion scores relate to Preliminary Overall Impact Scores. . . . 23

3.6 Simplified interference structure for two proposals $i$ and $j$ within a single SRG $k$. . . . 27

3.7 Estimated Overall Disparity and Disparity Remaining. . . . 39

3.8 Thresholded disparity results and Equalization Results for Applicant-Related Criterion Scores. . . . 42

4.1 Left: Raw data. Middle: Residuals from regressing the outcome on the predictor in the correct causal direction. Right: Residuals from regressing the predictor on the outcome in the incorrect direction. . . . 57

4.2 Simulation settings for linearity and Gaussianity. . . . 67

4.3 The first two rows presents results for the Balltrack datasets, and the third row presents results for the Temperature and Ozone dataset. . . . 74

5.1 Settings of linearity . . . 101

5.2 Distribution of the estimated directional detectability indices $I_{X\to Y,n}$ (blue) and $I_{Y\to X,n}$ (orange) under varying degrees of nonlinearity. The top-left panel corresponds to the linear case ($d$ = 1). The remaining panels correspond to $d$ = 1.2 (top-right), 1.25 (middle-left), 1.3 (middle-right), 1.4 (bottom-left), and 1.5 (bottom-right), representing progressively stronger departures from linearity. Solid blue and orange vertical lines indicate population values $I_{X\to Y}$ and $I_{Y\to X}$, respectively. . . . 103

5.3 Rows 1–3 correspond to the Population and Dietary Consumption, Balltrack, and Ozone–Temperature datasets, respectively. The first column shows scatterplots of the observed data. The second column shows histograms of the bootstrap distributions of $I_{X\to Y,n}$ (blue) and $I_{Y\to X,n}$ (orange); the solid vertical lines mark the corresponding estimates computed on the full dataset. Each histogram also reports $\widehat{P}_{\text{CDSP}}$. To the right of each histogram, we report the overall LiNGAM and CDSP results; LiNGAM bootstrap rates for the inferred direction are shown in brackets. . . . . 111

A.1 Distributions of Mean Preliminary Overall Impact Scores Conditioned on PI race. . . . . 133

A.2 Distributions of IRG and Administering Organization Conditioned on PI race. 134

A.3 Causal Diagram of NIH Peer Review Process after accounting for Selection Bias. . . . . 149

A.4 Thresholded disparity full results. . . . . 151

C.1 Distribution of the estimated directional detectability indices $I_{X\to Y,n}$ (blue) and $I_{Y\to X,n}$ (orange) under $d = 3$. Solid vertical lines indicate the corresponding population values ($I_{X\to Y}$ in blue and $I_{Y\to X}$ in orange). . . . . 165

# LIST OF TABLES


Table Number Page

2.1 Examples of causal estimands used in the fairness literature. Kusner et al., 2017 introduced Counterfactual Fairness, and Chiappa, 2019; Nabi and Shpitser, 2018 use mediation-based approaches. . . . . 6

3.1 NIH's descriptions of the five review criteria under the Enhanced Peer Review Process. . . . . 14

3.2 NIH's descriptions of the Review Criteria Under the Simplified Peer Review Framework. . . . . 16

4.1 Current state of statistical guarantees in functional causal discovery. . . . . 50

4.2 Causal discovery outcomes and their interpretation when using hypothesis tests in (4.7) to test for goodness-of-fit and independence test. . . . . 60

4.3 Simulation settings for varying linearity. . . . . 66

4.4 Simulation settings for varying levels of Gaussianity. . . . . 66

4.5 Comparison between the test-based approach and DirectLiNGAM under varying degrees of linearity assumption violation. . . . . 68

4.6 Comparison between the test-based approach and DirectLiNGAM under varying degrees of non-Gaussianity assumption violation. . . . . 69

5.1 Guidelines for interpretation of directional support $\widehat{P}_{\text{CDSP}}$ values. . . . . 96

5.2 Accuracies of CDSP and LiNGAM directional estimates for polynomial degrees $d \in \{1, 1.2, 1.25, 1.3, 1.4, 1.5\}$. . . . . 102

5.3 Accuracy of CDSP directional estimate for increasing sample size $N$ for $d = \{1.4, 1.5\}$. . . . . 106

5.4 Comparison of CDSP and LiNGAM on all 100 benchmark pairs. . . . . 108

5.5 Comparison of CDSP and LiNGAM on 34 approximately linear pairs. . . . . 109

A.1 Summary Statistics for Investigator and Environment Criterion Scores. . . . . 135

A.2 Summary Statistics for Proposal and PI Characteristics. Substantial differences and disparities are bolded. . . . . 136

A.3 Summary Statistics for Discussion and PI race. . . . . 136

A.4 Logistic regression model for selection into discussion as the outcome, with PI race, ethnicity, gender, average Preliminary Overall Impact Scores, and other covariates as fixed effects, and SRG included as a random effect. . . . . 138
A.5 Estimator bias based on simulations. . . . . 143
A.6 Summary of the interventions of interest, corresponding estimands, identification assumptions, and identification forms for the interventions considered. . . 144
A.7 Summary of the estimands of interest, identification assumptions, and identification forms for interventions targeting the applicant-related criterion scores under the Simultaneous and Normatively Backwards Evaluation Models. . . . 145
B.1 Test-based approach under varying degrees of model misspecification. . . . . 155

# ACKNOWLEDGMENTS

First and foremost, I want to thank my research advisor, Elena Erosheva, for her support and guidance throughout my PhD. I am especially grateful for the time, care, and thoughtfulness she invested in me as I navigated the challenges of graduate school and developed as an independent researcher. Her mentorship provided both structure and flexibility, shaping my dissertation work and the researcher I am today. It has truly been a privilege to work with her.

I am also deeply grateful to Fan Xia and Carole Lee, whose mentorship and collaboration throughout my dissertation projects helped shape my approach to research. Their continued support, feedback, and encouragement have been instrumental throughout this process. I thank Thomas Richardson and Carlos Cinelli for sparking my interest in causal inference through their courses on causal modeling, and I am especially grateful to Thomas Richardson for his mentorship as a TA advisor. I also thank the Statistical and Machine Learning Approaches for the Social Sciences working group for their thoughtful feedback and support on earlier iterations of the work included in this dissertation. I additionally thank my dissertation committee — Elena, Fan, Yen-Chi Chen, and Ting Ye — for their time, guidance, and insightful feedback. I would also like to thank my academic advisor, Ema Perkovic, for her support and advice during our advising meetings, and the Department of Statistics staff for their guidance and support in navigating the many administrative steps of the PhD.

I have been fortunate to be surrounded by friends who made this journey meaningful, joyful, and filled with encouragement and support. To Akshaj Kumar, Alana McGovern, Alex Kokot, Allison Traylor, Anna Gupta, Anupreet Porwal, Apara Venkat, Ashna Priyal, Jessica Kunke, Jillian Fisher, Medha Agarwal, Nikhila Jayaraman, Nitika Subramanian, Rahul Pun-

hani, Rishi Thakker, Ronak Mehta, Sagar Punhani, Saksham Jain, Sandra Sajeev, Sharika Hegde, Shivkumar Iyer, Vydhourie Thiyageswaran, and the many others who supported me along the way: thank you for the conversations, encouragement, laughter, and companionship throughout the many stages of graduate school. I am especially thankful for the friendships that helped me through the difficult moments and made the good moments even better.

Special thanks to my partner, James Buenfil, who has been my steady source of support, perspective, and joy throughout this journey. Thank you for believing in me, encouraging me through moments of uncertainty, and being my rock through the highs and lows.

I also want to thank Peter Freeman, Rebecca Nugent, and Alexandra Chouldechova from my time as an undergraduate at Carnegie Mellon. Their mentorship, teaching, and generosity fostered my love of learning and helped inspire me to pursue a PhD.

Finally, I thank my family for their unwavering belief in me. To my brother, Pratik Prakash, thank you for helping me in more ways than I can possibly mention throughout the PhD. You gave me a home in Seattle, both physically and emotionally, and I am deeply grateful for your support. To my parents, Sheetal Prakash and Prakash Subramanian, thank you for your constant love, support, and encouragement. Everything I have accomplished has been made possible by the foundation you gave me and the belief you have always had in me. And to my dog, Luna, thank you for bringing comfort, joy, and much-needed reminders to step away from work.

Lastly, I want to remember my maternal grandparents, Rajeshwari Nagarajan and V. Nagarajan, who sadly passed away during my PhD. They instilled in me a deep curiosity, a strong work ethic, and a love of learning that have guided me throughout this journey. I wish they could be here to see this milestone.

# DEDICATION

To my family

# Chapter 1
# INTRODUCTION

Causal inference methods have seen substantial theoretical and methodological advances, yet their application in real-world settings remains challenging. In practice, data often arise from complex systems where methods developed under idealized assumptions may not directly apply, making causal conclusions difficult to interpret and trust. This dissertation develops application-driven statistical methodologies for causal decomposition and causal discovery to address these challenges and support the reliable use of causal methods in practice.

### *1.0.1 Contributions to Statistical Methodology*

This dissertation develops statistical methodology for causal decomposition and causal discovery, motivated by the challenges of applying causal methods in complex, real-world settings. Chapter 2 provides background on causal inference, causal decomposition for disparity quantification, and causal discovery, establishing the foundation for the methodological developments that follow.

In Chapter 3, we extend the causal decomposition framework to the NIH peer review setting, addressing practical challenges such as interference between applications—since proposals are evaluated relative to others—and a hierarchical institutional structure, where applications are reviewed within panels that are themselves grouped into larger administrative units with constraints on how many applications can be discussed. We formalize causal estimands and identification assumptions for this setting and derive identification results for interventions relevant to NIH peer review. We also develop a Bayesian estimation procedure

that incorporates group-level structure. First, we use our causal decomposition framework to explore whether eliminating disparities and/or differences in attributes would reduce Black-white disparities in selection of proposals for panel discussion. Second, we conduct a thought experiment to assess the impact of discretizing evaluation criteria, motivated by proposed reforms to the peer review process. To carry out this thought experiment, we introduce a new estimand, the *Thresholded Disparity*, which captures the impact of discretizing evaluation criteria while accounting for the same complexities of the NIH peer review process (e.g., interference), and estimate it using our Bayesian procedure.

In Chapter 4, we develop statistical inference tools for causal discovery to address the lack of formal inference. We first provide a systematic review of the statistical guarantees of existing functional causal discovery methods across theory, software, and applied use, identifying a key gap: the absence of a unified inferential framework that applies broadly across model classes. To address this gap, we develop a hypothesis testing-based framework for bivariate causal discovery. This foundational framework yields explicit uncertainty quantification, characterizes distinct causal discovery outcomes, and provides diagnostic insight into potential model misspecification and lack of identifiability. We further introduce resampling-based procedures to assess the stability of causal direction estimates and provide practical recommendations for advancing statistical guarantees in functional causal discovery.

In Chapter 5, building on the test-based approach developed in Chapter 4, we introduce Causal Discovery via Statistical Power (CDSP), a framework that connects causal direction estimation with statistical power. We define notions of statistical power and effect size tailored to causal discovery and characterize when the data contain sufficient information to favor one causal direction over the other. Within this framework, we formalize an *effect-size asymmetry* assumption, under which the incorrect direction deviates more strongly from its null and is therefore easier to reject than the correct direction. We show that this asymmetry implies that the probability of detecting the correct causal direction exceeds that of the incorrect direction, yielding a statistical characterization of directional evidence. Building on these results, we develop an inferential procedure for estimating causal direction

along with a measure of directional support. Simulation studies and real data applications demonstrate robustness to model misspecification.

### *1.0.2 Contributions to Real-World Applications*

This dissertation is motivated by the gap between theoretical developments in causal inference and real-world applications. Chapter 3 focuses on the NIH R01 grant peer review process, a high-stakes evaluation system in which prior work has identified selection into panel discussion as the decision point contributing most to Black–white NIH funding disparities (Hoppe et al., 2019).

Using a causal decomposition framework, we evaluate what factors causally contribute to Black–white disparities in selection into panel discussion. In particular, we examine whether hypothetical interventions eliminating Black–white disparities and/or differences in attributes would reduce disparities in selection into discussion. In addition, we conduct a thought experiment to explore the potential impact of the NIH's transition to binary scoring for the Investigator and Environment criteria under the Simplified Peer Review Framework (NIH Center for Scientific Review, 2024b) on disparities in selection into discussion.

Under reasonable assumptions, we find that among the attributes considered, Preliminary Overall Impact Score was the only one that, after equalizing, could dissolve Black-white disparities in selection into discussion. Furthermore, analyses from our thought experiment suggest that changing to binary scoring of the Investigator-and-Environment factor may not have any effect on disparities in selection into discussion. Overall, under the assumptions of the causal decomposition framework, these results suggest that efforts aimed at reducing disparities in selection into discussion should focus on identifying upstream, policy-relevant, and practically feasible interventions for eliminating racial disparities in Preliminary Overall Impact Scores. In contrast, the transition to the Simplified Peer Review Framework is unlikely to substantially reduce disparities in selection into discussion.

More broadly, the methodologies developed in this dissertation enable the application of causal methods in settings where standard assumptions may not directly hold. In Chapter 3,

this enables evaluation of interventions in institutional decision-making processes such as peer review, where features such as interference between proposals, self-selection mechanisms, and structural constraints must be accounted for.

In Chapter 4, the proposed test-based causal discovery framework provides formal inference on causal directionality and a diagnostic tool for detecting potential violations of the assumptions underlying functional causal discovery methods. In Chapter 5, we introduce Causal Discovery via Statistical Power (CDSP), a statistically grounded framework for direction detection that is robust to model misspecification and supports inference through a measure of directional support. Across Chapters 4 and 5, we demonstrate these methods through simulations and applications to the Tübingen Cause–Effect Pairs (Mooij et al., 2016a), showing that they can improve upon a commonly used existing approach in realistic data settings.

Together, these contributions support the use of statistically grounded causal methods for reliable decision-making and scientific discovery in complex, real-world settings.

## *1.1 Organization of Dissertation*

To summarize, Chapter 2 provides important background on causal inference, causal decomposition for disparity quantification, and causal discovery. Chapters 3–5 present the main contributions of the dissertation. Specifically, Chapter 3 develops causal decomposition methods for NIH R01 peer review to study which factors causally contribute to Black–white disparities in selection into discussion and to assess how recent procedural changes may affect these disparities. Chapter 4 develops a hypothesis testing-based framework for causal discovery that enables statistical inference on causal direction and provides diagnostic insight into potential assumption violations. Chapter 5 builds on this foundational hypothesis testing-based framework and introduces Causal Discovery via Statistical Power (CDSP), which connects causal direction estimation with statistical power, characterizes when the data support one direction over the other, and is robust to model misspecification. Last, Chapter 6 discusses the complete work and proposes areas for future research.

# Chapter 2

# BACKGROUND

In this chapter, we provide key background information that will be useful for understanding the subsequent chapters of this dissertation. Specifically, we provide background on causal inference, causal decomposition for disparity quantification, and causal discovery.

## *2.1 Causal Inference*

We begin with a brief introduction to causal inference, as all subsequent chapters relate to this topic. Causal inference is a discipline that "considers the assumptions, study designs, and estimation strategies that allow researchers to draw causal conclusions based on data" (Hill & Stuart, 2015).

To draw causal conclusions from data, one must typically specify the underlying causal structure. When this structure is well understood (e.g., from randomized experiments), one can proceed to estimate causal quantities of interest. However, in many settings, the causal structure is not known a priori. In such cases, researchers may turn to causal discovery—the identification of causal relationships among variables from observational data. In Section 2.3, we review causal discovery as it relates to the work in Chapters 4 and 5.

Once a causal structure is specified, either through domain knowledge or causal discovery methods, it can be used to conduct causal analyses. One important application of causal inference is in quantifying fairness and disparities. In particular, causal methods can be used to identify factors that contribute to disparities and to evaluate policy interventions that may reduce them. In Section 2.2, we provide a more detailed review of causal methodologies for quantifying fairness and disparities as they relate to the work in Chapter 3.

| Estimand Name | Estimand | Interpretation |
|---|---|---|
| Counterfactual Fairness | $P(Y(A = a) = y \mid M = m, A = a) = P(Y(A = a') = y \mid M = m, A = a)$ $\forall\, a' \in A,\ y \in Y$ | Given $M = m$ and $A = a$ the probability of $Y = y$ equivalent to the probability of $Y = y$ if $A = a'$. |
| Natural Direct Effect | $E[Y(A = a, M(A = a')) - Y(A = a', M(A = a'))]$ | Measures the effect of the direct causal path $A \rightarrow Y$ |
| Natural Indirect Effect | $E[Y(A = a', M(A = a)) - Y(A = a', M(A = a'))]$ | Measures the effect of the mediator $M$ at levels $M(a)$, $M(a')$ on the outcome $Y$ had $A = a'$. |

Table 2.1: Examples of causal estimands used in the fairness literature. Kusner et al., 2017 introduced Counterfactual Fairness, and Chiappa, 2019; Nabi and Shpitser, 2018 use mediation-based approaches.

## *2.2 Causal Inference for Disparity Quantification*

In this section, we briefly review causal methodologies for quantifying fairness and disparities as they relate to the work in Chapter 3. Racial disparities are commonly defined as differences in outcomes across racial groups (Barocas et al., 2023; VanderWeele & Robinson, 2014). These differences may arise through multiple pathways, including factors associated with race as well as intermediate variables that lie on causal pathways from race to the outcome. The framework of causal inference provides tools for quantifying the mechanisms that contribute to these disparities and for evaluating interventions that may reduce them (VanderWeele & Robinson, 2014).

To formalize these questions, we use counterfactual variables, or potential outcomes, which denote the outcomes that would be observed under hypothetical interventions (Imbens & Rubin, 2015; Rubin, 1974). Let $Y$ denote the outcome of interest (e.g., peer review funding decisions), $A$ denote self-reported race, and $M$ denote a mediator—a variable that is affected by race and in turn affects the outcome (e.g., peer review scores). We use notation such as $Y(m)$ to represent these quantities.

Many approaches in the fairness literature define estimands that contrast counterfactual outcomes across levels of race (see Table 2.1) (e.g., Chiappa, 2019; Kusner et al., 2017; Nabi

and Shpitser, 2018). These include approaches that aim to ensure decisions or predictions remain unchanged under hypothetical changes to race, as well as approaches that examine how disparities arise through intermediate variables.

However, several researchers have argued that counterfactual variables involving race are not well-defined (e.g., Barocas et al., 2023; Hu, 2023; VanderWeele and Robinson, 2014). Race is not a manipulable treatment in the conventional sense, and its effects may reflect a combination of physical features, ancestry, and social and cultural factors (VanderWeele & Robinson, 2014). As a result, standard causal interpretations that rely on well-defined interventions are difficult to apply directly to race.

One approach proposed in the literature is to shift the focus from the effect of race itself to the effect of exposure to perceived race (Barocas et al., 2023). Under this perspective, the "treatment" corresponds to observable attributes (e.g., a PI's name) that may signal perceived race, and the estimand captures how decision-makers respond to such signals. While this approach introduces manipulable interventions, it also raises challenges, as perceptions of race can arise through multiple channels (e.g., names, speech, or other characteristics), and it is difficult to define coherent counterfactual scenarios that alter these signals while preserving all other relevant features (Barocas et al., 2023).

To address these issues, VanderWeele and Robinson (2014) propose focusing on mediating mechanisms that contribute to disparities. They introduce this framework to "provide a causal interpretation of race coefficients... without defining potential outcomes for race itself," noting that "there are... no reasonable hypothetical interventions on race." Instead, the framework asks "how much a racial inequality could be eliminated by intervening on a different variable," thereby shifting the focus to factors that are more amenable to intervention (VanderWeele & Robinson, 2014). This approach avoids causal interpretations of race itself while allowing for causal interpretation of how interventions on mediators may reduce disparities.

Within this framework, disparities can be decomposed into components corresponding to different intervention targets. To define the causal estimands of interest, VanderWeele

and Robinson (2014) use the potential outcomes framework (Rubin, 1974). Let $C$ denote baseline covariates. The potential outcome $Y(M = m)$ denotes the outcome if the mediator $M$ were set to value $m$. They propose a causal decomposition consisting of the observed disparity, disparity reduction, and disparity remaining:

$$\text{Observed Disparity: } E[Y|A = 0] - E[Y|A = 1] \tag{2.1}$$

$$\text{Disparity Reduction: } E[Y(M^*)|A = 1] - E[Y|A = 1] \tag{2.2}$$

$$\text{Disparity Remaining: } E[Y|A = 0] - E[Y(M^*)|A = 1] \tag{2.3}$$

Here, $M^* \sim P(M|A = 0, C)$ represents an intervention that equalizes the distribution of the mediator across racial groups. Under assumptions such as consistency, conditional ignorability, and positivity, these quantities are identifiable (VanderWeele & Robinson, 2014). The authors also discuss challenges that arise when conditioning on selected populations (Bareinboim & Pearl, 2012; VanderWeele & Robinson, 2014). Subsequent work has developed estimation procedures for these estimands (Jackson & VanderWeele, 2018; Park et al., 2020).

### *2.3 Causal Discovery*

In this section, we briefly review causal discovery methodologies relevant to Chapters 4 and 5. Identifying and understanding causal relations is fundamental to many scientific disciplines. For example, in biology, uncovering causal links in gene regulatory networks is crucial for gaining insight into biological processes and disease mechanisms (Jiang et al., 2023). While randomized experiments and carefully designed interventions can reveal causal relationships, these are often too expensive, time-consuming, or infeasible. As a result, reliable causal discovery methods that can extract evidence of causal structure from observational data are essential.

Causal discovery methods have been applied across the sciences (e.g., Glymour et al., 2019; Hu et al., 2018; Kotoku et al., 2020; Rosenström et al., 2012; Saito and Fujiwara, 2023; Zhang et al., 2012). These methods typically represent causal structure using a Directed Acyclic Graph (DAG), whose nodes correspond to random variables and whose edges represent direct causal dependencies. The acyclicity assumption rules out directed cycles and ensures that the graph encodes a well-defined causal ordering.

Two broad classes of causal discovery are constraint-based and function-based methods (Glymour et al., 2019). Constraint-based approaches—such as the Fast Causal Inference (FCI) algorithm (Spirtes et al., 2000) and the Peter–Clark (PC) algorithm (Spirtes et al., 2000)—search for patterns of conditional independence in the data. Their validity depends on the faithfulness assumption, which requires that all conditional independence relations present in the observed distribution are implied by the DAG. In general, constraint-based methods identify a Markov equivalence class rather than a unique DAG. These methods work best with large sample sizes, where tests of conditional independence have sufficient power (Glymour et al., 2019), and they provide no information in purely bivariate settings where no conditional independences exist.

Functional causal discovery methods take a different approach: they use structural assumptions on the functional form of the relationship between variables to identify the causal direction. In the bivariate case, for two random variables $X$ and $Y$, suppose the true causal direction is $X \to Y$. Then, functional causal models are of the form

$$Y = f(X, \eta; \boldsymbol{\phi}),$$

where $\eta$ is noise that is independent of $X$, $f \in \mathcal{F}$ is the functional form, and $\boldsymbol{\phi}$ is the parameter set. Given $\boldsymbol{\phi}$, some models additionally assume that $f$ is invertible so that $\eta$ can be recovered uniquely from $(X, Y)$. These methods identify the causal direction by exploiting asymmetries: when the model assumptions hold, the noise term is independent of the cause in the true direction but not in the reverse direction.

Prominent examples include the Linear Non-Gaussian Acyclic Model (LiNGAM) (Shimizu & Kano, 2008), where $Y = \beta X + \eta$ and the noise $\eta$ is non-Gaussian; the Additive Noise Model (ANM) (Hoyer et al., 2009), where $Y = f_{\text{AN}}(X) + \eta$; and the Post-Nonlinear Model (PNL) (Zhang & Chan, 2006; Zhang & Hyvarinen, 2012), where $Y = f_2(f_1(X) + \eta)$ with $f_2$ invertible. Shimizu et al. (2011, 2006), Hoyer et al. (2009), and Zhang and Hyvarinen (2012) formalize identifiability results for these models and provide algorithms for estimating the causal direction.

Functional causal discovery methods have been used widely in neuroscience, epidemiology, psychology, genetics, and other domains in the medical and social sciences. For example, Saito and Fujiwara (2023) used LiNGAM to study factors contributing to nitrogen oxide generation in a coal-fired power plant. Rosenström et al. (2012) used LiNGAM to examine causal relationships between sleep and depression, while Hu et al. (2018) applied ANMs for causal discovery in genetic studies. Song et al. (2017) applied ANM and PNL models to diverse bivariate datasets across geography, biology, physics, and economics. Despite this widespread use, the state of statistical guarantees for these methods remains underdeveloped; we provide a focused review of the current state of statistical guarantees in functional causal discovery in Chapter 4.

# Chapter 3

# CAUSAL DECOMPOSITION OF BLACK-WHITE DISPARITIES IN NIH'S SELECTION OF PROPOSALS FOR PANEL DISCUSSION

This work was conducted in collaboration with Dr. Elena A. Erosheva and Dr. Carole J. Lee. A revised version of the original manuscript received a 2026 Joint Statistical Meetings (JSM) Student Paper Award jointly from the Social Statistics, Government Statistics, and Survey Research Methods Sections of the American Statistical Association. This chapter provides the most complete account of the methodology.

## *3.1 Introduction*

The National Institutes of Health (NIH) "is the largest public funder of biomedical research in the world" (National Institutes of Health, 2026). The NIH funds research through a competitive peer review process. A major mechanism for this funding is the Research Project Grant (R01), which supports investigator-initiated research aligned with the mission of the NIH and its Institutes and Centers (National Institutes of Health, n.d.). The R01 grant "is the original and historically oldest grant mechanism used by NIH"(National Institute of Biomedical Imaging and Bioengineering, 2026). Research has demonstrated persistent Black-white disparities in funding at the NIH, even as of 2022 (Lauer & Roychowdhury, 2023), more than ten years after Ginther et al., 2011's first report. Portions of the funding gap have been attributed to different structural inequities in science. Ginther et al., 2011 found that, in submitted Biosketches, new and experienced Black Principal Investigators (PIs) publish in journals with lower impact factors, have fewer field-normalized citations, and report fewer publications than their white peers—and that these bibliometric differences

account for roughly half the funding gap. Lauer et al., 2021 discovered that part of the funding disparity could be attributed to the lower award rates at Centers and Institutes to which Black PI's disproportionately apply Lauer et al., 2021—and that this fully accounts for previously observed associations between PI race and proposal topic choice Hoppe et al., 2019. Inequities in mentoring and co-author networks are also thought to play a role: Warner et al., 2017 found that within the context of Harvard Medical School, an NIH applicant's network reach—measured as the intra-institutional number of the applicant's co-authors plus the intra-institutional number of their co-authors' co-authors—was associated with receiving an NIH R01 grant. Although the Black-white funding gap at NIH has narrowed in recent years, it has not been eliminated (Refer to Figure 8 of Lauer and Roychowdhury, 2024). The persistence of Black–white disparities raises concerns about equity in access to research funding and the potential loss of scientific contributions from underrepresented groups. The continued attention to this issue in recent work following Ginther et al., 2011 underscores its ongoing importance (e.g., Erosheva et al., 2020a; Lauer et al., 2021). However, most of this work relies on associational analyses and does not establish which factors causally contribute to disparities or which changes to the review process would meaningfully reduce them.

In this chapter, our objective is to use a causal decomposition framework to understand what factors causally contribute to reducing Black–white disparities in selection into discussion, and to assess how recent procedural changes may affect Black–white disparities in selection into discussion.

Research grant proposal applications must successfully surmount multiple hurdles to secure NIH R01 funding, as shown in Fig 3.1, with the first being selection into discussion by Scientific Review Groups (SRGs). NIH SRGs—study sections composed of subject-area experts who conduct the peer review of R01 applications—are organized within broader Integrated Review Groups (IRGs), which cluster related study sections by scientific discipline.

In the Enhanced Peer Review Process used by the NIH between January 25, 2009 and January 24, 2025 (NIH Center for Scientific Review, 2009), assigned reviewers evaluated the proposed research and the biosketch, which contains applicants' names, institutional affili-

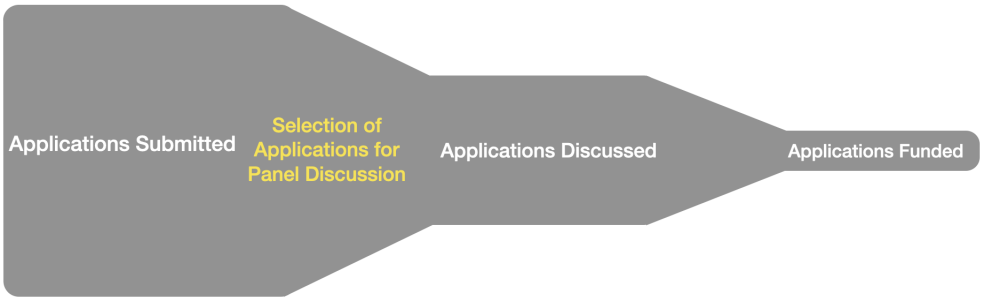



Figure 3.1: Overview of the NIH R01 peer review process. Figure widths reflect the approximate proportion of applications at each stage: all applications receive preliminary scores; roughly half are selected for discussion; and only a smaller subset ultimately receive funding. White text denotes the three major stages. Yellow text highlights a key review process—the selection of proposals for panel discussion. Adapted from Fig. 1 in Hoppe et al., 2019.

ations, and selected publications Hoppe et al., 2019; Nakamura et al., 2021, and provided preliminary (i.e., before panel discussion) criterion scores for Significance, Investigators, Innovation, Approach, and Environment as well as Preliminary Overall Impact Scores for each submitted application (see Table 3.1 for descriptions). Each preliminary criterion was scored on a scale from 1 to 9, with 1 being the best and 9 the worst. Preliminary Overall Impact Scores were used to nominate which applications—typically half based on Preliminary Overall Impact Scores—were not to be discussed. This process of nominating applications to not be discussed based on a Preliminary Overall Impact Score threshold is known as 'streamlining' (NIH NIAD, 2023). Applications that were not 'streamlined' would go on to be discussed.

Selection for panel discussion included an additional step. That is, after 'streamlining', panel reviewers assessed applications nominated not to be discussed in a randomized order to decide if any should be added to the discussion (NIH NIAD, 2023). For each proposal nominated not to be discussed, all reviewers without a conflict of interest must have agreed unanimously not to discuss that application; otherwise, applications would have been moved along to the discussion phase (NIH NIAD, 2023). All applications selected for SRG panel discussions were provided final Overall Impact Scores. At the funding step, these final

| **Enhanced Peer Review Process (for due dates before Jan 25, 2025)**<br>*All five scored criteria are considered in the Preliminary Overall Impact Score* | |
|---|---|
| **Criterion (Scored 1–9)** | **Description** |
| **Significance** | Does the project address an important problem or a critical barrier to progress in the field? If the aims of the project are achieved, how will scientific knowledge, technical capability, and/or clinical practice be improved? How will successful completion of the aims change the concepts, methods, technologies, treatments, services, or preventative interventions that drive this field? |
| **Innovation** | Does the application challenge and seek to shift current research or clinical practice paradigms by utilizing novel theoretical concepts, approaches or methodologies, instrumentation, or interventions? Are the concepts, approaches or methodologies, instrumentation, or interventions novel to one field of research or novel in a broad sense? Is a refinement, improvement, or new application of theoretical concepts, approaches or methodologies, instrumentation, or interventions proposed? |
| **Approach** | Are the overall strategy, methodology, and analyses well-reasoned and appropriate to accomplish the specific aims of the project? Are potential problems, alternative strategies, and benchmarks for success presented? If the project is in the early stages of development, will the strategy establish feasibility and will particularly risky aspects be managed? If the project involves clinical research, are the plans for (1) protection of human subjects from research risks, and (2) inclusion of minorities and members of both sexes/genders, as well as the inclusion of children, justified in terms of the scientific goals and research strategy proposed? |
| **Investigator(s)** | Are the PD/PIs, collaborators, and other researchers well suited to the project? If Early Stage Investigators or New Investigators, do they have appropriate experience and training? If established, have they demonstrated an ongoing record of accomplishments that have advanced their field(s)? If the project is collaborative or multi-PD/PI, do the investigators have complementary and integrated expertise? Are their leadership approach, governance, and organizational structure appropriate for the project? |
| **Environment** | Will the scientific environment in which the work will be done contribute to the probability of success? Are the institutional support, equipment, and other physical resources available to the investigators adequate for the project proposed? Will the project benefit from unique features of the scientific environment, subject populations, or collaborative arrangements? |

Table 3.1: NIH's descriptions of the five review criteria under the Enhanced Peer Review Process. 'Reviewers will provide' a preliminary 'overall impact/priority score to reflect their assessment of the likelihood for the project to exert a sustained, powerful influence on the research field(s) involved, in consideration of the following five core review criteria, and additional review criteria (as applicable for the project proposed)'. (NIH Center for Scientific Review, 2016, 2009).

Overall Impact Scores guided the NIH's Centers' and Institutes' funding decisions along with a variety of other factors such as pubic health burden, programmatic priorities, overall portfolio balance and budgets at the funding institutes (Hoppe et al., 2019; Nakamura et al., 2021).

On January 25th, 2025, the NIH introduced the new Simplified Peer Review Framework. The goals of the Simplified Peer Review Framework are to 1) simplify the peer review process to be more focused on the key questions necessary to assess scientific and technical merit, 2) reduce reviewer burden, and 3) mitigate the effect of reputation bias (NIH Center for Scientific Review, 2024b). Under this new system, instead of scoring five criteria individually, reviewers score three broad factors that encompass slightly updated versions of the previous review criteria (NIH Center for Scientific Review, 2023b), as shown in Table 3.2. A key simplification aimed at helping "put applicants on equal footing" involves combining the Investigator and Environment criteria into a single binary assessment to be scored as "gaps identified" or not (NIH Center for Scientific Review, 2024a). A major goal of this specific simplification is to "mitigate the effect of reputational bias" (NIH Center for Scientific Review, 2024a), a plausible contributor to Black-white disparities (Ginther et al., 2011).

In this chapter, we undertake causal analyses to evaluate which interventions could potentially decrease Black-white disparities in selection of proposals for panel discussion. We focus on this stage for two main reasons. First, this decision point culls the largest number of proposals. Second, this step in the process was shown to constitute "the largest single contribution" to Black-white disparities in NIH grant review and funding (Hoppe et al., 2019).

Despite the importance of this step of selection into discussion in the review process, this decision point has not been investigated from a causal analysis perspective. Most analyses aimed at understanding Black-white disparities in funding, including some of the most groundbreaking work (Erosheva et al., 2020a; Ginther et al., 2011; Hoppe et al., 2019; Lauer et al., 2021), have reported associations and, as such, refrained from drawing causal conclusions. For example, the only study that investigated selection of proposals for panel

| **Simplified Peer Review Framework (for due dates on/after Jan 25, 2025)**<br>*All three factors are considered in the Preliminary Overall Impact Score* | |
|---|---|
| **Factor (Score)** | **Evaluation Criteria** |
| **Factor 1: Importance of the Research (Scored 1–9)** | **Significance:** Evaluate the importance of the proposed research in the context of current scientific challenges and opportunities. Does the application address a significant knowledge gap, critical problem, or valuable conceptual/technical advance? Assess the rigor and justification of the scientific background.<br>**Innovation:** Evaluate the influence of innovation on the importance of the proposed research. Does the project apply novel concepts, methods, or technologies, or use existing approaches in novel ways? Projects without novelty may still be highly important to the field. |
| **Factor 2: Rigor and Feasibility (Scored 1–9)** | **Approach:** Evaluate the scientific quality of the work. Assess whether it is likely to yield compelling, reproducible results and whether the proposed studies are feasible within the proposed timeline.<br>**Rigor:** Evaluate experimental design, controls, sample size, plans for analysis, and whether biological variables (e.g., sex or age) are addressed. For human or animal studies, assess the rigor of interventions, outcome justification, generalizability, and population appropriateness.<br>**Feasibility:** Evaluate whether the research plan is realistic, including recruitment and retention strategies (for human studies), study milestones, and likelihood of achieving goals across diverse subgroups. |
| **Factor 3: Expertise and Resources (Binary: Appropriate/Gaps Identified)** | **Investigators:** Are the PD/PIs and collaborators well qualified? Do they have relevant experience, training, and complementary expertise?<br>**Environment:** Does the institutional setting, available resources, and infrastructure support project success?<br>Proposals flagged as having "gaps identified" require an explanation. This factor is not numerically scored but must be assessed for appropriateness. |

Table 3.2: NIH's descriptions of the Review Criteria Under the Simplified Peer Review Framework (NIH Center for Scientific Review, 2023b, 2024b). The Simplified Peer Review Framework reorganizes NIH's five individual review criteria (Table 3.1) into three broader factors: (1) Importance of the Research, (2) Rigor and Feasibility, and (3) Expertise and Resources. The most notable changes are: (i) Significance and Innovation are combined into a single 1–9 score under Factor 1, and (ii) Investigator and Environment are merged into a single binary score under Factor 3. These changes are detailed in NIH guidance and webinars (NIH Center for Scientific Review, 2023b, 2024b).

discussion (Hoppe et al., 2019) found significant associations between PI race, application resubmission, institutional prestige, and career stage and the outcome of selection of proposals for panel discussion; however, these associations estimated from observational data were not evaluated for their causal contribution to disparities in selection into discussion. Experimental efforts to identify causal mechanisms for racial disparities in peer review outcomes include a study that anonymized PI identities (Nakamura et al., 2021) and another study that swapped social identities of the names listed on otherwise identical proposals (Forscher et al., 2019). However, these studies focused primarily on the causal role of perceived applicant race and did not consider other factors that could impact disparities in scoring and funding. Identifying hypothetical interventions that could reduce or eliminate racial disparities in selection into discussion could inform where policies could be directed most effectively.

In Section 3.2, we start by confirming Hoppe et al., 2019's finding of substantial Black-white disparity in selection into discussion with our data and illustrate variation in these disparities across NIH's Integrated and Scientific Review Groups. Then, to gain insight into the scoring and selection process, we extend the causal decomposition framework introduced by VanderWeele and Robinson, 2014 for the peer review context. Section 3.4 outlines the causal decomposition methodology, its assumptions, and its adaptation to the peer review context. Importantly, the causal decomposition framework does not estimate the causal effect of race. Rather, it begins by estimating the observed Black–white disparity in selection into discussion and then asks how this disparity would change under hypothetical interventions of certain attributes (VanderWeele & Robinson, 2014). Section 3.5 presents the estimation procedure. In Section 3.6.1, we use the causal decomposition framework to quantify: (i) the observed Black–white disparity in selection into discussion; (ii) how much of this disparity would be reduced if the distribution of a given attribute (e.g., PI career stage or institutional prestige) were equalized across groups; and (iii) how much of the disparity would remain after such equalization.

In Section 3.6.2, we examine counterfactual questions about how procedural changes akin

to the Simplified Peer Review Framework's move to replace the Investigator and Environment 1-9 scores with one binary Factor 3 score might affect disparities in selection for discussion. We single out this procedural change in our analysis as it represents an intervention introduced by the NIH with the stated goal of "mitigat[ing] the effect of reputational bias"NIH Center for Scientific Review, 2024a, a factor that may contribute to Black–white disparities in peer review (Ginther et al., 2011). While we do not have access to data from NIH's new Simplified Peer Review, we can still consider a thought experiment in which we use available data from the NIH's Enhanced Peer Review to binarize the Investigator and Environment criteria separately since, after deciding on each Investigator and Environment criteria score, one can relate these two binarized scores to the combined Factor 3 score as outlined in Fig 3.2. To implement our thought experiment, we use two conceptual extremes regarding how criterion scores relate to Preliminary Overall Impact Score since that mechanism is unknown.

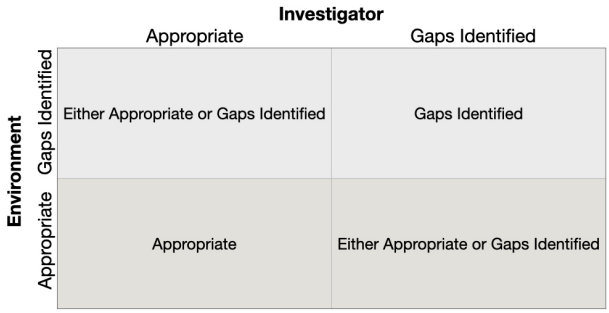



Figure 3.2: Relationship between two binarized criteria scores Investigator and Environment (margins) and the new combined binary Investigator and Environment Score (cells).

As with any statistical analysis, our causal decomposition analysis relies on assumptions, most of which are not directly testable from the observed data. Thus, the causal interpretations of our results should be understood as conditional on the assumptions underlying the causal decomposition framework. In Section 3.4, we describe these assumptions in technical detail and discuss their plausibility in the context of NIH peer review.

Under the assumptions outlined in Section 3.4, our causal decomposition analysis yields several causal conclusions. First, eliminating racial disparities and/or differences in one of

the attributes, in particular Preliminary Overall Impact Score, can dissolve the disparity in selection into discussion. This result was not guaranteed a priori because, as we discuss in the paper and Appendix A.2, selection into discussion is not deterministically driven by Preliminary Overall Impact Scores. Second, we find that the impact of changes in the Simplified Peer Review Framework on disparities in selection into discussion depends on how reviewers evaluate the binary criteria as well as on the underlying causal relationship between the criterion scores and Preliminary Overall Impact Scores. The results from our thought experiment show that two conceptual extremes for this causal relationship can theoretically lead to different effects that some may find surprising (see Figs 3.8a and Appendix A.7). However, if the criteria are evaluated using cutoffs that reflect historical discussion rates at the NIH, our thought experiment suggests that neither extreme would substantially alter Black-white disparities in selection into discussion. Thus, under the assumptions of the causal decomposition framework, we conclude that while reductions in racial disparities in Preliminary Overall Impact Scores may impact disparities in selection into discussion, the transition to the Simplified Peer Review Framework is unlikely to do so.

### *3.2 Preliminary analysis*

Our dataset consists of preliminary peer review scores and several PI and application characteristics for a subset of R01 submissions by Black and white PIs from 2014-2016 (Erosheva et al., 2020a), and is publicly available on the Open Science Framework (Erosheva et al., 2020b) (https://osf.io/4d6rx/). The dataset contains all submissions from Black PIs and a random sample of submissions from white PIs. 22% of the PIs in the data are identified as Black, while 78% are white. Additional details on our dataset and its construction are provided in Section 3.3.

We observe that 54% of proposals are selected for discussion, aligning with the NIH's standard that approximately half of proposals are discussed; however, there is a great deal of variability across Integrated Review Groups (IRGs) (NIH Center for Scientific Review, 2023a). We note that 56.4% (Cred. Interval = [54.4%, 58.2%]) of white PIs are selected

for panel discussion while only 43.9% (Cred. Interval = [40.6%, 47.3%]) of Black PIs are selected[1], leading to a substantial 12.3% (Cred. Interval = [11.3%, 16.3%]) disparity in selection for discussion toward Black PIs. This is similar to the findings of Hoppe et al., 2019, who observed a 13.4% disparity in their analysis of grant applications submitted between 2011 and 2015. Fig 3.3 presents discussion rates and Black-white disparities in discussion rates by IRG. We observe considerable variability, with discussion rates across IRGs ranging from 40% to 57%. Discussion rates further vary by SRG; Fig 3.3 (bottom panel) shows SRG discussion rates for SRGs within the largest IRG range from 41% to 71%.

Discussion rates vary substantially based on Preliminary Overall Impact Scores, as well as Investigator and Environment criterion scores (see Table A.1). Specifically, almost no proposals with preliminary criteria scores above 5 are discussed, while the vast majority of proposals with Preliminary Overall Impact Scores of 1 (the highest score) are discussed. However, both NIH documentation and our data suggest that while selection for panel discussion is primarily driven by Preliminary Overall Impact Score, the process is not entirely deterministic. As outlined in Appendix A.2, our analysis demonstrates that while these scores play a central role, there may be other additional factors that influence which applications are ultimately selected for discussion.

### *3.3 Data Sources*

We analyze publicly available preliminary peer review data on NIH R01 grant applications reviewed by the Center for Scientific Review (CSR) from 2014–2016 (Erosheva et al., 2020a). The original dataset contains 4,651 applications with proposal-, investigator-, and review-level information, including PI race, gender, ethnicity, degree type, career stage, institutional funding bin, application type, amended status, Integrated and Scientific Review Group identifiers, administering Institute or Center, and both preliminary and final scoring outcomes.

[1]Because we observe all Black PI applications, whereas the white PI data comprise both a matched subset of all white PIs and a randomly sampled subset (see Erosheva et al., 2020a for details on the publicly available dataset), the credible intervals are conservative and not directly comparable across groups.

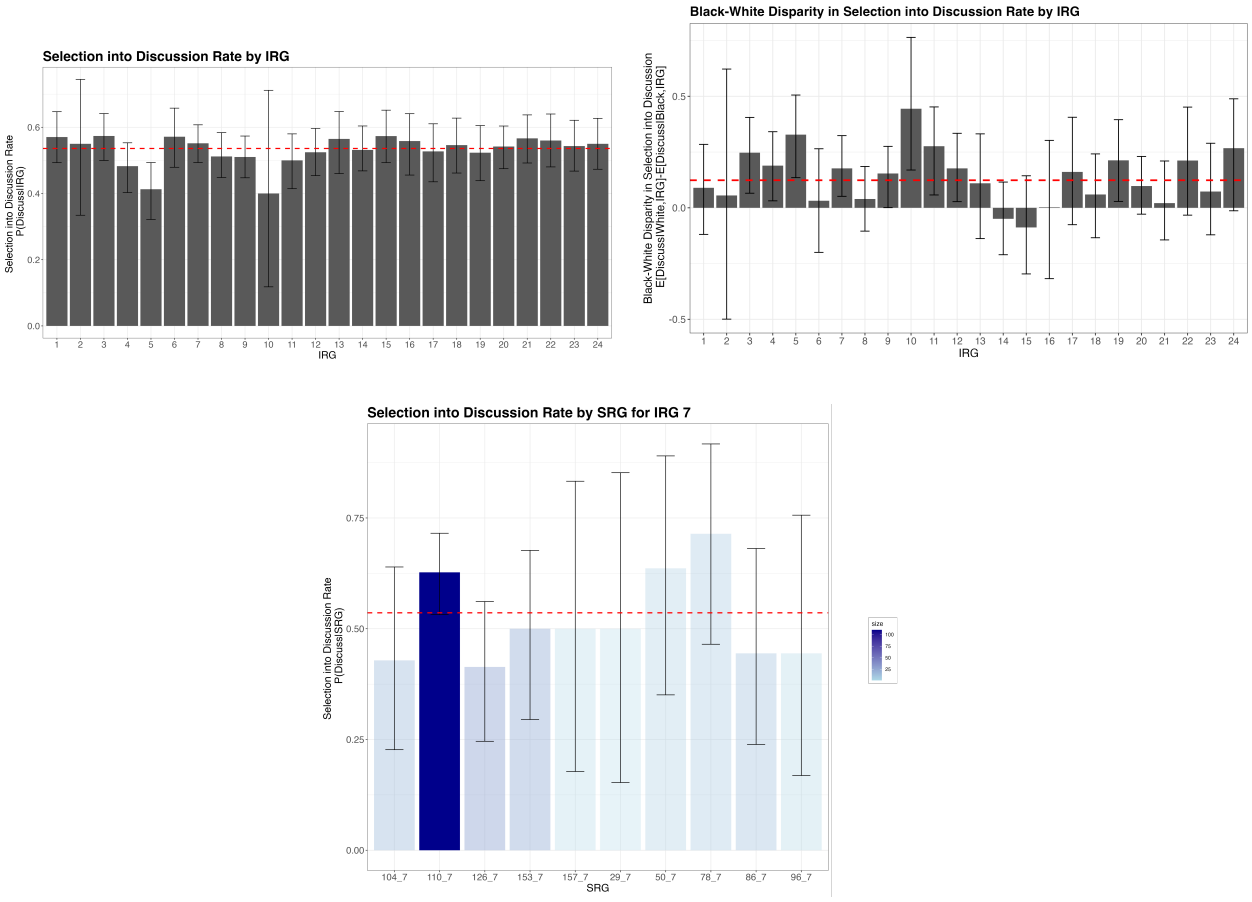



Figure 3.3: Discussion rates (top); Black-white disparities in selecting into discussion by IRG (middle); and estimated discussion rates and the associated 95% Credible Intervals for SRGs within the largest IRG (IRG 7) (bottom). Red dashed line is the estimate of the overall discussion rate of 0.54 (0.51, 0.56) and the overall disparity of -0.12 (-0.16, -0.08). Credible Intervals were obtained using a Bayesian bootstrap with 500 posterior samples.

We restructure the data to the proposal level to focus on selection into discussion. Preliminary scores are averaged across reviewers for each proposal, reflecting NIH procedures in which initial decisions are based on average Preliminary Overall Impact Scores (NIH Center for Scientific Review, 2018). We restrict the sample to applications with three reviewers—the most common configuration—for consistency. The binary outcome indicating selection into discussion is defined based on the availability of a final Overall Impact Score, which is assigned only to proposals that are discussed. This yields a final analytic sample of 3,687 proposals (22% from Black PIs and 78% from white PIs).

### *3.4 Causal Decomposition Analysis*

In this section, we summarize the causal decomposition methodology introduced by VanderWeele and Robinson (2014) and describe how we extend it in the context of NIH peer review. We first construct a causal diagram or directed acyclic graph (DAG) (Figure 3.4) informed by domain knowledge of the NIH peer review process (NIH Center for Scientific Review, 2023a); the presence of directed edges indicates possible causal relationships, while the absence of an edge reflects an assumed lack of causal influence. Key variables include: PI Race (denoted $A$), selection into discussion ($Y$), protected attributes not currently under examination ($W$), application and PI characteristics ($Z$) and contextual factors ($L$). See Figure 3.4 caption for the list of variables in each group.

We extend the causal decomposition framework of VanderWeele and Robinson (2014) to evaluate how changes in the distribution of specific attributes—such as Preliminary Overall Impact Scores or degree category—contribute to Black–white disparities in selection into discussion. Their framework provides "a causal interpretation of race coefficients... without defining potential outcomes for race itself," emphasizing that there are no reasonable hypothetical interventions on race. Instead, they ask "how much a racial inequality could be eliminated by intervening on a different variable," thereby shifting focus to factors that are manipulable or policy-relevant.

While it may not be feasible to experimentally alter some attributes, we can still con-

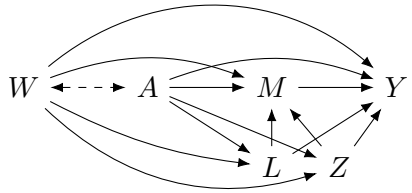
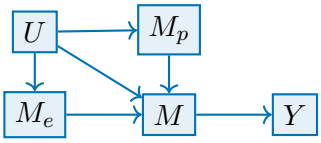
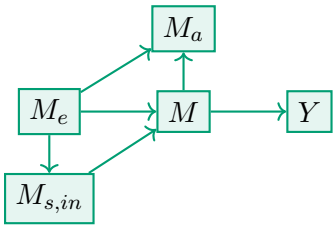



Figure 3.4: Causal diagram of the NIH peer review process. $A$: PI race ($A$=0 white, $A$=1 Black); $Y$: selection of proposals for discussion; $W$: ethnicity and gender (protected attributes not under examination); $M$: Preliminary Overall Impact Score; $Z$: application and PI characteristics (Application Type, Amended Status, Career Stage, Degree Category, and Funding Bin); $L$: contextual factors (IRG, SRG, and Administering Organization). The dotted bi-directional arrow between $W$ and $A$ indicates an association rather than a causal link.




(a) Simultaneous Evaluation Model.




(b) Normatively Backwards Evaluation Model.

Figure 3.5: Simplified DAGs representing two conceptual extremes for how criterion scores relate to Preliminary Overall Impact Scores. Panel a (Simultaneous Evaluation Model) reflects NIH documentation, where all criterion scores are evaluated in parallel and influenced by underlying PI ability ($U$). Panel b (Normatively Backwards Evaluation Model) represents a scenario where some criterion scores influence the Preliminary Overall Impact Score while others (e.g., Approach $M_a$) are evaluated afterward. In both DAGs, $M_e$ denotes applicant-related scores, $M_p$ proposal-related scores, and $Y$ selection into discussion.

sider hypothetical interventions that shift their distributions to gain insight into potential policy effects. For example, we might consider a hypothetical intervention that equalizes Preliminary Overall Impact Scores between Black and white PIs (i.e., removes any Black-white disparities in Preliminary Overall Impact Scores). Analyzing such a scenario can help us understand how much of the disparity in selection for discussion is driven by the score differences, and guide policies that would target underlying factors.

We define three estimands to quantify disparities in selection into discussion. The **Observed Disparity (OD)** (Eq. 3.1) captures the difference in discussion rates between white and Black PIs. The **Disparity Reduction (DRed)** (Eq. 3.2) represents the change in disparity under a hypothetical intervention that equalizes the distribution of an attribute (e.g., Preliminary Overall Impact Scores) between groups. To formalize this, we let $M^*$ denote a value drawn from the conditional distribution of Preliminary Overall Impact Scores for white PIs, given gender, ethnicity, IRG, SRG, and Administering Organization: $P(M \mid A = 0, W, L)$. Using the potential outcomes framework (Rubin, 1974), which allows us to define outcomes under hypothetical scenarios, we define $Y(M^*)$ as an indicator for whether a proposal would be selected for discussion after the hypothetical intervention to equalize the distribution of Preliminary Overall Impact Scores between white and Black PIs. The **Disparity Remaining (DRem)** (Eq. 3.3) captures the portion of the disparity that persists after such an intervention. Consistent with NIH's streamlining practice, all potential outcomes (e.g., $Y(M^*)$) are defined such that the discussion rate within each SRG $k$—that is, the proportion of proposals selected for discussion—is fixed at its observed value $d_k$.

$$\text{Observed Disparity: } E[Y \mid A = 0] - E[Y \mid A = 1] \tag{3.1}$$

$$\text{Disparity Reduction: } E[Y(M^*) \mid A = 1] - E[Y \mid A = 1] \quad s.t. \quad E[Y_k(M^*)] = d_k \tag{3.2}$$

$$\text{Disparity Remaining: } E[Y \mid A = 0] - E[Y(M^*) \mid A = 1] \quad s.t. \quad E[Y_k(M^*)] = d_k \tag{3.3}$$

We also consider a thought experiment in the form of a hypothetical intervention mo-

tivated by NIH's Simplified Peer Review Framework: binarizing the Investigator and Environment criterion scores. To assess its impact, we define the **thresholded disparity (TD)** (Eq. 3.4), which measures the disparity that would remain if these scores were converted from 1–9 scales to binary indicators (e.g., "appropriate" vs. "gaps identified").

$$\text{Thresholded Disparity: } E[Y(M_{i,e}^t)|A=0] - E[Y(M_{i,e}^t)|A=1] \quad s.t. \quad E[Y_k(M_{i,e}^t)] = d_k \tag{3.4}$$

Formally, $M_{i,e}^t$ denotes the binarized Investigator and Environment criterion scores (as a stacked vector), based on a predefined threshold $t$. If $M_{i,e} \leq t$, then $M_{i,e}^t = [1,1]$ (strong score); otherwise, $M_{i,e}^t = [5,5]$ (poor score). The potential outcome $Y(M_{i,e}^t)$ represents whether a proposal would be selected for discussion under this binarized scoring.

To carry out this thought experiment, we consider two conceptual extremes for how criterion scores relate to Preliminary Overall Impact Scores: the **Simultaneous Evaluation Model** and the **Normatively Backwards Evaluation Model**. The Simultaneous Evaluation Model reflects the structure implied by NIH documentation, where all criterion scores are assessed in parallel and influenced by an underlying latent PI ability ($U$). Figure 3.5a shows an example DAG for this model. In contrast, the Normatively Backwards Evaluation Model captures a scenario in which some criterion scores influence the Preliminary Overall Impact Score, while the Approach score ($M_a$) is evaluated afterward, reflecting a possible ordering or dependency in how reviewers form their assessments (Figure 3.5b).

The following assumptions are sufficient for nonparametric identification of the Disparity Reduction, Disparity Remaining, and thresholded disparity estimands using observational data:

1. **Consistency (SRG level):** For proposal $i$ in SRG $k$, if $\mathbf{M}_k = \mathbf{m}_k$, then $Y_{ki} = Y_{ki}(\mathbf{m}_k)$.

2. **No unmeasured confounding:** The effect of $M$ on $Y$ is unconfounded at the SRG level conditional on observed covariates $(W, Z, L)$.

3. **Positivity:** For all relevant values of $(W, Z, L)$, there is positive probability of observing each level of $M$.

4. **Partial interference:** Interference occurs only within SRGs, not between SRGs.

The above assumptions 1-4 are reasonable in the NIH peer review setting. The **consistency** assumption follows from the scoring process: the observed discussion outcome corresponds to the outcome under the realized score. The **no unmeasured confounding** assumption requires that, conditional on observed covariates (i.e., gender, ethnicity, IRG, SRG, and Administering Organization), there are no unobserved variables that influence both the intervened-on attribute (i.e., Preliminary Overall Impact Score) and selection into discussion. A key concern for this assumption is the presence of unobserved aspects of proposal quality or evaluation context that may affect both preliminary scores and discussion decisions. Many dimensions of proposal quality are reflected in the criterion scores and Preliminary Overall Impact Score themselves; however, it is possible that some aspects of proposal quality or evaluation context are not fully captured by these scores and remain unobserved. To mitigate confounding, we adjust for covariates selected based on the causal diagram (Fig. 3.4), including structural, application, and investigator characteristics. In particular, conditioning on IRG, SRG, and Administering Organization accounts for systematic differences in reviewer composition and evaluation standards, while application and investigator characteristics capture key observable dimensions of proposal quality and/or review context. While these adjustments address major sources of observed variation, the no unmeasured confounding assumption cannot be verified and may be violated if important aspects of proposal quality or evaluation context are not fully captured. The causal diagram (Fig. 3.4) clarifies the assumptions under which identification holds, and our results should be interpreted in light of this assumed causal structure. The **positivity** assumption is supported by the presence of both Black and white PIs across relevant subgroups in the data.

The set of covariates used for adjustment is determined by the directed acyclic graph

and corresponding identification forms. These covariates are included to account for differences in observed context that may influence both the attribute being intervened on and the outcome. Conditioning on these variables allows us to compare outcomes between Black and white PIs who are similar along these observed dimensions. For example, when evaluating interventions on preliminary impact scores, we adjust for PI characteristics (e.g., career stage, degree category, and prior funding), proposal characteristics (e.g., application type and amendment status), and structural features of the review process (e.g., IRG and administering organization), with SRG included as a random effect. Full details of the covariate sets corresponding to each intervention are provided in Tables A.6 and A.7 in the Appendix.

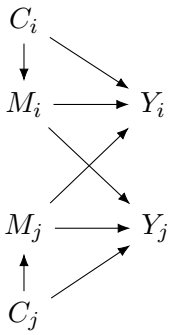



Figure 3.6: Simplified interference structure for two proposals $i$ and $j$ within a single SRG $k$.

The **partial interference** assumption reflects the NIH streamlining procedure: proposals are evaluated relative to others within the same SRG, so interventions on one proposal's score may affect discussion outcomes of other proposals in that SRG, but not those in different SRGs. This motivates a partial interference framework (Hudgens & Halloran, 2008; Ogburn & VanderWeele, 2014).

Let $n$ denote the total number of proposals and $N$ the number of SRGs. Proposals are partitioned into SRGs indexed by $k$, with $n_k$ proposals in SRG $k$. Let $\mathbf{M}_k = (M_{k1}, \ldots, M_{kn_k})$ denote the vector of scores within SRG $k$. Let $\mathbf{m} = (\mathbf{m}_1, \ldots, \mathbf{m}_N)$ denote a global assignment vector across all SRGs. For proposal $i$ in SRG $k$, define the potential outcome $Y_{ki}(\mathbf{m}_k)$ as

the outcome that would be observed under an intervention that sets $\mathbf{M}_k = \mathbf{m}_k$.

Under partial interference,

$$Y_{ki}(\mathbf{m}) = Y_{ki}(\mathbf{m}_k),$$

so outcomes depend only on interventions within SRG $k$.

To illustrate the implied dependence structure, Figure 3.6 presents a simplified DAG involving two proposals within a single SRG. For notational simplicity, the SRG index $k$ is suppressed in the figure. Let $C_{ki} = (W_{ki}, A_{ki}, Z_{ki}, L_{ki})$ denote observed covariates for proposal $i$.

The DAG encodes that both $M_{ki}$ and $M_{kj}$ may affect $Y_{ki}$, and that confounding of these relationships is captured by $C_{ki}$. Under the no unmeasured confounding assumption at the SRG level, and the absence of cross-unit confounders beyond $C_{ki}$, this implies, under the assumptions above, that

$$Y_{ki}(\mathbf{m}_k) \perp\!\!\!\perp \mathbf{M}_k \mid C_{ki},$$

which permits nonparametric identification of the causal effect of $\mathbf{M}_k$ on $Y_{ki}$ under partial interference.

Formal identification proofs for the Disparity Reduction, Disparity Remaining, and thresholded disparity estimands appear in Appendix A.5, and the effects of self-selection on these estimands are discussed in Appendix A.6.

## *3.5 Estimation*

### *3.5.1 Estimation Framework*

To estimate the Disparity Reduction, Disparity Remaining, and thresholded disparity, we consider only interventions that keep the discussion rate constant. This ensures that all aspects of review other than the distributions of the intervened attributes remain unchanged. Let $d_k$ denote the observed discussion rate for SRG $k$, and $n_k$ the number of proposals in

SRG $k$. Define

$$Y_k(\cdot) = \frac{1}{n_k} \sum_{i=1}^{n_k} Y_{ki}(\cdot)$$

as the proportion of proposals selected for discussion in SRG $k$. In order to estimate the Disparity Reduction, Disparity Remaining and thresholded disparity, we must estimate Eqs. (3.5) and (3.6).

$$\mathbb{P}\big(Y_{ki}(\mathbf{M}_k^t) = 1 \mid A_{ki} = 0\big), \quad \text{s.t.} \quad \mathbb{E}\big[Y_k(\mathbf{M}_k^t)\big] = d_k. \tag{3.5}$$

$$\mathbb{P}\big(Y_{ki}(\mathbf{M}_{e,k}^t) = 1 \mid A_{ki} = 0\big), \quad \text{s.t.} \quad \mathbb{E}\big[Y_k(\mathbf{M}_{e,k}^t)\big] = d_k. \tag{3.6}$$

Under the identifying assumptions described in Section 3.4, the Disparity Reduction, Disparity Remaining and thresholded disparity can be written as functionals of the observed data distribution. To estimate these functionals under interventions that preserve the observed discussion rate within each SRG, we introduce the *shifted estimator*.

We begin by specifying a working model for the probability of discussion. Let $C_{ki}$ denote a vector of observed covariates for proposal $i$ in SRG $k$, chosen to capture the variables required for identification, including those that account for interference within SRGs. The specific covariates included in $C_{ki}$ depend on the identification form associated with each estimand (e.g., thresholded disparity versus disparity reduction), but we use a common notation for simplicity. Following Erosheva et al. (2020a), we incorporate an SRG-specific random effect and fit a logistic regression model of the form

$$\hat{p}_{k,i} = \frac{1}{1 + \exp\big(-(\beta_0 + \beta_a A_{ki} + \beta_c^\top C_{ki} + b_k)\big)},$$

where $b_k$ is an SRG-specific random effect.

To enforce the constraint that the discussion rate in each SRG equals the observed rate $d_k$, we introduce an SRG-specific shift parameter $\gamma_k$. The shifted predicted probability is

defined as

$$\hat{p}_{k,i}^{\text{shift}} = \frac{1}{1 + \exp\left(-\left(\beta_0 + \gamma_k + \beta_a A_{ki} + \beta_c^\top C_{ki} + b_k\right)\right)}.$$

The shift parameter $\gamma_k$ is chosen to satisfy

$$\frac{1}{n_k} \sum_{i=1}^{n_k} \hat{p}_{k,i}^{\text{shift}} = d_k, \qquad k = 1, \ldots, N,$$

which ensures that, within each SRG, the average shifted predicted probability matches the observed discussion rate. Intuitively, this estimator enforces the marginal constraint by shifting the intercept, aligning the overall discussion rate with $d_k$ while preserving the relative ordering of proposals within each SRG.

Under the working model, the shifted predicted probabilities $\hat{p}_{k,i}^{\text{shift}}$ approximate the conditional expectation under the model. Averaging over the empirical distribution of the covariates yields a plug-in estimate of the corresponding marginal expectations $\mathbb{E}[Y(\mathbf{M}_e^t) \mid A = a]$ and $\mathbb{E}[Y(\mathbf{M}^t) \mid A = a]$.

Thus, the Disparity Reduction, Disparity Remaining, and thresholded disparity can all be expressed through a common plug-in form:

$$\widehat{\Delta} = \frac{1}{n_{A=0}} \sum_{k=1}^{N} \sum_{i:A_{ki}=0} \hat{p}_{k,i}^{\text{shift}} - \frac{1}{n_{A=1}} \sum_{k=1}^{N} \sum_{i:A_{ki}=1} \hat{p}_{k,i}^{\text{shift}}. \tag{3.7}$$

The specific estimand (e.g., Disparity Reduction) is determined by the definition of the shifted probabilities $\hat{p}_{k,i}^{\text{shift}}$, which depend on the corresponding intervention and identification form.

### *3.5.2 Bayesian Implementation*

We estimate Disparity Reduction, Disparity Remaining and Thesholded Disparity using a Bayesian framework, which provides a coherent way to model the joint distribution of outcomes and mediators, incorporate structural constraints required for the intervention, and propagate uncertainty through the intervention and aggregation steps. Specifically, we combine parametric models for the outcome and mediator distributions with a group-level

Bayesian bootstrap and use Monte Carlo integration to obtain posterior summaries. In order to estimate Disparity Reduction and Disparity Remaining, we follow Algorithm 1; here, we illustrate the procedure by intervening on average Preliminary Overall Impact Scores. In the first step of Algorithm 1, we model selection into discussion $Y$ using a Bernoulli likelihood, with the mean modeled by a logistic regression including fixed effects for the PI's race, ethnicity, gender, and the proposal's average Preliminary Overall Impact Scores, the mean and variance of scores from other proposals within the same SRG, application type, amended status, career stage, degree category, funding bin, IRG, and Administering Organization. SRG is included as a random effect to account for differences across SRGs (Erosheva et al., 2020a).

To account for interference, we allow a proposal's discussion outcome to depend on the scores of other proposals within the same SRG, but assume no interference across SRGs. To make estimation tractable, we do not model the full vector of other proposals' scores directly. Instead, conditional on a proposal's own score and observed covariates, we assume that the effect of other proposals' scores on selection into discussion is captured by the mean and variance of those scores within the same SRG, denoted $\mu_{M^*_{k,(-i)}}$ and $\sigma^2_{M^*_{k,(-i)}}$. This working modeling assumption reduces the dimensionality of the problem. Intuitively, we assume $\mu_{M^*_{k,(-i)}}$ and $\sigma^2_{M^*_{k,(-i)}}$ summarize the overall level and dispersion of scores among competing proposals in the same SRG. We assess the adequacy of this approximation (i.e., representing the influence of other proposals' scores using only their mean and variance) in simulation studies described in Appendix A.4.

In the second step of Algorithm 1, we model the average Preliminary Overall Impact Scores using a linear mixed-effects model with Gaussian residuals. Specifically, for proposal $i$ in SRG $k$,

$$M_{ki} \mid A_{ki}, W_{ki}, L_{ki} \sim \mathcal{N}(\mu_{ki}, \sigma^2), \quad \text{with} \quad \mu_{ki} = \beta_0 + \beta_a A_{ki} + \beta_w W_{ki} + \beta_l L_{ki} + b_k,$$

where race, ethnicity, gender, IRG, and Administering Organization are included as fixed

**Algorithm 1** Bayesian Estimation Procedure for Disparity Reduction and Remaining.

1. Model $Y$ using a Bernoulli likelihood with SRG as a random effect and other attributes as fixed effects.

2. Model $M$ using a Gaussian likelihood with SRG as a random effect.

3. For each proposal $i$ in SRG $k$:

   (a) Draw 100 samples from $M^*_{ki} \sim P(M_{ki} \mid A_{ki} = 0, W_{ki}, L_{ki})$.

   (b) For each draw, compute summaries $\mu_{M^*_{k,(-i)}}$ and $\sigma^2_{M^*_{k,(-i)}}$ using the sampled intervened scores.

   (c) Estimate

   $$P\left(Y_{ki}(\mathbf{M}^*_k) = 1 \mid A_{ki}, M_{ki} = M^*_{ki}, \mu_{M^*_{k,(-i)}}, \sigma^2_{M^*_{k,(-i)}}, Z_{ki}, W_{ki}, L_{ki}\right)$$

   for each sample and average over the Monte Carlo samples.

   - To improve efficiency, we summarize the scores of other proposals within each SRG using $\mu_{M^*_{k,(-i)}}$ and $\sigma^2_{M^*_{k,(-i)}}$, rather than conditioning on each individual score.

4. Apply the shifted estimator to the resulting predicted probabilities so that $E[Y_k(M^*)] = d_k$.

5. Jointly model all other variables using a group-level Bayesian bootstrap and compute $E[Y(M^*) \mid A = 1]$ via posterior averaging.

6. Compute Disparity Reduction and Remaining.

effects, and $b_k \sim \mathcal{N}(0, \tau^2)$ is a random effect for SRG that captures between-SRG variability. In the next step of Algorithm 1, we approximate the counterfactual outcomes under the intervention using Monte Carlo integration.

In the third step of Algorithm 1, for each proposal, we use Monte Carlo integration to average over draws of the intervened score $M^*_{ki} \sim P(M_{ki} \mid A_{ki} = 0, W_{ki}, L_{ki})$. For each draw, we compute the corresponding leave-one-out summaries $\mu_{M^*_{k,(-i)}}$ and $\sigma^2_{M^*_{k,(-i)}}$, and input these leave-one-out summaries into the logistic regression model to estimate $P\Big(Y_{ki}(\mathbf{M}^*_k) = 1 \mid A_{ki}, M_{ki} = M^*_{ki}, \mu_{M^*_{k,(-i)}}, \sigma^2_{M^*_{k,(-i)}}, Z_{ki}, W_{ki}, L_{ki}\Big)$. Averaging over the Monte Carlo samples yields proposal-level predicted probabilities. We then adjust these probabilities using the shifted estimator so that $E[Y_k(M^*)] = d_k$. In the final steps of Algorithm 1, we model the joint distribution of $W$, $L$, and $Z$ using a group-level Bayesian bootstrap at the SRG level, and compute $E[Y(M^*) \mid A = 1]$ via posterior averaging. We repeat the model fitting, intervention, and prediction steps for 1,000 draws, obtaining 1,000 estimates of the Disparity Reduction and Disparity Remaining, which are used to construct 95% credible intervals.

In Algorithm 2, we estimate the thresholded disparity using a procedure analogous to that described above for Disparity Reduction and Disparity Remaining (Algorithm 1). As described when introducing the thresholded disparity in Eq. (3.4), this estimand requires binarizing the Investigator and Environment criterion scores relative to a threshold. We therefore consider thresholds over the support of each score, applied separately to the Investigator and Environment criteria, to indicate whether each score exceeds the threshold. This allows us to assess how the Thresholded Disparities vary across different cutoff levels. In the first step of Algorithm 2, for each pair of thresholds $(t_{\text{invest}}, t_{\text{envir}})$, selection into discussion $Y$ is modeled using a Bernoulli likelihood, with the mean specified via logistic regression. To define the intervention at a given threshold pair, we threshold the Investigator

and Environment criterion scores as:

$$M_{e,ki}^{t} = \begin{cases} 1, & M_{e,ki}^{(\text{invest})} \leq t_{\text{invest}} \text{ and } M_{e,ki}^{(\text{envir})} \leq t_{\text{envir}}, \\ 5, & \text{otherwise.} \end{cases} \tag{3.8}$$

In the next step of Algorithm 2, for each SRG, we apply the thresholding defined in Eq. (3.8) to all proposals and use the resulting thresholded scores to compute summaries of the other proposals. Specifically, to obtain predicted probabilities of selection under the intervention, we include $M_{e,ki}^{t}$ along with the leave-one-out mean and variance of the thresholded scores of other proposals, denoted $\mu_{M_{e,k,(-i)}^{t}}$ and $\sigma^2_{M_{e,k,(-i)}^{t}}$, among the covariates in the logistic regression model (with the full set of covariates under the Normatively Backwards and Simultaneous Evaluation models specified in Algorithm 2). The summaries $\mu_{M_{e,k,(-i)}^{t}}$ and $\sigma^2_{M_{e,k,(-i)}^{t}}$ capture within-SRG interference. To enforce the fixed discussion rate $d_k$, we use the shifted estimator. In the final steps of Algorithm 2, we model the remaining variables jointly via a group-level Bayesian bootstrap and compute $E[Y(M_e^t) \mid A = 0]$ and $E[Y(M_e^t) \mid A = 1]$ to estimate the thresholded disparity. We carry out the model fitting, intervention, and prediction steps for 1,000 draws, obtaining 1,000 estimates of the thresholded disparity, which are used to construct 95% credible intervals.

In our Bayesian estimation framework (illustrated in Algorithms 1 and 2), we use weakly informative priors because the data are sufficiently informative for the likelihood to dominate (Gabry et al., 2019). Fixed-effect parameters in the logistic regression have $N(0, 4)$ priors, and SRG random effects follow $N(\mu_{srg_y}, \sigma_{srg_y})$ with $\mu_{srg_y} \sim N(0, 4)$ and $\sigma_{srg_y} \sim \text{Exponential}(1)$. For the linear regression modeling average scores, fixed-effect parameters have $N(0, 10)$ priors, with analogous priors for SRG random effects. Predictive posterior and MCMC diagnostic checks reveal no issues.

Moreover, in our Bayesian estimation framework, summaries such as $\mu_{M_{k,(-i)}}$ and $\sigma^2_{M_{k,(-i)}}$ are, by construction, highly correlated within SRGs. However, because our target of inference is the marginal distribution $P(Y \mid A)$, rather than conditional relationships involving

**Algorithm 2** Bayesian Estimation Procedure for thresholded disparity.

1. For each pair of thresholds $(t_{\text{invest}}, t_{\text{envir}})$, where $t_{\text{invest}}, t_{\text{envir}} \in \{1, \ldots, 9\}$:

   (a) Model selection into discussion $Y$ using logistic regression with SRG as a random effect and other attributes as fixed effects:

   - For the Normatively Backwards Evaluation Model, model

   $$Y_{ki} \mid A_{ki}, M_{e,ki}, \mu_{M_{e,k,(-i)}}, \sigma^2_{M_{e,k,(-i)}}, W_{ki}, L_{ki}, Z_{ki}.$$

   - For the Simultaneous Evaluation Model, model

   $$Y_{ki} \mid A_{ki}, M_{e,ki}, \mu_{M_{e,k,(-i)}}, \sigma^2_{M_{e,k,(-i)}}, M_{p,ki}, W_{ki}, L_{ki}, Z_{ki}.$$

   (b) Define the thresholded Investigator and Environment criterion scores by

   $$M^t_{e,ki} = \begin{cases} 1, & M^{(\text{invest})}_{e,ki} \leq t_{\text{invest}} \text{ and } M^{(\text{envir})}_{e,ki} \leq t_{\text{envir}}, \\ 5, & \text{otherwise.} \end{cases}$$

   (c) Compute summaries $\mu_{M^t_{e,k,(-i)}}$ and $\sigma^2_{M^t_{e,k,(-i)}}$, and use them together with $M^t_{e,ki}$ to obtain predicted probabilities of selection under the intervention:

   - For the Normatively Backwards Evaluation Model,

   $$P\left(Y_{ki}(\mathbf{M}^t_{e,k}) = 1 \mid A_{ki}, \mathbf{Z}_k, \mathbf{W}_k, \mathbf{L}_k, M^t_{e,ki}, \mu_{M^t_{e,k,(-i)}}, \sigma^2_{M^t_{e,k,(-i)}}\right).$$

   - For the Simultaneous Evaluation Model,

   $$P\left(Y_{ki}(\mathbf{M}^t_{e,k}) = 1 \mid A_{ki}, \mathbf{Z}_k, \mathbf{M}_{p,k}, \mathbf{W}_k, \mathbf{L}_k, M^t_{e,ki}, \mu_{M^t_{e,k,(-i)}}, \sigma^2_{M^t_{e,k,(-i)}}\right).$$

   (d) Apply the shifted estimator so that $E[Y_k(M^t_e)] = d_k$.

   (e) Jointly model the remaining variables using a group-level Bayesian bootstrap and compute $E[Y(M^t_e) \mid A = 0]$ and $E[Y(M^t_e) \mid A = 1]$ by posterior averaging.

   (f) Compute the thresholded disparity.

these summaries, such within-group dependence has limited impact on inference. To further account for this dependence, we implement a group-level Bayesian bootstrap, assigning resampling weights at the SRG level (Cameron et al., 2008; Rubin, 1981). Results closely match those from the standard Bayesian bootstrap, suggesting that within-group dependence does not substantially affect our estimates.

We conducted a simulation study to assess the statistical bias of the shifted estimator and evaluate the impact of interference by comparing estimates with and without accounting for interference. The simulations showed that, after accounting for interference, the shifted estimator is approximately unbiased. Moreover, both simulation results and empirical analyses suggest that interference due to streamlining has minimal effect on the estimates, implying that partial interference is unlikely to introduce substantial bias in this setting. Additional details are provided in Appendix A.4.

## *3.6 Results*

### *3.6.1 Causal decomposition analysis of hypothetical interventions equalizing Black-white differences in attributes*

In this section, we focus on several variables that may influence a proposal's chances of being discussed: the Preliminary Overall Impact Score, application type, amended status, degree category, funding bin, career stage, Administering Organization, and IRG, which is a proxy for scientific area. Given that distributions of these variables differ substantially by race, we investigate – via causal decomposition analyses – whether hypothetical interventions of eliminating these disparities and/or differences would influence racial disparities in selection into discussion.

Observed disparities and/or differences in the investigator and application characteristics by race are substantial. The observed distribution of mean Preliminary Overall Impact Scores differs between Black and white PIs, with white PIs tending to receive lower (better) scores. We also observe notable racial differences in other key attributes: white PIs have

a 5.1% higher rate of renewed applications, 7.9% more hold a Ph.D., while 4.3% and 3.4% more Black PIs hold an M.D. or other degrees, respectively. Additionally, 9.2% more Black PIs are Early Stage Investigators, 6.5% more are non-Early-Stage-Investigators, and 15.7% more white PIs are Experienced Investigators. Lastly, the distributions of Administering Organization and IRGs also differ between Black and white PIs. We also observe differences in the PI's institution's funding bin which is a proxy for its prestige, where lower bins indicate higher prestige (Erosheva et al., 2020a). These differences highlight potential contributors to disparities in selection for discussion and motivate the use of hypothetical interventions to assess whether targeting these factors could reduce those disparities. Tables and figures in Appendix A.1 provide further details on these comparisons.

We examine several hypothetical interventions eliminating racial disparities and/or differences in: (1) mean Preliminary Overall Impact Score, (2) application type and amended status jointly, (3) degree category, funding bin, and career stage jointly, and (4) Administering Organization and IRG jointly. These interventions are motivated by observed racial differences in the distributions of these attributes and are grounded in previous research that examined success in securing NIH funding (Erosheva et al., 2020a; Ginther et al., 2011; Hoppe et al., 2019; Lauer et al., 2021). Ginther et al., 2011 found that those with strong priority scores (scores given after SRG panel discussion), were equally likely to receive funding, irrespective of the PI's race. Hoppe et al., 2019 found significant marginal effects for (1) application resubmission, with new applications being significantly less likely to be discussed and awarded than resubmitted applications; (2) institution prestige, with applications from less prestigious institutions (those that receive less funding) being significantly less likely to be discussed and awarded; (3) career stage, with early career investigators being significantly more likely to be discussed and awarded; and (4) topic choice, suggesting that preferences for topics "less likely to excite the enthusiasm of the scientific community" could lead to lower discussion and funding rates, potentially creating a vicious cycle. Subsequently, Lauer et al., 2021 refuted Hoppe et al., 2019's claim that topic choice is the primary reason for the funding gap and instead suggested that differences in funding rates between Institutes and

Centers to which PI applications are assigned play a larger role.

We estimate the expected Disparity Reduction and Disparity Remaining in selection into discussion under hypothetical interventions aimed to equalize the attribute(s) of interest among Black and white PIs (Jackson & VanderWeele, 2018; VanderWeele & Robinson, 2014). To do this, we randomly assign the attribute(s) under consideration among Black PIs such that they follow the same distribution as white PIs conditional on gender, ethnicity, IRG, SRG, and Administering Organization. (If the intervention under consideration is for IRG and Administering Organization jointly, we do not include IRG, SRG, and Administering Organization among conditioning variables). To estimate the disparity reduced, we use a causal decomposition analysis, assuming the causal structure in Fig. 3.4 and accounting for SRG-level discussion rates, so that comparisons are made within an SRG-constrained selection process. We use logistic regression to model the selection process. We use either linear or binary/multinomial logistic regressions, as appropriate, to model the attribute(s) under consideration. For the joint interventions, we first model the distribution of one of the attributes and then model the conditional distribution of the other attributes given the previously modeled attributes.

Under the assumptions of the causal decomposition framework, our findings, summarized in Fig 3.7, indicate that equalizing mean Preliminary Overall Impact Scores would dissolve racial disparities (Disparity Remaining Estimate = 0.1%, 95% Cred. Interval = [-2.7%, 2.9%]); however, equalizing distributions of application type, amended status, degree category, funding bin, career stage, IRG, and/or Administering Organization by race results in minimal reductions in the disparity in selection into discussion.

### *3.6.2 Thought experiment on binary Investigator and Environment scoring under the Simplified Peer Review Framework*

On January 25th, 2025, NIH implemented the Simplified Peer Review Framework (NIH Center for Scientific Review, 2023b). Tables 3.1 and 3.2 demonstrate the key differences between the Enhanced Peer Review Process and Simplified Peer Review Framework. A

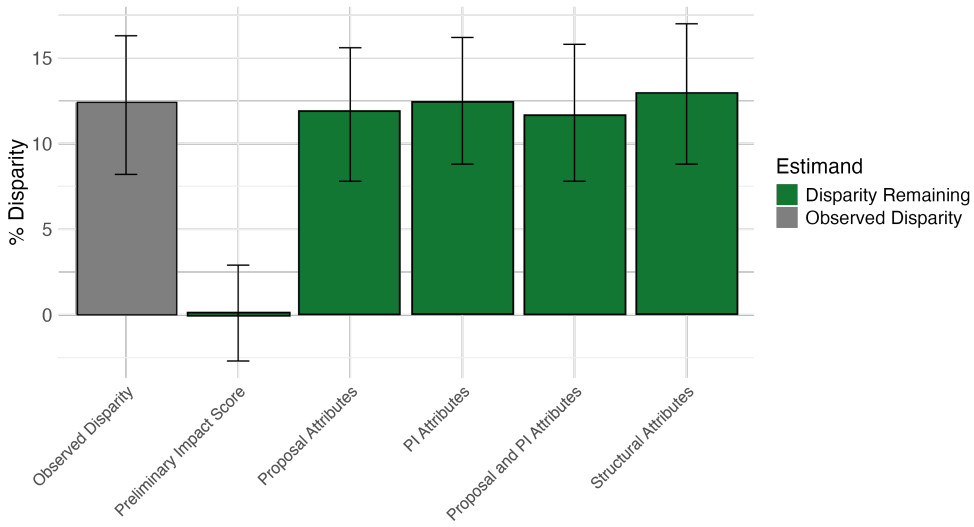



Figure 3.7: Estimated Overall Disparity and Disparity Remaining. Estimated Disparity Remaining for Preliminary Overall Impact Scores, proposal attributes (application type and amended status), PI attributes (career stage, degree category, and institutional funding bin – a proxy for institution prestige), and structural attributes (IRG and Administering Organization). Positive values indicate disparity against Black PIs. Included 95% Credible Intervals. Estimates and intervals were obtained using Bayesian estimation.

key simplification is combining the evaluation of the Investigator and Environment into a single binary score, labeled as Factor 3, which will be rated as either "appropriate" or "gaps identified". This "will help put applicants on equal footing", aiming to "mitigate the effect of reputational bias" (NIH Center for Scientific Review, 2024a), a plausible contributor to Black-white disparities (Ginther et al., 2011). Although we cannot fully assess changes due to combining the Investigator and Environment into a single binary score, we can use our causal decomposition framework to assess effects of binarization of both Investigator and Environment criterion scores on racial disparities in selection into discussion. Assuming that reviewers determine binary appropriateness based on the previous score being below a certain threshold, we consider a thought experiment by binarizing the Investigator and Environment scores separately at those thresholds while keeping all other criterion scores as they are. We do not know exactly where the psychological threshold for "appropriate" versus "gaps identified" will fall relative to the previous 9 point scale, so we perform a range of analyses with different thresholds. If the threshold is very high (recall that higher scores on the 9 point scale are worse), almost all proposals will be marked as "appropriate", essentially leading to parity. Similarly, if the threshold is very low, almost all proposals will be marked as "gaps identified", also resulting in parity. In what follows, we refer to the Investigator and Environment criterion scores as applicant-related scores, and the Significance, Innovation, and Approach criterion scores as proposal-related scores.

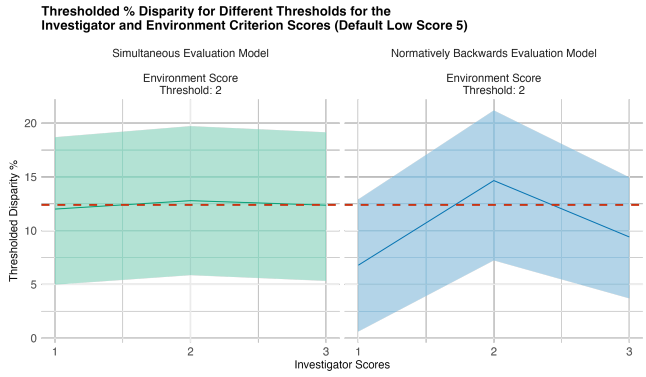
We propose a thought experiment where reviewers map their numerical scores (1-9) for rating Investigator and Environment criteria onto binary "appropriate" or "gaps identified" decisions. Each of the three reviewers would give one of these two designations. To model the impact onto discussion, we assume that "appropriate" would correspond to a numeric Investigator/Environment score that indicates the proposal will likely proceed to be discussed. Likewise, "gaps identified" would correspond to a score that reflects a low chance of selection into discussion. As seen in Table A.1, proposals with scores of 1 are often discussed, and proposals with scores of 5 and above are rarely discussed. Therefore, we map a score of 1 to an "appropriate" decision and a score of 5 to a "gaps identified" decision. When we

average these binary decisions across the three reviewers, we get the following scores: 1 (all three reviewers find the proposal "appropriate"), 2.33 (two reviewers select "appropriate" and one selects "gaps identified"), 3.67 (one reviewer selects "appropriate" and two select "gaps identified"), and 5 (all three reviewers select "gaps identified"). Based on the distribution of Investigator and Environment criterion scores (Fig A.1), scores of 1 and 2.33 are more common, while 3.67 and 5 occur less frequently. Table A.1 shows that proposals with scores around 1 or 2.33 are more likely to be discussed.

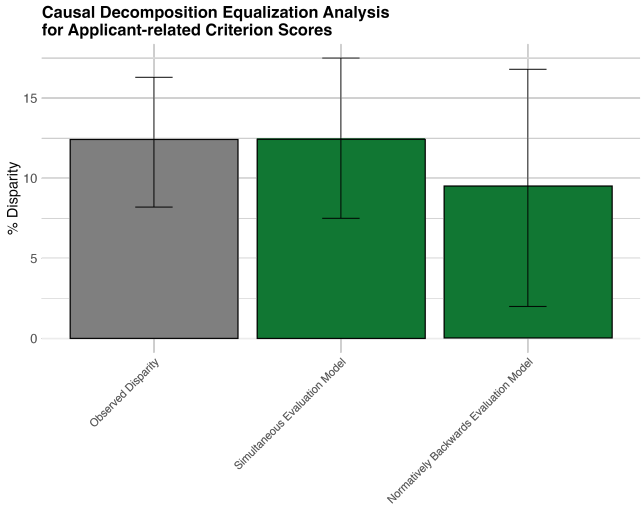
We implement this thought experiment under two conceptual models introduced in Section 3.4 (Figure 3.5): the *Simultaneous Evaluation Model*, consistent with NIH documentation, and the *Normatively Backwards Evaluation Model*, which allows for sequential assessment of criteria.

Fig 3.8a shows results for the Environment criterion score threshold of 2 and Investigator criterion score thresholds from 1 to 3. We focus on this range because higher thresholds are less realistic given historical NIH discussion rates. The combination of an Investigator criterion score threshold of 3 and Environment criterion score threshold of 2 is most plausible, as approximately 50% of proposals with these scores or lower are discussed. We note that thresholded disparities are consistent across Environment criterion score thresholds; therefore, we present results only for the Environment criterion score threshold of 2. Complete results over the full range of threshold combinations are provided in the Appendix (Fig A.7). In what follows, we offer further explanation and intuition for our findings under both the Simultaneous and Normatively Backwards Evaluation Models.

For the Simultaneous Evaluation Model, we use the estimand identified in Eq. S9. Notably, this model conditions on proposal-based criterion scores. Under this model, there is minimal variation across different Investigator and Environment criterion score thresholds because, conditional on the proposal-related criterion scores, the applicant-related criterion scores have minimal impact on selection into discussion, as indicated by the coefficients in the logistic regression model. As a result, regardless of whether reviewers were strict, moderate, or lenient in judging proposals as "acceptable" versus "gaps identified," disparities in




(a) Thresholded disparity results.




(b) Equalization results for Applicant-related Criterion Scores.

Figure 3.8: **Thresholded disparity Results:** Each plot shows the percentage of thresholded disparity for an Environment criterion score threshold of 2, with the x-axis representing varying Investigator criterion score thresholds (1-3), based on historical NIH discussion rates. The blue line represents the thresholded disparity under Model 3.5b, with the blue shaded region indicating the 95% Credible Intervals. The green line represents the thresholded disparity under Model 3.5a, with the green shaded region indicating the corresponding 95% Credible Intervals. The red dotted line marks the overall observed disparity, which is 12.3% (11.3%, 16.3%). For both models, applications with scores below the Environment and Investigator criterion score thresholds were treated as if they had received a score of 5. **Equalization Results for Applicant-Related Criterion Scores:** This section presents the estimated overall disparity and Disparity Remaining after equalizing applicant-related criterion scores under the Simultaneous and Normatively Backwards Evaluation Models. Positive values indicate disparity against Black PIs. Estimates and their 95% Credible Intervals are obtained using Bayesian estimation.

selection into discussion remain largely unchanged from the originally observed levels.

For the Normatively Backwards Evaluation Model, we use the estimand from Eq. S8. Notably, this model does not account for proposal-based criterion scores. Our results show that, if reviewers were very strict, deeming only top-tier Investigator scores as "appropriate", disparities in selection into discussion would become slightly lower than observed, as this strictness would apply to all PIs. Our exploratory analysis supports these results (see Table A.1), as substantial disparities still exist at an Investigator criterion score of 1. If reviewers were to use the Investigator criterion score threshold of 2, meaning that proposals with Investigator criterion scores of 1 or 2 would be considered "appropriate" while scores above 2 would indicate "gaps identified," this would result in the largest disparities in selection into discussion compared to all other thresholds. This finding can be traced back to data (see Table A.1) that shows that most white PIs receive scores of 1 or 2, while many fewer Black PIs achieve these scores, resulting in the largest Black-white disparity by the Investigator criterion score breakdowns. If reviewers were to use the Investigator criterion score threshold of 3, the disparity in selection into discussion would be similar to that at the threshold of 1. This is because, as shown in Table A.1, the disparity at a score of 3 is the same as that at a score of 1. For higher thresholds (not shown in Fig 3.8a, refer to Fig A.4), disparities decrease slightly and level out for thresholds of 5 and above, as few PIs score above 4. Overall, if reviewers were to binarize Investigator and Environment criterion scores, the expected thresholded disparity differs from the observed disparity, but not substantially, especially for the relevant thresholds of 1 to 3.

We connect the binarization analysis to the equalization analysis by noting that extremely lenient/strict thresholds effectively treat nearly all Investigator and Environment criterion scores as "acceptable"/"gaps identified" for both Black and white PIs, thereby equalizing their distributions. The findings from our binarization analysis align with those from hypothetical interventions that equalize applicant-related criterion scores, as shown in Fig 3.8a. The identification formulas for these equalization interventions under the Simultaneous and Normatively Backwards Evaluation Models are provided in Table A.6. Under the Simul-

taneous Evaluation Model, equalization has no impact on disparities – consistent with the binarization analysis results. Under the Normatively Backwards Evaluation Model, equalization leads to small, but not statistically substantial, reductions in disparities, again echoing the binarization analysis results. These results indicate that even under extreme thresholds, where applicant-related scores are effectively equalized across groups, disparities in selection into discussion persist.

Overall, under both the Normatively Backwards and Simultaneous Evaluation models, our thought experiment—simulating the NIH's shift to binary Investigator and Environment scores—did not lead to a meaningful reduction in disparities. The joint evaluation of Investigator and Environment falls into the patterns described in Fig 3.2. Our conclusions are conditional on the assumed plausible mappings between scoring systems and reviewer behavior, as well as on the assumptions of the causal decomposition framework. Under these conditions, the transition to the Simplified Peer Review Framework is unlikely to substantially reduce disparities.

### *3.7 Discussion*

Following on Hoppe et al., 2019's finding of substantial Black-white disparity in selection into discussion, we employ a causal decomposition methodology (1) to evaluate what factors drive Black-white disparities in selection for panel discussion and (2) to assess the potential impact of the move to the Simplified Peer Review Framework on these disparities.

In our work, we illustrate that there is quite a bit of variability in both discussion rates and amount of disparity across NIH's Integrated and Scientific Review Groups. This variability, along with NIH's documented practices of "rescuing" applications for discussion (NIH NIAD, 2023), suggests that selection for discussion is not entirely deterministic based on Preliminary Overall Impact Scores (see Appendix A.2).

Nonetheless, under the assumptions of the causal decomposition framework, our analysis suggests that eliminating racial disparities in Preliminary Overall Impact Scores can potentially dissolve disparities in selection into discussion. This could be because the selection

process is either primarily driven by Overall Impact scores and/or because the variability in discussion selection practices is too infrequent to detect differences. Given the causal influence of Preliminary Overall Impact Scores on selection into discussion, future work should focus on identifying upstream, policy-relevant, and practically feasible interventions for closing disparities in these scores. In the context of NIH grant review, experiments have provided mixed results about the impact of interventions on perceived PI race on Preliminary Overall Impact scores: on the one hand, an experiment that redacted names on proposals submitted by Black, matched white, and random white PIs (400 proposals for each group) found that anonymization halved the difference in mean Preliminary Overall Impact scores for proposals submitted by Black versus white PIs (Nakamura et al., 2021); in contrast, a Goldberg-style study (Goldberg, 1968) that swapped the perceived-race of applicant names on otherwise identical applications (which included 24 funded and 24 unfunded NIH grant proposals) found little to no difference in assigned Preliminary Overall Impact Scores (Forscher et al., 2019). Future research is needed to resolve these discrepancies and to evaluate whether inequities in Preliminary Overall Impact Scores could be causally explained by structural inequities that have been found to account for downstream gaps in funding, including inequities in bibliometric profiles (Ginther et al., 2011) and co-author networks (Ginther et al., 2011; Warner et al., 2017). Note that research topics disproportionately studied by Black PIs is not listed here because, although such proposals receive less funding, this was found to be due to differences in funding award rates at the Centers and Institutes to which they are assigned rather than differences in peer review scores (Lauer et al., 2021).

Under the assumptions of the causal decomposition framework, our analysis also finds that eliminating disparities in application type, amended status, degree category, institution prestige, career stage, Integrated Review Groups (IRGs) and/or Administering Organization has little potential to reduce disparity in selection into discussion. This finding is akin to that from an earlier correlational analysis that found that preliminary criterion scores fully explained racial disparities in Preliminary Overall Impact Scores (Erosheva et al., 2020a).

In our thought experiment, the causal decomposition analysis indicates that the impact

of moving to binary scoring for the Investigator and Environment criteria in the new Simplified Peer Review Framework depends on how reviewers evaluate the criteria as well as on the underlying causal relationship between the criterion scores and Preliminary Overall Impact Scores. According to our thought experiment, if the process operates as intended in NIH documentation, binarization produces essentially no change in Black–white disparities. Under an alternative, backward evaluation of scores, binarization can have some effect, but the changes supported by the data were not substantial. Under the assumptions of the causal decomposition framework and the scoring scenarios considered in our thought experiment, the results suggest that while efforts to eliminate racial disparities in Preliminary Overall Impact Scores may greatly reduce disparities in selection for discussion, transition to the Simplified Peer Review Framework is unlikely to do so.

Our study has several important limitations. First, due to data privacy restrictions, we did not have access to the full set of applications from 2014 to 2016. Our dataset is only representative of those applications with complete demographic information. We cannot be certain that applicants who did not provide their demographic information were not different from those that did. Second, our thought experiment related only to binarizing the Investigator and Environment criteria scores in the Simplified Peer Review Framework. While it is conceptually straightforward to imagine that reviewers might still assess Investigator and Environment separately before forming a combined binary judgment (see Fig 3.2), it is not clear how to operationalize the mapping for another important change of combining Significance and Innovation scores into one single Factor 1 score on a 1-9 scale. Third, our analyses relied on patterns in data collected under the previous Enhanced Peer Review framework, and one cannot be certain that data patterns that were present under NIH's Enhanced Peer Review will persist under the new Simplified Peer Review Framework.

Lastly, while our study includes important applicant- and application-level factors such as career stage, institutional prestige, and whether an application was amended, other potentially influential factors are missing. In particular, we do not have bibliometric profiles (Ginther et al., 2018) or mentorship network measures for applicants (Warner et al., 2017),

which have been found to explain a significant portion of the Black/white R01 funding gap. While we believe the variables we included are sufficient to identify our causal quantities of interest, incorporating additional factors may improve the precision and robustness of our estimates.

This chapter demonstrates how a causal decomposition framework could be used to pilot policy initiatives via specifying the updated estimands and identification assumptions and verifying whether those assumptions hold in practice as we did in the second part of the results section. Although the analyses in this chapter focused specifically on selection of proposals for panel discussion, our causal decomposition framework could be extended to study hypothetical interventions at other stages such as funding decisions, if the assumptions outlined in Section 3.4 remain reasonable.

Overall, conditional on the assumptions of the causal decomposition framework and on the assumed plausible mappings between scoring systems and reviewer behavior, our findings suggest that interventions such as binarizing the Investigator and Environment criterion scores, as proposed in the Simplified Peer Review Framework, are unlikely to improve Black–white disparities in selection into discussion. We find that, among the proposal and investigator characteristics considered, racial disparities in Preliminary Overall Impact Scores play the central role, suggesting that future work should focus on identifying upstream, policy-relevant, and practically feasible interventions that could reduce disparities in these scores and move selection into discussion toward parity.

# Chapter 4

# HYPOTHESIS TESTING FOR FUNCTIONAL CAUSAL DISCOVERY

This work was conducted in collaboration with Dr. Elena A. Erosheva and Dr. Fan Xia.

## *4.1 Introduction*

Despite their widespread use, functional causal discovery methods remain limited by the absence of statistical guarantees—uncertainty quantification, error control, and consistency properties that form the basis of statistical inference (e.g., Wasserman, 2010). In particular, many causal discovery methods lack accompanying inferential frameworks: they rely on deterministic outcomes by returning a single, potentially misleading point estimate of the causal direction. In addition, functional causal discovery depends on strong assumptions—such as specific functional forms—to identify the causal model from population information (e.g., Hoyer et al., 2009; Shimizu et al., 2006). As these assumptions are likely to be violated in practice to various degrees, there is a clear need for inferential tools that quantify uncertainty and provide information about the extent of possible assumption violations.

In this chapter, we begin by providing the first focused review of statistical guarantees in functional causal discovery, a family of causal discovery methods that identify causal direction by positing specific functional relationships between variables. Existing surveys of the literature describe model classes and algorithms, but to our knowledge, none systematically examine whether current methods provide uncertainty quantification, error control, or tools for diagnosing assumption violations. Our review shows that these inferential components—central to classical statistical methodology—receive limited attention in existing functional causal discovery approaches. To address these gaps, we then propose a test-based framework

for functional causal discovery.

In Section 4.2, we review the state of statistical guarantees provided by existing functional causal discovery methods. In Section 4.3, we introduce a principled way to assess causal discovery evidence and to recognize when the data do not support a definitive causal direction. We do so by repurposing goodness-of-fit and independence tests into a formal hypothesis-testing procedure for causal direction, which we term the *test-based* approach for functional causal discovery. In Section 4.4.2, we demonstrate the use and behavior of the test-based approach through simulations that vary the degree of assumption violation, as well as through real-data applications. Finally, in Section 4.5, we situate the test-based approach within the broader causal discovery literature, offer practical guidance for applied researchers, and discuss directions for future work on strengthening statistical guarantees in functional causal discovery.

## *4.2 Current State of Statistical Guarantees in Functional Causal Discovery*

We summarize the state of statistical guarantees provided by existing functional causal discovery methods in Table 4.1. To simplify the exposition, we view the current literature as comprising several distinct components. The first group are papers presenting core methodological frameworks such as the Linear Non-Gaussian Acyclic Model (LiNGAM), Additive Noise Model (ANM), and Post-Nonlinear (PNL) model (Peters et al., 2014; Shimizu et al., 2006; Zhang & Hyvarinen, 2012). The second group is comprised of papers that focus on software implementations of these methods, providing practical tools for applied researchers (Ikeuchi et al., 2023; Kalainathan et al., 2020; Kalisch et al., 2012; Zheng et al., 2024). The third group are papers that aim to establish formal statistical guarantees for causal discovery (Komatsu et al., 2010b; Thamvitayakul et al., 2012a; Wang et al., 2025). The fourth group is a set of benchmarking and applied studies that use these methods on real datasets to assess their empirical performance (Hu et al., 2018; Jiao et al., 2018; Mooij et al., 2016a; Motokawa et al., 2020; Rosenström et al., 2012; Song et al., 2017). We consider the model class each paper addresses to understand the extent to which these statistical guarantees apply within

Table 4.1: Current state of statistical guarantees in functional causal discovery. Our proposed test-based approach is included and boxed for emphasis. A check (✓) indicates the feature is explicitly supported; a cross (✗) indicates it is not; a triangle (△) indicates partial or limited support.

| | Inference | | | | Decision | | Model |
|---|---|---|---|---|---|---|---|
| **Method / Paper** | Any inferential quantity | Formal inferential framework | Multiple forms of inference | Impact of assumption violations understood | Deterministic decision enforced | Allows inconclusive outcome | Model class coverage |
| **LiNGAM / ANM / PNL** | | | | | | | |
| Shimizu et al., 2011 | ✗ | ✗ | ✗ | ✗ | ✓ | ✗ | LiNGAM |
| Hoyer et al., 2009 | ✓ | ✗ | ✗ | △ | ✗ | ✓ | ANM |
| Peters et al., 2014 | ✓ | ✗ | ✗ | △ | △ | △ | ANM |
| Zhang and Hyvarinen, 2012 | ✓ | ✗ | ✗ | △ | ✗ | ✓ | PNL |
| **Software** | | | | | | | |
| Ikeuchi et al., 2023 | ✓ | ✗ | ✗ | ✗ | △ | △ | LiNGAM |
| Kalisch et al., 2012 | ✗ | ✗ | ✗ | ✗ | ✓ | ✗ | LiNGAM |
| Kalainathan et al., 2020 | ✗ | ✗ | ✗ | ✗ | ✓ | ✗ | ANM |
| Zheng et al., 2024 | ✓ | ✗ | ✗ | ✗ | △ | △ | LiNGAM/ANM/PNL |
| **Inferential Tools** | | | | | | | |
| Komatsu et al., 2010b | ✓ | ✓ | ✗ | ✗ | ✓ | ✗ | LiNGAM |
| Thamvitayakul et al., 2012a | ✓ | ✓ | ✗ | ✗ | ✓ | ✗ | LiNGAM |
| Wang et al., 2025 | ✓ | ✓ | ✗ | ✓ | ✗ | ✓ | LiNGAM/ANM |
| **Test-based Approach (proposed)** | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | LiNGAM/ANM/PNL |
| **Benchmarks/Applications** | | | | | | | |
| Mooij et al., 2016a | ✓ | ✗ | ✗ | △ | △ | △ | ANM |
| Rosenström et al., 2012 | ✓ | ✓ | ✗ | ✗ | ✓ | ✗ | LiNGAM |
| Motokawa et al., 2020 | ✓ | ✓ | ✗ | ✗ | ✓ | ✗ | LiNGAM |
| Jiao et al., 2018 | ✓ | ✗ | ✗ | △ | △ | △ | ANM |
| Hu et al., 2018 | ✓ | ✗ | ✗ | △ | △ | △ | ANM |
| Song et al., 2017 | ✓ | ✗ | ✗ | ✗ | ✓ | ✗ | ANM/PNL |

*Notes.* "Any inferential quantity" includes $p$-values, bootstrap frequencies, or confidence sets. "Formal inferential framework" means that these quantities are developed and interpreted within a formalized inferential framework, such as a hypothesis-testing or confidence-set procedure. "Multiple forms of inference" indicates that multiple complementary inferential tools are integrated within a coherent framework, as recommended for reliable inference by Wasserstein et al. (2019). △ denotes partial or limited discussion or support.

the broader domain of functional causal discovery.

The papers in Table 4.1 are intended to be representative rather than exhaustive lists, with the level of coverage varying by category: the methodological papers highlight the core model families, the software papers reflect widely used implementations, the inferential papers largely capture the existing formal work, and the applied studies illustrate case studies in the literature.

We discuss statistical guarantees by distinguishing whether the methodology relies on a clearly defined inferential quantity—such as a $p$-value, confidence interval, or bootstrap frequency—and has a well-specified inferential framework—such as a formal hypothesis testing or a bootstrap framework—that situates the inferential quantity within a broader statistical inference structure.

The importance of both clearly defined inferential quantities and coherent inferential frameworks echoes broader guidance for statistical practice. The American Statistician's statements on $p$-values (Wasserstein & Lazar, 2016; Wasserstein et al., 2019) emphasize that reliable inference is achieved by integration of multiple complementary tools within a framework that makes assumptions and uncertainty explicit, rather than by reliance on "using bright-line rules for justifying scientific claims or conclusions" which "can lead to erroneous beliefs and poor decision making" (Wasserstein et al., 2019, p. 2). In the context of causal discovery, reliance on a single inferential device—as a $p$-value, test statistic, or confidence set—may therefore limit interpretability and robustness. Accordingly, we assess whether each method incorporates multiple forms of inference, consistent with recommended best practices (Wasserstein et al., 2019).

We further distinguish between methods that always enforce a deterministic conclusion—typically by comparing a single numerical summary such as a test statistic or a p-value—and methods that allow for an inconclusive outcome when the available evidence is insufficient. The ASA's Statement on $p$-values (Wasserstein & Lazar, 2016) emphasizes that a $p$-value alone does not measure the strength of evidence nor the probability that a hypothesis is true, underscoring the broader principle that inferential summaries should not be used as

automatic decision rules in isolation. From this perspective, methods that accommodate inconclusive results more faithfully reflect the uncertainty inherent in statistical inference, whereas deterministic rules risk overstating the evidence when assumptions are violated and/or when the signal is weak.

The core methodological frameworks (LiNGAM/ANM/PNL) provide population-level identifiability and, for some procedures, asymptotic consistency guarantees (Mooij et al., 2016a; Peters et al., 2014; Shimizu et al., 2006; Zhang & Hyvarinen, 2012). However, these results rely on assumptions that might be often violated in practice. The core LiNGAM papers (Shimizu et al., 2011, 2006) do not define a formal inferential quantity or framework; instead, they enforce a directional decision by comparing dependence or independence measures—such as mutual information—used as comparative "scores."

The use of statistical tests to assess residual–predictor independence, and to permit inconclusive causal discovery outcomes when both directions are rejected or fail to reject, has been discussed in the ANM and PNL contexts (Hoyer et al., 2009; Zhang & Hyvarinen, 2012). However, these discussions remain largely conceptual: the testing framework itself is not formally developed, and the meaning of each test outcome—particularly how rejection or non-rejection should be interpreted for causal direction identification—is not clearly specified. This lack of a well-defined inferential structure has led to heuristic uses of $p$-values in several studies and software implementations, such as selecting the direction with the larger $p$-value (Peters et al., 2014), rather than interpreting the $p$-values within a principled statistical testing framework.

Overall, none of the core methodological papers we reviewed formalize an inferential framework or incorporate multiple complementary inferential tools. They tend to either enforce a deterministic directional decision (Peters et al., 2014; Shimizu et al., 2011, 2006) or, at most, contain limited discussion of assumption violations and potential inconclusive outcomes (ANM/PNL) (Hoyer et al., 2009; Peters et al., 2014; Zhang & Hyvarinen, 2012).

Existing software implementations (e.g., (Kalainathan et al., 2020; Kalisch et al., 2012)) often follow a similar sequence of steps. That is, for a given a bivariate dataset of size $N$

on a pair of variables $X$ and $Y$, models are first fitted separately in each direction ($X \to Y$ and $Y \to X$); second, a dependence measure between the predictor and residuals (e.g., mutual information or Hilbert Schmidt Independence Criterion (HSIC) (Gretton et al., 2008)) is computed; and finally, the direction with the smaller dependence is taken as "causal." These procedures provide no quantification of uncertainty. While some implementations report per-direction $p$-values (Ikeuchi et al., 2023; Zheng et al., 2024), it is left to the practitioner whether to interpret them inferentially or simply use them as comparative scores, as recommended by Peters et al. (2014).

There is a paucity of work aimed at building inferential tools for causal discovery (e.g., Komatsu et al., 2010b; Thamvitayakul et al., 2012a; Wang et al., 2025). The existing work in this area can largely be grouped into two categories: methods that estimate bootstrap-type reliability measures and those that construct confidence sets for causal structures or directionality.

Komatsu et al. (2010b) extended the LiNGAM framework with a multiscale bootstrap procedure to assess the stability of estimated causal orderings, which is implemented in the `lingam` Python package (Ikeuchi et al., 2023). Their algorithm repeatedly refits LiNGAM on bootstrap samples, computes the mutual information between residuals as a dependence measure, and estimates the bias-corrected selection probability with which the same ordering is recovered. The resulting values represent approximately unbiased bootstrap probabilities that indicate how consistently LiNGAM selects a given causal ordering across resampled datasets. Similarly, Thamvitayakul et al. (2012a) adopt a resampling-based approach within the Direct-LiNGAM algorithm to estimate the uncertainty of directional signals, quantifying how frequently a particular edge orientation appears across bootstrap samples. While these methods represent important early steps toward incorporating uncertainty into causal discovery, they remain limited: the resulting probabilities depend on the assumed LiNGAM model class and can be misleading when the underlying assumptions are violated.

Among existing methods that produce confidence sets for causal orderings, Wang et al. (2025) provides one of the most general inferential frameworks currently available for un-

certainty quantification in causal discovery. Their method applies to identifiable structural equation models with additive errors, encompassing both linear and nonlinear additive-noise models. Wang et al. (2025) construct a confidence set of causal orderings by inverting a goodness-of-fit test. In the bivariate case, the set may contain one direction, neither, or both, reflecting whether the data provide sufficient evidence to distinguish the models. The procedure also offers some insight into model-class misspecification: an empty confidence set indicates that the "model class does not capture the data-generating process" (Wang et al., 2025, p. 2). However, this discussion is limited to that specific case, and further examination of other types of assumption violations could broaden the framework's interpretability. Conceptually, this approach parallels a hypothesis-testing framework, since confidence sets and tests are dual procedures. However, it does not currently extend to the PNL model class, where noise enters non-additively.

Turning to the final group of papers—benchmarking and applied studies—we examine how these works handle inference and decision-making in practice. For example, Mooij et al. (2016a) present the 103 cause–effect benchmark pairs for causal discovery and apply ANM-based algorithms to these pairs. They provide some discussion of $p$-values and inconclusive outcomes but do not describe any formal inferential framework. In particular, they discuss using the HSIC independence measure as a heuristic score, selecting the model with the smaller HSIC estimate, rather than interpreting the corresponding $p$-values within a principled inferential structure (Mooij et al., 2016a).

Some applications use inferential tools to quantify uncertainty; however, these papers inherit the same limitations as the underlying inferential procedures. For example, several applications of LiNGAM-based methods exist in practice (e.g., (Kotoku et al., 2020; Motokawa et al., 2020; Rosenström et al., 2012; Saito & Fujiwara, 2023)), and some of these employ bootstrap methods to quantify uncertainty (e.g., (Motokawa et al., 2020; Rosenström et al., 2012)). Yet, as previously discussed, bootstrap-based approaches can be misleading when model assumptions are violated. Similarly, some applications of ANM and PNL-based methods use permutation tests to obtain $p$-values (e.g., Hu et al., 2018; Jiao et al., 2018).

In these approaches, permutation is used to approximate the null hypothesis that no causal direction is favored: by breaking the directional relationship between the variables while preserving their marginal distributions—whether by permuting one variable, as in Jiao et al. (2018), or permuting joint samples, as in Hu et al. (2018)—the permuted datasets mimic settings in which neither functional model holds. The directional models are then re-fit and the dependence statistics are recomputed on each permuted dataset to form empirical null distributions against which the observed statistics are compared to obtain $p$-values. However, these works do not explicitly discuss how each decision outcome should be interpreted within a formal hypothesis-testing framework, and hypothesis testing for causal discovery remains largely informal and often only implicitly addressed (Hoyer et al., 2009). Other applications (e.g., Song et al., 2017) report $p$-values but rely on direct $p$-value comparisons to determine the favored direction.

Based on our review, we identify a clear need for a formal inferential framework that can be paired with any functional causal discovery model class (LiNGAM, ANM, or PNL). Such a framework should integrate multiple complementary inferential tools within a coherent structure that makes both assumptions and uncertainty explicit. In this chapter, we aim to address this gap through our proposed *test-based approach*.

## *4.3 A Hypothesis-Testing Framework for Bivariate Functional Causal Discovery*

In this section, we develop a framework for functional causal discovery that is grounded in classical hypothesis testing. This framework, which we term the *test-based approach*, repurposes standard goodness-of-fit and independence tests into a formal procedure for causal direction detection and inference. As summarized in Table 4.1, the test-based approach provides a formal inferential framework that integrates multiple complementary forms of inference and includes an explicit "inconclusive" outcome whose behavior under assumption violations is well understood. It can be paired with any major functional causal discovery model class, including PNL and its special cases, LiNGAM and ANM. These aspects place

it in a favorable position within the broader literature in terms of its statistical guarantees.

Given a bivariate dataset of size $N$ on a pair of variables $X$ and $Y$, the test-based framework infers directionality through setting up two hypothesis tests, one for each direction, each conducted with controlled Type I error. In addition to the Type I error control provided by the underlying hypothesis tests, we introduce a resampling-based procedure that estimates and constructs confidence intervals for the frequencies with which the test-based approach yields each causal discovery outcome—(a) rejecting the tests for both directions $X \to Y$ and $Y \to X$, (b) failing to reject both tests, (c) rejecting only the test for $X \to Y$, or (d) rejecting only the test for $Y \to X$. This procedure adds an additional layer of uncertainty quantification beyond the formal guarantees of the tests themselves. Consequently, the test-based approach provides formal statistical guarantees for functional causal discovery, whereas most existing methods yield only deterministic decisions without any accompanying assessment of uncertainty (e.g., Peters et al., 2014; Shimizu et al., 2011). We illustrate the methodology when it is paired with the bivariate Post-Nonlinear (PNL) model class, which encompasses the Additive Noise Model (ANM) and LiNGAM as special cases.

### *4.3.1 Setup*

Suppose we have a dataset of size $N$ on a pair of variables $X$ and $Y$, where a causal direction exists. In bivariate causal discovery, the task is to distinguish between $X \to Y$ and $Y \to X$. Under Assumptions A1–A5, the causal direction is identified at the population level under the PNL model class.

**Assumptions A1–A5** (Assumptions for Identifiability)**.**

*(A1) The data-generating mechanism is correctly specified within the assumed functional model class.*

*(A2) $X$ and $Y$ have an acyclic relationship.*

*(A3) $X$ and $Y$ are unconfounded, meaning they share no common causes.*

*(A4) The data are independent and identically distributed (i.i.d.).*

*(A5) None of the five unidentifiable cases characterized by Zhang and Hyvarinen (2012) (summarized in Appendix B.1) occur; these include, for example, situations where the relationship between $X$ and $Y$ is linear and all random variables are Gaussian.*

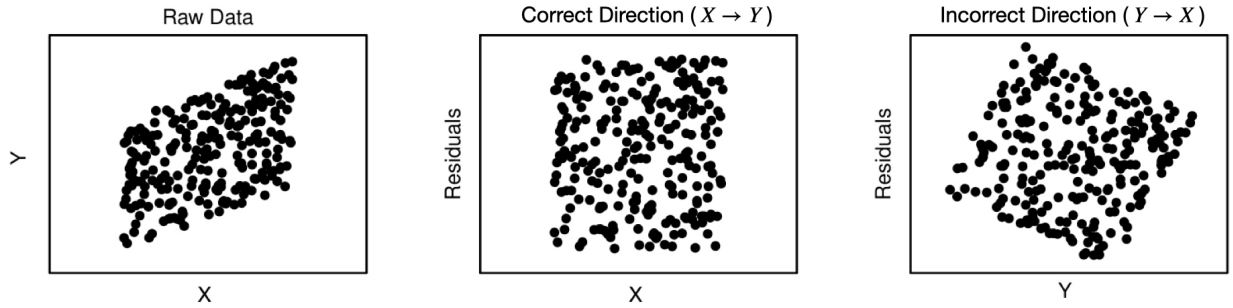



Figure 4.1: Left: Raw data. Middle: Residuals from regressing the outcome on the predictor in the correct causal direction. Right: Residuals from regressing the predictor on the outcome in the incorrect direction. Adapted from Wang et al. (2025).

When Assumptions A1–A5 are satisfied, the noise term in the true causal direction is independent of the predictor but not the other way around. This property can be exploited to infer the causal direction. To make this dependence structure explicit, assume the true data-generating mechanism is given by:

$$Y = f_2(f_1(X) + \eta), \tag{4.1}$$

where $f_1$ and $f_2$ are the generating functions, and $f_2$ is invertible. The noise term $\eta$ is, by construction, independent of the predictor $X$.

Now, suppose $f_1$ is also invertible. Then we can write:

$$X = f_1^{-1}(f_2^{-1}(Y) - \eta). \tag{4.2}$$

Next, consider attempting to represent the data-generating mechanism in the reverse

direction, $Y \to X$. Specifically, we consider representations of the form

$$X = g_2\big(g_1(Y) + \xi\big), \tag{4.3}$$

where $g_1$ and $g_2$ are unknown functions and $g_2$ is assumed to be invertible. Let $\mathcal{G}_1$ and $\mathcal{G}_2$ denote the function classes from which $g_1$ and $g_2$ are chosen (e.g., linear functions under LiNGAM).

To select a particular representation within this class, we define $(g_1^\star, g_2^\star)$ as population risk minimizers under a loss function $\ell$:

$$(g_1^\star, g_2^\star) \in \arg\min_{g_1 \in \mathcal{G}_1,\, g_2 \in \mathcal{G}_2} \mathbb{E}\big[\ell\big(g_2^{-1}(X) - g_1(Y)\big)\big]. \tag{4.4}$$

Given $(g_1^\star, g_2^\star)$, we define the reverse-direction residual as

$$\xi := g_2^{\star -1}(X) - g_1^\star(Y), \tag{4.5}$$

so that (4.3) holds by construction.

The functions $(g_1^\star, g_2^\star)$ represent the best-fitting functions within the chosen model classes when attempting to express the data-generating process in the reverse direction, and depend on both the model class and the loss function $\ell$. They are introduced to illustrate how dependence arises when fitting the incorrect causal direction. According to (4.2), the residual $\xi$ can be written as a deterministic function of $(Y, \eta)$. Since $Y$ itself is generated from $\eta$ via (4.1), we generally have $Y \not\perp\!\!\!\perp \xi$. Thus, if $X \to Y$ is the true data-generating mechanism, then $X \perp\!\!\!\perp \eta$ but typically $Y \not\perp\!\!\!\perp \xi$, except in the well-known unidentifiable cases under Assumption A5 (e.g., the linear-Gaussian setting).

Figure 4.1 illustrates this asymmetry property between the correct and incorrect causal directions. Additional intuition for how $\xi$ behaves under different properties (e.g., additive $f_1$) is provided in Appendix B.2.

As shown above, the PNL model relies on independence between residuals and predictors

to infer the causal direction—a property that holds only when the fitted model is correctly specified. Thus, assessing the causal direction naturally involves evaluating both model adequacy (goodness-of-fit) and independence, recasting causal discovery as a joint problem of goodness-of-fit and independence testing. Building on this observation, the next section introduces the *test-based approach*, a hypothesis-testing framework for statistical inference in functional causal discovery.

### *4.3.2 Test-based Approach*

The test-based approach translates the causal discovery task of selecting between the directions $X \to Y$ and $Y \to X$ into the following pair of hypotheses:

$$\begin{aligned} H^0_{X\to Y}&: X \to Y, \qquad H^1_{X\to Y}: \text{ otherwise}, \\ H^0_{Y\to X}&: Y \to X, \qquad H^1_{Y\to X}: \text{ otherwise}. \end{aligned} \tag{4.6}$$

Under the key assumptions of PNL causal discovery (A1–A5), the noise is independent of the predictor in the true causal direction. Thus, the pair of hypotheses from (4.6) can be reformulated to test both independence and goodness of fit in each direction. Specifically, for prespecified function classes $\mathcal{F}_1, \mathcal{F}_2, \mathcal{G}_1, \mathcal{G}_2$, we consider:

$$\begin{aligned} &H^0_{X\to Y}: \exists\, f_1 \in \mathcal{F}_1,\ f_2 \in \mathcal{F}_2 \text{ s.t. } Y = f_2(f_1(X)+\eta),\ f_2 \text{ invertible},\ X \perp \eta, \quad H^1_{X\to Y}: \textit{otherwise}; \\ &H^0_{Y\to X}: \exists\, g_1 \in \mathcal{G}_1,\ g_2 \in \mathcal{G}_2 \text{ s.t. } X = g_2(g_1(Y)+\xi),\ g_2 \text{ invertible},\ Y \perp \xi, \quad H^1_{Y\to X}: \textit{otherwise}. \end{aligned} \tag{4.7}$$

Here, the function classes $\mathcal{F}_1, \mathcal{F}_2, \mathcal{G}_1, \mathcal{G}_2$ can encode modeling assumptions (e.g. linearity) and are typically taken to be the same across directions (i.e., $\mathcal{F}_1 = \mathcal{G}_1, \mathcal{F}_2 = \mathcal{G}_2$). One can choose among different independence tests to conduct the hypothesis testing in (4.7). For example, we use the test introduced by Sen and Sen (2014), which simultaneously assesses (i) independence between the residuals and the predictor using the Hilbert–Schmidt independence criterion and (ii) the goodness of fit within the specified function classes. When

incorporated into a hypothesis-testing framework, the test-based approach provides a principled way to assess whether a proposed directional model is compatible with the data, in contrast to methods that rely solely on comparing dependence measures (scores). This enables detection of situations where model assumptions are not satisfied, leading to inconclusive outcomes.

Table 4.2: Causal discovery outcomes and their interpretation when using hypothesis tests in (4.7) to test for goodness-of-fit and independence test.

| **Test Outcome** | **Interpretation** |
|---|---|
| Reject both $H^0_{X\to Y}$ and $H^0_{Y\to X}$. | *Inconclusive* outcome; can arise from model misspecfication. |
| Fail to reject both $H^0_{X\to Y}$ and $H^0_{Y\to X}$. | *Inconclusive* outcome; can arise from identifiability issues or small sample size. |
| Reject $H^0_{X\to Y}$ and fail to reject $H^0_{Y\to X}$. | Favors the direction $Y \to X$. |
| Fail to reject $H^0_{X\to Y}$ and reject $H^0_{Y\to X}$. | Favors the direction $X \to Y$. |

Table 4.2 summarizes how different assumption violations correspond to each possible outcome under the test-based approach. Rejecting $H^0_{X\to Y}$ while failing to reject $H^0_{Y\to X}$ favors the direction $X \to Y$, whereas rejecting $H^0_{Y\to X}$ while failing to reject $H^0_{X\to Y}$ favors $Y \to X$. When both null hypotheses are rejected, the outcome is inconclusive, which occurs whenever neither causal direction provides an adequate representation of the data under the assumed model families. A common reason for this is model misspecification (i.e., violation of Assumption A1). While the Post-Nonlinear (PNL) model class is flexible enough to represent a wide range of data-generating mechanisms, practical implementations rely on estimated functions (e.g., $\hat{f}_1, \hat{f}_2$) that may not adequately capture the true relationships. As a result, the test-based procedure can still detect apparent model misspecification when sample size is limited or when the functional form lacks sufficient flexibility to approximate the true data-generating process. Moreover, for additive-noise models (ANMs) and LiNGAM, which are special cases of PNL, both null hypotheses may be rejected if the true data-generating process lies outside those restricted model classes (e.g., when the errors are not additive or the model is not linear, respectively). Conversely, when neither null hypothesis is rejected, the outcome is also inconclusive, reflecting situations in which the available data do not

provide sufficient statistical power to distinguish between directions under the assumed model families. Inconclusive outcomes can arise from small sample sizes or identifiability issues. In the LiNGAM setting, for instance, this occurs under Gaussianity, where the model becomes unidentifiable (Shimizu & Kano, 2008).

In addition to inferential guarantees provided by hypothesis testing, we recommend incorporating a resampling-based procedure that estimates the frequency of each causal discovery outcome (Table 4.2) and constructs associated confidence intervals for the test-based approach. For example, if one were to resample with replacement $S$ times, the rate of favoring the direction $X \to Y$ would be

$$\frac{1}{S}\sum_{s=1}^{S}\mathbf{I}_s(\text{Reject } H^0_{X\to Y} \text{ and fail to reject } H^0_{Y\to X} \text{ for replication } s). \tag{4.8}$$

This frequency can be viewed as an estimate of the probability of each outcome, since it is the sample mean of the corresponding indicator function. Conditional on the observed dataset $D_N$, the bootstrap replicates are i.i.d., so by the conditional law of large numbers and central limit theorem, (4.8) is consistent and asymptotically normal, allowing for confidence intervals to be constructed using the normal approximation (Casella & Berger, 2021).

This resampling step quantifies uncertainty in the causal direction by assessing the stability of each outcome across bootstrap replicates. Conceptually, this is a bootstrap procedure, much like those used in the LiNGAM literature (e.g., Komatsu et al., 2010b; Thamvitayakul et al., 2012a), in that we repeatedly draw bootstrap datasets and recompute the causal conclusion. In the bivariate setting, both approaches estimate how often each causal outcome would be recovered under repeated sampling. However, the test-based framework generalizes this idea by resampling the data and re-running both hypothesis tests on each bootstrap dataset. Each replicate therefore yields a pair of hypothesis test outcomes (e.g., Reject $H^0_{X\to Y}$ and fail to reject $H^0_{Y\to X}$ for replication $s$) and the resulting bootstrap distribution estimates the joint distribution of these two hypothesis-test outcomes. In contrast, LiNGAM bootstrap procedures resample the data, refit the LiNGAM algorithm on each

bootstrap sample, and record only the resulting causal ordering; the bootstrap distribution therefore reflects how often the algorithm selects $X \to Y$ versus $Y \to X$ across replicates. Moreover, LiNGAM bootstrap methods assess the stability of estimated causal coefficients or graph structures within a specific model class, whereas our approach applies the same resampling principle directly to the joint hypothesis-test outcomes.

As summarized in Table 4.1, the test-based approach occupies a favorable position among existing functional causal discovery methods by providing a unified inferential framework that integrates two complementary forms of inference within a hypothesis-testing paradigm. The discussion below elaborates on how these advantages arise.

The test-based approach extends and improves upon using LiNGAM, ANM, and PNL causal discovery algorithms in several important ways. Rather than relying solely on the comparison of dependence measures or test statistics, the test-based approach obtains two $p$-values corresponding to hypothesis tests in each direction. The results of these hypothesis tests convey uncertainty in the direction estimate and provide insights into assumption violations. Although two tests are performed, they correspond to different modeling assumptions—one under $X \to Y$ and another under $Y \to X$—and their outcomes are interpreted jointly. Because these hypotheses are logically coupled, standard multiple-comparison adjustments are neither required nor appropriate; the inferential target is the pattern of rejections rather than the marginal significance of each test.

When assumptions hold, the two $p$-values are perfectly correlated, meaning that, with enough samples, rejecting in one direction would imply a failure to reject in the other. Therefore, given no assumption violations and a sufficiently large sample size, the test-based approach would provide the same information as a traditional functional causal discovery method.

However, one cannot be sure that strict assumptions always hold in practice. Under assumption violations, the two $p$-values would not be perfectly correlated, and the test-based approach would be able to convey this information by either rejecting or failing to reject in both directions, resulting in an inconclusive directionality (see Table 4.2). In contrast, many

implementations of functional causal discovery methods output a single causal direction (e.g., $X \to Y$) without providing the option for an inconclusive outcome (e.g., Ikeuchi et al., 2023; Kalainathan et al., 2020; Kalisch et al., 2012).

As also reflected in Table 4.1, previous work (e.g., Hoyer et al., 2009; Mooij et al., 2016a; Peters et al., 2014; Shimizu et al., 2011; Zhang and Hyvarinen, 2012) has not formalized an inferential framework that integrates hypothesis testing with additional inferential quantities to provide statistical guarantees in functional causal discovery. Unlike the bootstrap-based approaches of Komatsu et al. (2010b) and Thamvitayakul et al. (2012a) that assess stability under the assumption that the LiNGAM causal model is correctly specified, our framework explicitly incorporates information about model assumptions through formal hypothesis tests. Moreover, unlike the confidence-set approach proposed by Wang et al. (2025), our framework can be paired with the broader Post-Nonlinear (PNL) model class and offers multiple forms of uncertainty quantification—both through hypothesis testing itself and through the estimated rates of causal discovery outcomes. We now turn to numerical results to demonstrate the test-based approach in practice.

## *4.4 Numerical Results*

In this section, we illustrate the use and behavior of the test-based approach through both simulated and real data. In the simulations studies in Section 4.4.1, we pair the test-based framework with the LiNGAM and PNL model classes and assess its performance in controlled settings with varying degrees of assumption violations.

For real data analyses in Section 4.4.2, we pair the approach with LiNGAM to demonstrate its application in real settings where one might ordinarily apply LiNGAM-based causal discovery and where the extent of assumption violations is unknown. In this section, we limit our illustrations to the LiNGAM model class. This is because allowing for more extensive PNL or ANM model classes would require additional machine-learning steps to estimate the functions $f_1$ (and $f_2$ in the PNL case). Such model-selection decisions—choosing architectures, tuning hyperparameters, and optimizing nonlinear predictors—are outside the

scope of this chapter and would distract from our main emphasis on illustrating inferential guarantees provided by the test-based approach.

### *4.4.1 Simulations*

In our simulations, we pair the test-based approach to two model classes, LiNGAM and PNL. For both classes, the simulations illustrate how the hypothesis-testing formulation in (4.7) can be implemented in practice within a given model class and how the resampling step yields empirical causal outcome rates that quantify uncertainty. However, the scope and emphasis of the simulation studies differ across the two model classes.

Specifically, for the LiNGAM model class, we use simulations to demonstrate how the test-based approach paired with LiNGAM behaves under both model misspecification and unidentified settings. In this case, we further compare the resulting decisions and resampling-based outcome rates to those obtained from DirectLiNGAM with bootstrap rates, as is commonly used in applications (see Table 4.1). This comparison highlights differences in the inferential information provided by the two approaches, particularly in settings where model assumptions fail.

For the PNL model class, we focus exclusively on the behavior of the test-based approach under varying degrees of model misspecification. We do not consider unidentified PNL settings, as the linear–Gaussian case already provides the most intuitive and canonical example of non-identifiability in functional causal models, which is examined in the LiNGAM simulations. As a result, the PNL simulations are used to assess how the test-based approach responds to misspecification within a flexible nonlinear model class. In addition, because existing PNL-based methods do not provide a comparable bootstrap-based inferential procedure, we do not include a direct comparison to alternative PNL methods with inference in this setting.

#### *4.4.1.1 LiNGAM Model Class Simulations*

When paired with LiNGAM, the hypothesis tests in the test-based approach can be formalized as follows:

$$\begin{aligned} H^0_{X\to Y} &: X \perp \eta, \text{ relationship between } X \text{ and } Y \text{ is linear}, \quad H^1_{X\to Y} : \text{otherwise}, \\ H^0_{Y\to X} &: Y \perp \xi, \text{ relationship between } Y \text{ and } X \text{ is linear}, \quad H^1_{Y\to X} : \text{otherwise}. \end{aligned} \tag{4.9}$$

We view this as a special case of (4.7) in which $f_1$ and $g_1$ are restricted to be linear and $f_2$ and $g_2$ are identity maps, yielding linear structural equations with additive noise.

To study the behavior of the test-based approach paired with LiNGAM and DirectLiNGAM, we consider two forms of assumption violations: (i) varying degrees of linear model misspecification (violation of Assumption A1) and (ii) data-generating processes that span a range of non-Gaussianity settings, including the linear–Gaussian setting, which is unidentified under Assumption A5.

For each simulation setting, we generate $M = 100$ independent datasets, each of size $N = 3000$, from the same data-generating process under the true causal direction $X \to Y$. For each of the $M$ simulated datasets, we run the test-based approach to jointly test for independence and goodness-of-fit of the linear model, as formalized in (4.9), using the combined independence and goodness-of-fit test based on the Hilbert–Schmidt Independence Criterion (HSIC) proposed by Sen and Sen (2014). We compare the results with those from DirectLiNGAM,[1] ensuring comparability by also using HSIC as the independence measure within DirectLiNGAM. For each of the $M$ simulated datasets, we additionally compute the causal outcome rates (4.8) for the test-based approach and the bootstrap rates for DirectLiNGAM following the procedure of Thamvitayakul et al. (2012a).

[1]We use DirectLiNGAM rather than ICA-LiNGAM because DirectLiNGAM, under its assumptions, is guaranteed to converge to the correct causal ordering in a finite number of iterations as the sample size tends to infinity (Shimizu et al., 2011). This guarantee does not hold for ICA-LiNGAM. In the bivariate case, DirectLiNGAM fits linear models in both directions and determines the causal direction by assessing the independence between the predictor and its residuals. For each direction, the algorithm computes a user-specified measure of independence and selects the direction that minimizes this measure.

Table 4.3: Simulation settings for varying linearity.

| | | |
|---|---|---|
| Linear | Polynomial = 1 | $Y = \text{sign}(X-a)\lvert X-a\rvert\beta + \eta$ |
| Moderately nonlinear | Polynomial = 1.5 | $Y = \text{sign}(X-a)\lvert X-a\rvert^{1.5}\beta + \eta$ |
| Severely Nonlinear | Polynomial = 3 | $Y = \text{sign}(X-a)\lvert X-a\rvert^{3}\beta + \eta$ |

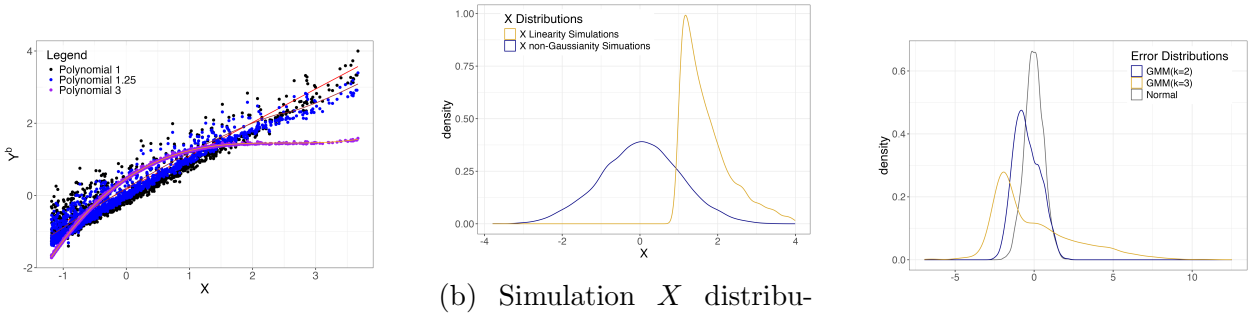
Table 4.4: Simulation settings for varying levels of Gaussianity. Here $k$ is the number of mixture components in the Gaussian mixture model (GMM).

| | |
|---|---|
| Gaussian | $X \sim N(0,1), \eta \sim N(\mu_1, \sigma_1)$ |
| Slightly non-Gaussian | $X \sim N(0,1), \eta \sim \text{GMM}(k=2)$ |
| Non-Gaussian | $X \sim N(0,1), \eta \sim \text{GMM}(k=3)$ |

Table 4.3 summarizes the functional form used to generate $Y$ from $X$ under the three levels of linearity assumption violations considered: no violations (where $Y$ is a linear function of $X$), moderate violations (where $Y$ is a polynomial function of $X$ with a polynomial power of 1.5), and severe violations (where $Y$ is a polynomial function of $X$ with a polynomial power of 3). We ensure all other assumptions, besides linearity, are met. We sample $X$ from a truncated exponential distribution (Figure 4.2b), which ensures that $Y$ is not normally distributed. Figure 4.2a illustrates the simulated data under each linearity setting.

Table 4.4 summarizes the noise distributions used to generate $\eta$, corresponding to three levels of increasing Gaussianity, or equivalently, decreasing degrees of violation of the non-Gaussianity assumption, considered in our simulations: (1) non-Gaussian errors generated from a three-component Gaussian mixture model (GMM), (2) slight violations from a two-component GMM, and (3) Gaussian errors. To ensure a controlled progression toward Gaussianity, mixture parameters were chosen so that the three-component mixture produces a visibly multimodal, asymmetric distribution, while the two-component mixture produces a unimodal but still heavy-tailed and slightly skewed distribution. Specifically, for the three-component case, mixture means $(-2, 0, 2)$ and unequal variances $(0.5, 1, 3)$ with mixture weights $(0.4, 0.2, 0.4)$ yield a clearly non-Gaussian, multimodal shape. For the two-component case, mixture means $(-1, 0.25)$ and variances $(0.5, 0.6)$ with weights $(0.6, 0.4)$

generate a near-unimodal density with heavier tails and mild skewness relative to a standard Gaussian. Finally, under the Gaussian setting, the noise, $\eta$, is drawn from a single standard normal distribution. We ensure that all other assumptions, aside from non-Gaussianity, are satisfied by sampling $X$ from a Gaussian distribution (Figure 4.2b). When both $X$ and the noise $\eta$ are Gaussian, the resulting $Y$ is also Gaussian, placing the model in the linear–Gaussian setting where causal direction is not identified. Figure 4.2c illustrates the corresponding error distributions under each non-Gaussianity setting.




(a) Settings of linearity.

(b) Simulation $X$ distribution.

(c) Error ($\eta$) distributions.

Figure 4.2: Simulation settings for linearity and Gaussianity.

Tables 4.5 and 4.6 summarize the results for the test-based approach and DirectLiNGAM under each level of linearity and non-Gaussianity assumption violation, respectively. For the test-based approach, we report decision rates across the $M$ replications, along with the average outcome rates aggregated over the $M$ simulated datasets. For DirectLiNGAM, we likewise report decision rates across the $M$ replications and the average bootstrap resampling rates aggregated over the $M$ simulated datasets.

Simulation results under varying degrees of linearity violation show that the test-based approach appropriately signals assumption violations, frequently yielding inconclusive outcomes, whereas DirectLiNGAM tends to select an incorrect causal direction when linearity assumptions are violated. When there are no assumption violations, both DirectLiNGAM and the test-based approach correctly identify the causal direction $X \to Y$ across all repli-

| | Test-based approach | | DirectLiNGAM | |
|---|---|---|---|---|
| Degree of misspecification | Decision rate | Average outcome rates | Decision rate | Average bootstrap rates |
| Linear (polynomial 1) | **$X \to Y$: 100%**<br>$Y \to X$: 0%<br>Inconclusive: 0% | Reject both: 3.1%<br>Fail to reject both: 0%<br>Reject only $X \to Y$: 0%<br>Reject only $Y \to X$: 96.9% | **$X \to Y$: 100%**<br>$Y \to X$: 0% | $X \to Y$: 99.6%<br>$Y \to X$: 0.4% |
| Moderately nonlinear (polynomial 1.5) | **$X \to Y$: 0%**<br>$Y \to X$: 0%<br>Inconclusive: 100% | Reject both: 100%<br>Fail to reject both: 0%<br>Reject only $X \to Y$: 0%<br>Reject only $Y \to X$: 0% | **$X \to Y$: 0%**<br>$Y \to X$: 100% | $X \to Y$: 0%<br>$Y \to X$: 100% |
| Nonlinear (polynomial 3) | **$X \to Y$: 0%**<br>$Y \to X$: 0%<br>Inconclusive: 100% | Reject both: 100%<br>Fail to reject both: 0%<br>Reject only $X \to Y$: 0%<br>Reject only $Y \to X$: 0% | **$X \to Y$: 0%**<br>$Y \to X$: 100% | $X \to Y$: 0%<br>$Y \to X$: 100% |

Table 4.5: Comparison between the test-based approach and DirectLiNGAM under varying degrees of linearity assumption violation. Rows correspond to no violation (linear), moderate violation (polynomial exponent 1.5), and severe violation (polynomial exponent 3). For the test-based approach, we report the decision and average outcome rates across $M$ simulated datasets. For DirectLiNGAM, we report the decision and average bootstrap rates.

cations. For DirectLiNGAM, the average bootstrap rate for the correct direction is 99.6%. For the test-based approach, the average outcome rate corresponding to the correct direction is quite high (96.9%), with all other average causal outcome rates being quite small. Under moderate and severe model misspecification, DirectLiNGAM incorrectly favors the reverse direction $Y \to X$, with average bootstrap rates of 100% in both settings. In contrast, across all replications, the test-based approach yields an inconclusive result, as both hypothesis tests are rejected—indicating assumption violations due to model misspecification. The average causal outcome rates further support this conclusion, with the average rate of rejecting both directions equal to 100% and all other average rates equal to 0%. Together, the moderate and severe model misspecification settings illustrate how the inferential framework provided by the test-based approach enables a more informative interpretation of uncertainty. Specifically, the test-based approach correctly signals that no causal direction can be reliably inferred under assumption violations and therefore yields an inconclusive result. By contrast, while DirectLiNGAM incorporates a bootstrap-based inferential procedure, it

fails to indicate the presence of assumption violations and instead favors an incorrect causal direction due to its enforced deterministic decision rule.

| | Test-based approach | | DirectLiNGAM | |
|---|---|---|---|---|
| Degree of Gaussianity | Decision rate | Average outcome rates | Decision rate | Average bootstrap rates |
| $GMM(k = 3)$ | **$X \to Y$: 100%**<br>$Y \to X$: 0%<br>Inconclusive: 0% | Reject both: 2.7%<br>Fail to reject both: 0%<br>Reject only $X \to Y$: 0%<br>Reject only $Y \to X$: 97.3% | **$X \to Y$: 100%**<br>$Y \to X$: 0% | $X \to Y$: 99.1%<br>$Y \to X$: 0.9% |
| $GMM(k = 2)$ | **$X \to Y$: 70%**<br>$Y \to X$: 0%<br>Inconclusive: 30% | Reject both: 5.2%<br>Fail to reject both: 24.9%<br>Reject only $X \to Y$: 1.1%<br>Reject only $Y \to X$: 68.8% | **$X \to Y$: 96%**<br>$Y \to X$: 4% | $X \to Y$: 95.2%<br>$Y \to X$: 4.8% |
| Gaussian | **$X \to Y$: 0%**<br>$Y \to X$: 0%<br>Inconclusive: 100% | Reject both: 1.2%<br>Fail to reject both: 97.6%<br>Reject only $X \to Y$: 0.4%<br>Reject only $Y \to X$: 0.8% | **$X \to Y$: 45%**<br>$Y \to X$: 55% | $X \to Y$: 42.9%<br>$Y \to X$: 57.1% |

Table 4.6: Comparison between the test-based approach and DirectLiNGAM under varying degrees of non-Gaussianity assumption violation. Rows correspond to no violation (errors from a $GMM(k = 3)$), slight violation (errors from a $GMM(k = 2)$), and strong violation (Gaussian data and errors i.e. direction is unidentified). For the test-based approach, we report the decision and average outcome rates across $M$ simulated datasets. For DirectLiNGAM, we likewise report the decision and average bootstrap-based stability rates across $M$ simulated datasets.

Similarly to the (non-)linearity settings, simulation results under varying degrees of non-Gaussianity assumption violation also show that the test-based approach appropriately signals when the causal direction is unidentified, yielding inconclusive outcomes, whereas DirectLiNGAM tends to oscillate between the two directions when the non-Gaussianity assumption is violated. When there are no or only slight violations of the non-Gaussianity assumption and all other assumptions are satisfied, both DirectLiNGAM and the test-based approach correctly identify the causal direction $X \to Y$ for the majority of replications (100% for both under no violations, and 96% for DirectLiNGAM and 70% for the test-based approach under slight violations). For DirectLiNGAM, the average bootstrap rates for the correct direction are 99.1% and 95.2% under no and slight violations, respectively. Under no assumption violations, the test-based approach yields an average outcome rate of 97.3% for

the correct direction, with all other average causal outcome rates being quite small. Under slight violations, the average rate of correctly identifying the causal direction decreases to 70%, while the average rate of failing to reject both hypotheses increases to 24.9%, indicating mild departures from the non-Gaussianity assumption. Under severe assumption violations, when the data are Gaussian, DirectLiNGAM favors the incorrect direction $Y \to X$ in 55% of replications, with an associated average bootstrap rate of 57.1%. In contrast, across all replications, the test-based approach yields an inconclusive result, as both hypothesis tests fail to be rejected—indicating assumption violations due to Gaussianity. The average causal outcome rates further support this interpretation, with the rate of failing to reject both hypotheses equal to 97.6% and all other average rates comparatively small. Together, these settings illustrate how the inferential framework provided by the test-based approach enables a more informative interpretation of uncertainty when identifiability breaks down. Specifically, the test-based approach correctly signals that no causal direction can be reliably inferred under violations of the non-Gaussianity assumption and therefore yields an inconclusive result. By contrast, while the bootstrap in DirectLiNGAM provides a limited form of uncertainty quantification—the reduced average rate of 57.1% under severe violations—it still favors the incorrect direction in the majority of replications. Because DirectLiNGAM enforces a deterministic decision and does not allow for inconclusive outcomes, this behavior may still lead a practitioner to favor the incorrect direction.

#### *4.4.1.2 PNL Model Class Simulations*

When paired with PNLs, we use the hypothesis tests in (4.7), with $f_1$ and $f_2$ restricted to a specified nonlinear family.

We use our simulations to study the behavior of the test-based approach paired with PNLs under varying degrees of model misspecification (violation of Assumption A1). We refrain from considering unidentifiable regimes (Assumption A5), since the linear–Gaussian case—the simplest unidentifiable regime—has already been studied for the LiNGAM model class, where we show that the test-based approach correctly signals non-identifiability. Ex-

tending these experiments to PNL models would not provide additional insight.

For each model misspecification setting, we generate $M = 100$ independent datasets, each of size $N = 3000$, from the same data-generating process under the true causal direction $X \to Y$, given by

$$Y = f_2\big(f_1(X) + \eta\big),$$

which corresponds to a post-nonlinear (PNL) causal model. In our simulations, we specify logarithmic $f_1(x) = \log(x)$ and hyperbolic tangent $f_2(u) = \tanh(u)$ transformations. We use the same marginal distribution for $X$ and the same error distributions $\eta$ as in the linearity simulations for LiNGAM model class (see Figures 4.2b and 4.2c).

To assess the effect of model misspecification, we consider three pairs of fitted models for $f_1$ and $f_2$, corresponding to correct specification, partial misspecification, and full misspecification. First, in the correctly specified PNL scenario, the fitted models match the true forms of $f_1$ and $f_2$. Second, in a partially misspecified scenario, we fit a linear model for $f_1$ while correctly specifying $f_2$. Third, in a fully misspecified scenario, we fit linear models for both $f_1$ and $f_2$. These scenarios allow us to isolate the impact of progressively stronger deviations from the true PNL model on the behavior of the test-based approach.

We find that, across these misspecification settings, the test-based approach behaves analogously to the LiNGAM linear misspecification experiments. When the PNL model is correctly specified, the test-based approach recovers the true causal direction $X \to Y$ in 100% of the $M$ replications. Under partial and full misspecification, it yields inconclusive outcomes in 100% of the $M$ replications, indicating that no reliable causal conclusion can be drawn under these misspecified models. Thus, the PNL simulations corroborate the conclusions from the LiNGAM misspecification experiments: when the model is correctly specified, the test-based approach recovers the correct direction, whereas under misspecification it appropriately signals a lack of reliable evidence through inconclusive outcomes. A table summarizing these results is provided in Appendix B.3.

In summary, our simulation studies confirm that the test-based approach provides a

principled and informative inferential framework for functional causal discovery. Across model classes, the simulations show that the test-based approach correctly identifies the causal direction when model assumptions are satisfied. Through hypothesis testing and causal outcome rates, the framework quantifies uncertainty in the inferred direction and yields inconclusive outcomes when the causal direction is unidentified—corresponding to failure to reject both directions—or when model misspecification is present, as indicated by rejection of both directions.

By contrast, the commonly used DirectLiNGAM with bootstrap provides limited inferential information. While DirectLiNGAM with bootstrap summarizes the stability of a single directional decision across resamples, it does not permit inconclusive outcomes or provide insight about potential assumption violations. Under moderate and severe linear model misspecification, DirectLiNGAM favors the incorrect direction, and in the linear–Gaussian setting the causal direction is unidentified, leading DirectLiNGAM to oscillate between the two directions at near 50% rates. The test-based approach, on the other hand, permits inconclusive outcomes and provides insight into potential assumption violations. In the LiNGAM model class simulations, it is able to signal potential violations of two key assumptions—linearity (Assumption A1) and non-Gaussianity (Assumption A5). The PNL model class simulations further demonstrate that this behavior is not specific to LiNGAM: when PNL assumptions hold, the test-based approach favors the correct causal direction, and when model misspecification is present, it shifts toward inconclusive outcomes rather than enforcing a potentially misleading decision.

Our simulation results motivate the analysis of real-world datasets, where the data-generating mechanism and the validity of modeling assumptions are typically unknown.

### *4.4.2 Real Data*

We apply the test-based approach to three real-world datasets from the Cause–Effect Pairs Benchmark (Tübingen pairs) (Mooij et al., 2016a): two Balltrack datasets and the Ozone and Temperature dataset. These examples were chosen because their scatterplots suggested

approximately linear relationships. Therefore, one might reasonably expect approaches based on the LiNGAM model class to be well suited for these data, making them informative test cases for illustration.

Through this analysis, we demonstrate how the test-based approach paired with LiNGAM provides a richer inferential framework than DirectLiNGAM. Specifically, separate hypothesis tests for each causal direction allow practitioners to distinguish between confident directional conclusions and inconclusive cases due to possible assumption violations—insights that are not available from DirectLiNGAM, even when supplemented with bootstrap rates (Thamvitayakul et al., 2012a).

The first Balltrack dataset contains 94 measurements collected using a physical ball track equipped with two pairs of light barrier sensors: the first sensor ($X$) records the ball's initial speed, and the second ($Y$) records its speed at a later position along the track. The experiment was designed and conducted for the Tübingen Cause–Effect Pairs benchmark suite (Mooij et al., 2016a). The initial portion of the track has a steep slope, making the initial speed highly sensitive to the exact release position. To introduce natural variability and avoid systematic positioning, some runs used positions chosen by an adult, while the remainder used positions chosen by a four-year-old child. Following Mooij et al. (2016a), we treat the causal direction as known, with the initial speed causing the final speed. A scatterplot of initial versus final speed (Figure 4.3) suggests an approximately linear relationship. Nevertheless, the DirectLiNGAM point estimate infers the incorrect causal direction when applied to this dataset, and the associated bootstrap procedure supports this incorrect direction in 61% of resamples. In contrast, the test-based approach supports the correct causal direction. Specifically, the causal outcome rate for the correct direction is 62%, while the remaining outcome rates indicate slight assumption violations, potentially due to model misspecification, as evidenced by a 29% rate of rejecting both directions. This example illustrates that relying on the DirectLiNGAM point estimate alone is not advisable and that, even when supplemented with bootstrap rates, it can still be misleading. By contrast, the test-based approach not only recovers the correct causal direction but also provides diagnostic information about

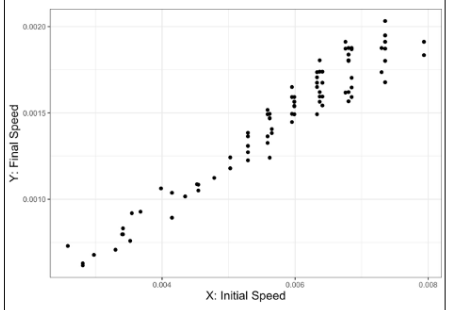
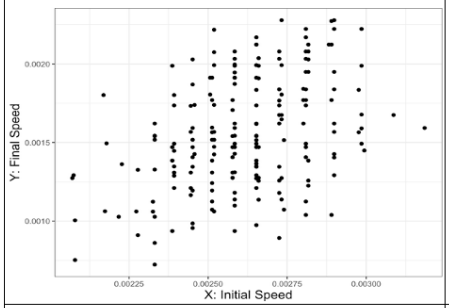
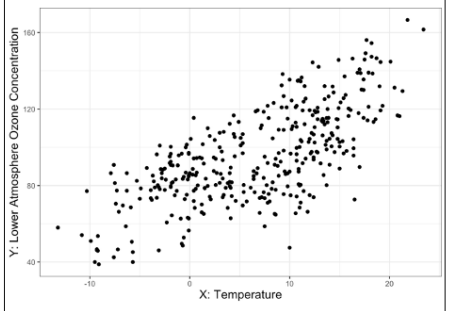
| Dataset | Test-based Approach | | LiNGAM | |
|---|---|---|---|---|
| | Decision | Outcome rates (95% CI) | Decision | Bootstrap rates |
|   | $X \to Y$ (correct) | Reject both: 29% [20%, 38%]<br>Fail to reject both: 9% [3%, 15%]<br>Reject only $X \to Y$: 0%<br>Reject only $Y \to X$: 62% [52%, 72%] | $Y \to X$ (incorrect) | $X \to Y$: 39%<br>$Y \to X$: 61% |
|   | Inconclusive (fail to reject both) | Reject both: 18% [11%, 26%]<br>Fail to reject both: 50% [40%, 60%]<br>Reject only $X \to Y$: 22% [14%, 30%]<br>Reject only $Y \to X$: 10% [4%, 16%] | $Y \to X$ (incorrect) | $X \to Y$: 39%<br>$Y \to X$: 61% |
|   | Inconclusive (reject both) | Reject both: 100%<br>Fail to reject both: 0%<br>Reject only $X \to Y$: 0%<br>Reject only $Y \to X$: 0% | $Y \to X$ (incorrect) | $X \to Y$: 48%<br>$Y \to X$: 52% |

Figure 4.3: The first two rows presents results for the Balltrack datasets, and the third row presents results for the Temperature and Ozone dataset. From left to right, the columns display the scatterplot of the data, the test-based decision and outcome rates (resampling proportions with 95% confidence intervals; CIs are omitted when the estimated variance is zero), and the corresponding LiNGAM decision with associated bootstrap rates.

potential assumption violations.

The second Balltrack dataset contains 202 measurements collected under a setup similar to the first Balltrack dataset with two sensors $X$ and $Y$ recording the initial and a later speed along the track. However, in this setup, the track had a shorter acceleration zone, which limited the extent to which the ball can increase its speed. As with the first Balltrack dataset, the experiment was designed and conducted for the Tübingen Cause–Effect Pairs benchmark suite (Mooij et al., 2016a), and we treat the causal direction as known, with the initial speed causing the final speed. A scatterplot of initial versus final speed (Figure 4.3) suggests an approximately linear relationship. Here, the DirectLiNGAM point estimate infers the incorrect causal direction when applied to this dataset, and the associated bootstrap procedure supports this incorrect direction in 61% of resamples. By contrast, the test-based approach yields an inconclusive result by failing to reject either causal direction, with a causal outcome rate of 50%. This indicates potential identifiability or small sample size issues, an informative finding which is preferable to LiNGAM's enforced deterministic (and incorrect) decision.

The Ozone and Temperature dataset contains 365 daily mean temperatures and lower-atmosphere ozone concentrations measured in Chaumont, Switzerland, during 2009. The data were collected by the Swiss Federal Office for the Environment (FOEN) Federal Office for the Environment (FOEN), 2009. Following Mooij et al. Mooij et al., 2016a, we assume the causal direction is known, with changes in temperature causing changes in lower-atmosphere ozone concentrations. When plotting ozone concentrations against temperature in Figure 4.3, the relationship between the two variables appears approximately linear. As shown in Figure 4.3, applying DirectLiNGAM to this dataset yields the incorrect causal direction, with a bootstrap rate of 52%. In contrast, the test-based approach yields an inconclusive result by rejecting both causal directions, indicating potential model misspecification—consistent with the behavior observed in our simulation studies (Table 4.5). This outcome is preferable to LiNGAM's enforced deterministic (and incorrect) decision, as it not only avoids a misleading conclusion but also provides diagnostic information about possible model misspecification.

Overall, the real-data analyses further illustrate how the test-based approach provides a richer inferential framework than DirectLiNGAM with bootstrap. In particular, it produces directional conclusions when supported by the data and inconclusive outcomes otherwise, together with outcome rates and confidence intervals that quantify uncertainty across resamples. This distinction—between enforcing a single deterministic direction and allowing uncertainty-aware, inconclusive conclusions—is particularly important in applied settings, where the data-generating mechanisms and the presence or extent of assumption violations are typically unknown. In such scenarios, inconclusive outcomes are especially informative: rejecting both causal directions can signal potential model misspecification, while failing to reject either direction may indicate lack of identifiability or insufficient sample size under the assumed model class. By permitting and interpreting these distinct inconclusive outcomes, the test-based approach provides diagnostic insight that is not available from LiNGAM-based causal discovery methods.

## *4.5 Discussion*

Causal discovery methods have pioneered data-driven approaches for identifying and understanding causal relationships, a goal that is fundamental across many scientific disciplines. However, applied researchers using causal discovery methods should be mindful of the (lack of) statistical guarantees in existing approaches such as LiNGAM, ANM, or PNL.

In this chapter, we reviewed the state of statistical guarantees provided by existing functional causal discovery methods. As summarized in Table 4.1, this review revealed a clear gap: the absence of a formal, general inferential framework in functional causal discovery. As the first step toward filling this gap, we then developed the test-based approach to functional causal discovery that establishes a formal inferential framework by combining two complementary forms of inference: hypothesis testing and uncertainty quantification for each competing causal discovery outcome. Together, these components provide a principled way to quantify uncertainty in the estimated causal direction and provide diagnostic insight for cases when a determinate causal direction is not supported by the data.

Specifically, the inferential framework provided by the test-based approach enables practitioners to (i) report evidence supporting a causal direction when such conclusion is supported by the data, (ii) quantify uncertainty in that conclusion, and (iii) obtain diagnostic insight into potential assumption violations—such as model misspecification (Assumption A1) and non-identifiability (Assumption A5)—particularly when the causal direction is inconclusive. Importantly, these forms of inference are available across major functional causal model classes. As a result, the test-based approach to functional causal discovery addresses a gap in the existing literature, where this combination of qualities—uncertainty quantification, diagnostic insight, and applicability across model classes—is largely unavailable (see Table 4.1).

The lesson for applied statisticians is threefold. First, causal discovery should always be undertaken with explicit awareness of its modeling assumptions and of the consequences when those assumptions fail. Second, inconclusive results or signs of assumption violation should not be viewed as failures but as statistically meaningful outcomes that can guide model refinement, data collection, or reformulation of causal hypotheses. Third, causal discovery should be paired with explicit uncertainty quantification, and users should understand the nature and scope of the inferential information provided by these tools. The test-based approach represents a first step toward incorporating these key principles into a coherent inferential framework.

Within the test-based approach, we primarily discussed the impact of model misspecification and non-identifiability (e.g., the linear–Gaussian setting). An important direction for future work is to extend this framework to characterize how other types of assumption violations—such as unobserved confounding, cyclicity, and non-i.i.d. data—affect the behavior of the two hypothesis tests in (4.6) and the resulting causal discovery outcomes. Understanding how these violations, individually or in combination, influence test outcomes would further enhance the diagnostic scope and practical utility of the test-based approach.

While we studied the test-based approach in the bivariate setting in this chapter, extending the framework to multivariate settings is another important direction for future work.

One very simple multivariate setting in which the framework can directly apply is when the causal structure is known for all variables except for two variables of interest, $X$ and $Y$, and the effects of all other variables on $X$ and $Y$ are understood. In this case, the framework can be used to study the directionality between $X$ and $Y$ with only minor modifications. Specifically, one could first remove the effects of the known confounders on $X$ and $Y$, respectively, via, e.g., linear regression. The resulting residuals $r_X$ and $r_Y$ can then be treated as a new pair of bivariate variables to be directly used in the test-based approach. Extending the test-based approach to a general multivariate setting in which the causal structure among variables is unknown is a substantially more complex problem. Such an extension would require evaluating a large collection of hypothesis tests corresponding to the many possible parent sets for each variable. The number of these tests—and the resulting space of possible causal outcomes—grows combinatorially with the number of variables, making coherent inference and uncertainty summarization challenging. Moreover, the introduction of confounders substantially complicates the estimation effort, as it necessitates modeling and removing nuisance components for multiple variables. Developing scalable procedures and principled methods for summarizing uncertainty in a multivariate outcome space remains an important avenue for future research.

Finally, the inferential foundations established by the test-based functional causal discovery approach can be extended using ideas from the broader statistical inference literature—such as methods for power analysis (Casella & Berger, 2021; Lehmann & Romano, 2005), resampling (Efron & Tibshirani, 1993; Good, 2005), and model diagnostics (Atkinson, 1985; Buja et al., 2009; Cook, 1977)—to further strengthen the inferential foundations of causal discovery and promote its reliable use in practice.

# Chapter 5

# CAUSAL DISCOVERY VIA STATISTICAL POWER (CDSP)

This work was conducted in collaboration with Dr. Elena A. Erosheva and Dr. Fan Xia.

## *5.1 Introduction*

In this chapter, we focus on bivariate functional causal discovery that reduces to determining the causal direction between two variables by fitting regressions in both directions. The bivariate setting is important because it is the simplest setting in which causal discovery identifiability mechanisms can be studied. It can also be considered a building block for more complex multivariate settings. Bivariate causal discovery methods are often evaluated on benchmark datasets consisting of cause–effect pairs (Mooij et al., 2016b).

In a bivariate setting, functional causal models posit that the data-generating process for two random variables $X$ and $Y$ and the true causal direction $X \to Y$ can be written as

$$Y = f(X, \eta), \tag{5.1}$$

where $\eta$ is a noise term that is independent of $X$ and $f \in \mathcal{F}$ is a function belonging to a specified model class.

Functional causal discovery methods rely on strong identifiability assumptions to identify the causal direction from observational data (Hoyer et al., 2009; Shimizu et al., 2006).

**Assumptions A1–A5** (Assumptions for Identifiability)**.** *In the bivariate setting, the identifiability assumptions are as follows:*

*(A1) The data-generating mechanism is correctly specified within the assumed functional model class.*

*(A2)* *$X$ and $Y$ have an acyclic relationship.*

*(A3)* *$X$ and $Y$ are unconfounded, i.e., they share no common causes.*

*(A4)* *The data are independent and identically distributed (i.i.d.).*

*(A5)* *The data-generating process does not fall into the few unidentifiable cases characterized by Zhang and Hyvarinen (2012); for example, when the relationship between $X$ and $Y$ is linear and all variables are Gaussian.*

Under Assumptions A1–A5, the independence between cause variable $X$ and noise variable $\eta$ holds only in the true direction and fails in the reverse direction. This asymmetry enables identification of a unique causal direction from observational data. When Assumptions A1–A5 hold, the causal direction can be identified; however, impacts of assumption violations on traditional causal discovery approaches are unclear. In addition, most existing causal discovery methods do not provide statistical guarantees—that is, statistical inference for the estimated direction.

In practice, assumption violations such as model misspecification due to deviations from the assumed functional form are often unavoidable. This has been widely recognized; for example, in linear regression, the consequences of assumption violations have been extensively studied, motivating the use of diagnostic tools and uncertainty quantification to assess model fit and guide interpretation (Kutner et al., 2005; Weisberg, 2005).

In contrast, current work on uncertainty quantification in functional causal discovery has been limited. It can be broadly divided into two categories: resampling-based procedures that assess stability of estimated causal structures and inferential approaches based on statistical tests. Bootstrap-based methods extend causal discovery algorithms by repeatedly refitting models on resampled datasets to assess stability of inferred causal orderings (Komatsu et al., 2010a; Thamvitayakul et al., 2012b); however, as we will see in this chapter, resulting reliability measures can be misleading when the underlying modeling assumptions are violated. Wang et al. (2025) develop statistical inference via confidence sets for causal

orderings by inverting goodness-of-fit tests in identifiable structural equation models with additive errors. The confidence sets offer insights into model-class misspecification: an empty confidence set indicates that the "model class does not capture the data-generating process" (Wang et al., 2025, p. 2). While this approach provides a formal method for uncertainty quantification for both bivariate and multivariate functional causal discovery, it does not extend to more general functional causal models such as the post-nonlinear model. Additionally, both resampling-based and inferential methods are developed under the assumption that the working models are correctly specified and may fail to provide reliable directional evidence under mild violations, as robustness to such assumption violations is not explicitly examined.

More broadly, much of the existing causal discovery literature emphasizes identifiability under idealized modeling assumptions, with comparatively limited attention paid to uncertainty quantification and robustness under assumption violations. A central challenge in applying causal discovery in practice is that identifiability assumptions—such as those related to model specification (Assumption A1)—are rarely satisfied exactly, making it essential to develop methods that remain informative under such violations.

To address this challenge, we develop *Causal Discovery via Statistical Power (CDSP)*, a framework that connects causal direction estimation with statistical power concepts. CDSP provides a statistical perspective on causal discovery by introducing notions of statistical power and effect size tailored to this setting. Within this framework, we introduce the *effect-size asymmetry assumption*, which posits that the incorrect direction deviates more strongly from its null and is therefore easier to reject than the correct direction. We show that this asymmetry implies that the probability of detecting the correct causal direction exceeds that of detecting the incorrect direction, yielding a statistical characterization of directional evidence. We further leverage the effect-size asymmetry assumption to formulate causal direction estimation as a problem of assessing directional detectability. We develop the CDSP procedure to estimate the causal direction via effect-size asymmetry and to provide statistical inference for this estimate. Through simulation studies, we demonstrate that

CDSP is robust to model misspecification. Analyzing 100 cause–effect benchmark pairs (Mooij et al., 2016b), we find that CDSP yields materially improved performance, reducing the false discovery rate by approximately 18% relative to LiNGAM, both across all pairs and when restricted to pairs with approximately linear relationships—a setting in which methods based on the linear model class are expected to perform well.

Next, we introduce the CDSP framework and its theoretical foundations in Section 5.2, followed by numerical results in Section 5.3. We conclude with a discussion of our contributions and open questions for future research in Section 5.4.

## *5.2 Causal Discovery via Statistical Power (CDSP)*

In this section, we introduce Causal Discovery via Statistical Power (CDSP), a framework that connects causal direction estimation with statistical power. We begin by formulating a hypothesis-testing framework for bivariate causal discovery and defining a corresponding notion of directional power. We then show that this notion of power depends on the relative magnitudes of directional effect sizes, leading to the effect-size asymmetry assumption. Using asymptotic arguments, we establish that the effect-size asymmetry assumption is equivalent to favoring the correct direction with higher probability. Importantly, the effect-size asymmetry assumption does not explicitly depend on the presence of model misspecification. We then present the CDSP procedure, the implementation of the proposed framework for direction estimation, and discuss inference for the estimated direction through a measure of directional strength.

### *5.2.1 Motivating the Role of Power, Effect Size, and Assumptions in Causal Discovery*

**Assumption 5.2.1** (Standing Assumption)**.** *Assume that the true causal direction is $X \to Y$, and the data-generating mechanism is*

$$Y = m(X) + \eta,$$

*where $m$ is an arbitrary (unknown) generating function. The noise $\eta$ is, by construction, independent of $X$.*

*To assess the reverse direction $Y \to X$, we attempt to represent $X$ as a function of $Y$ by fitting a model of the form*

$$X = l(Y) + \xi,$$

*where $l$ is restricted to lie in a specified model class (e.g., linear functions under LiNGAM). The function $l$ is defined as the population risk minimizer,*

$$l = \arg\min_{h \in \mathcal{L}} R(h),$$

*where $R(h)$ denotes the population risk. The reverse-direction residual is defined as*

$$\xi := X - l(Y).$$

*Thus $(l, \xi)$ represents the best-fitting approximation to the true data-generating process within the chosen model class when expressing the data in the reverse direction.*

Functional causal discovery methods assume that the regression function lies within the specified model class. Under this assumption, independence between the predictor and the noise term in the correct causal direction, but generally not in the incorrect one, creates an asymmetry that can be exploited for causal discovery. However, this independence cannot be tested directly because the noise term is unobserved. Instead, the noise terms, denoted by $\hat{\eta}$ and $\hat{\xi}$, must be estimated from regression models fit in the two possible directions. Measures of dependence between $\hat{\eta}$ and $X$, and between $\hat{\xi}$ and $Y$, are then compared to determine the causal direction. However, simply comparing the estimated dependence measures in the two directions does not account for the uncertainty in these empirical quantities. We therefore formulate causal direction detection as a pair of hypothesis tests corresponding to the two possible causal directions in (5.2),

$$
\begin{aligned}
H^0_{X\to Y} : X \perp \eta,\ m \in \mathcal{M}, \quad H^1_{X\to Y} : \text{otherwise},\\
H^0_{Y\to X} : Y \perp \xi,\ l \in \mathcal{L}, \quad H^1_{Y\to X} : \text{otherwise}.
\end{aligned}
\tag{5.2}
$$

The null hypotheses $H^0_{X\to Y}$ and $H^0_{Y\to X}$ jointly assess (i) independence between the predictor and residual and (ii) the adequacy of the model classes $\mathcal{M}$ and $\mathcal{L}$ for the $X \to Y$ and $Y \to X$ directions, respectively. Their alternatives, $H^1_{X\to Y}$ and $H^1_{Y\to X}$, correspond to any violation of these conditions.

Here, $\mathcal{M}$ denotes the model class used to represent the generating mechanism $m$ in the $X \to Y$ direction, while $\mathcal{L}$ denotes the model class used to approximate the reverse-direction regression function $l$. In practice, these model classes are often taken to be the same (e.g., both linear), but they need not coincide in general.

Under the assumptions of functional causal discovery methods (Assumptions A1–A5), the model class corresponding to the true causal direction is correctly specified[1], the two tests are assessing only (i), and therefore the test results can directly inform causal direction. When either or both $\mathcal{M}$ or $\mathcal{L}$ are misspecified, rejection of a null hypothesis may arise from residual dependence, model misspecification, or both. This motivates decomposing each composite alternative into subcases that distinguish these possibilities in (5.3).

$$
\begin{aligned}
H^{1a}_{X\to Y} : X \not\perp \eta,\ m \in \mathcal{M}, \qquad H^{1a}_{Y\to X} : Y \not\perp \xi,\ l \in \mathcal{L},\\
H^{1b}_{X\to Y} : X \perp \eta,\ m \notin \mathcal{M}, \qquad H^{1b}_{Y\to X} : Y \perp \xi,\ l \notin \mathcal{L},\\
H^{1c}_{X\to Y} : X \not\perp \eta,\ m \notin \mathcal{M}, \qquad H^{1c}_{Y\to X} : Y \not\perp \xi,\ l \notin \mathcal{L}.
\end{aligned}
\tag{5.3}
$$

Intuitively, alternatives that depart more strongly from the null should be easier to detect and therefore yield higher power, whereas alternatives that represent weaker or more limited departures should generally yield lower power. In our setting, when model misspecification (Assumption A1) is the sole source of violation, the correct causal direction satisfies the independence condition and can violate the null only through the model misspecification.

[1]Model misspecification occurs when the true data-generating relationship does not follow the assumed functional form and therefore cannot be represented within the specified model class.

Consequently, if the model class in the correct direction is only mildly misspecified, the corresponding alternative may remain close to the null and thus be harder to detect. This motivates studying power under the derived alternatives, as different forms of null violation can have different detectability and therefore provide evidence for causal direction.

In classical hypothesis testing (Casella & Berger, 2021; Cohen, 1988; Lehmann & Romano, 2005), the *statistical power* of a test is defined as

$$\mathbb{P}(\text{reject } H_0 \mid H_1 \text{ is true}),$$

for a given null hypothesis $H_0$ and a given alternative hypothesis $H_1$.

Because causal discovery involves comparing two directional hypotheses, the notion of statistical power must be adapted to account for this paired testing structure. In particular, power corresponds to the probability that the procedure favors the correct direction.

**Definition 5.2.1** (Directional Power)**.** *Directional power is the probability that the causal discovery method correctly favors $X \to Y$, namely*

$$\mathbb{P}\Big( \big(\textit{fail to reject } H^0_{X\to Y}\big) \wedge \big(\textit{reject } H^0_{Y\to X}\big) \,\Big|\, X \to Y \Big). \tag{5.4}$$

To understand how this probability depends on the strength of directional evidence, we next introduce a directional effect size. Let $T_{X\to Y}$ and $T_{Y\to X}$ denote the test statistics corresponding to the tests of $H^0_{X\to Y}$ and $H^0_{Y\to X}$, respectively.

**Definition 5.2.2** (Directional Effect Size)**.** *Let $\theta_{X\to Y}$ and $\theta_{Y\to X}$ denote the population values of the test statistics $T_{X\to Y}$ and $T_{Y\to X}$, respectively. Let $\sigma_{X\to Y}/\sqrt{n}$ and $\sigma_{Y\to X}/\sqrt{n}$ denote their corresponding standard deviations.*

*Define the directional effect size for testing $H^0_{X\to Y} : \theta_{X\to Y} = \theta_{0,X\to Y}$ as*

$$\Delta_{X\to Y,n} = \frac{(\theta_{X\to Y} - \theta_{0,X\to Y})\sqrt{n}}{\sigma_{X\to Y}}.$$

*Similarly, the directional effect size for testing* $H^0_{Y\to X} : \theta_{Y\to X} = \theta_{0,Y\to X}$ *is*

$$\Delta_{Y\to X,n} = \frac{(\theta_{Y\to X} - \theta_{0,Y\to X})\sqrt{n}}{\sigma_{Y\to X}}.$$

*In our setting, we assume that under the null hypotheses of independence, the population values of the test statistics are zero (i.e.,* $\theta_{0,X\to Y} = \theta_{0,Y\to X} = 0$*). This assumption holds for commonly used dependence measures such as the Hilbert–Schmidt Independence Criterion (HSIC) (Gretton et al., 2008). Under this assumption, the directional effect sizes simplify to*

$$\Delta_{X\to Y,n} = \frac{\theta_{X\to Y}\sqrt{n}}{\sigma_{X\to Y}}, \qquad \Delta_{Y\to X,n} = \frac{\theta_{Y\to X}\sqrt{n}}{\sigma_{Y\to X}}. \tag{5.5}$$

**Assumption 5.2.2** (Traditional Asymmetry Assumption)**.** *Traditionally, functional causal discovery methods rely on an asymmetry between the two directions, which can be expressed in terms of the population test statistics as*

$$\theta_{Y\to X} \;>\; \theta_{X\to Y}.$$

*Under the standard identifiability assumptions (Assumptions A1–A5), the correct direction* $X \to Y$ *yields* $\theta_{X\to Y} = 0$*, while the reverse direction* $Y \to X$ *yields* $\theta_{Y\to X} > 0$*. Thus, this inequality arises as a direct consequence of the model assumptions when they are satisfied.*

Although this Assumption 5.2.2 is not typically stated explicitly in the literature, it is implicitly relied upon. For example, many functional causal discovery methods compare test statistics or dependence measures across directions and select the direction with the smaller value (e.g., Mooij et al. (2016b), Peters et al. (2017), and Shimizu et al. (2006)).

Building on the traditional asymmetry assumption (Assumption 5.2.2), we incorporate the variability of the test statistics through standardization and account for the corresponding critical values, so that the comparison reflects how strongly each direction departs from its null, yielding a refined comparison in Assumption 5.2.3.

**Assumption 5.2.3** (Effect-Size Asymmetry)**.** *Let $q_{X\to Y,n}$ and $q_{Y\to X,n}$ denote the directional standardized critical values,*

$$q^{\alpha}_{X\to Y,n} = \frac{\sqrt{n}\, c^{\alpha}_{X\to Y,n}}{\sigma_{X\to Y}}, \qquad q^{\alpha}_{Y\to X,n} = \frac{\sqrt{n}\, c^{\alpha}_{Y\to X,n}}{\sigma_{Y\to X}}, \tag{5.6}$$

*where $c^{\alpha}_{X\to Y,n}$ and $c^{\alpha}_{Y\to X,n}$ denote the level-$\alpha$ critical values under the null hypotheses $H^0_{X\to Y}$ and $H^0_{Y\to X}$, respectively. We say that the effect-size asymmetry assumption holds if*

$$\frac{\theta_{Y\to X}}{\sigma_{Y\to X}} > \frac{\theta_{X\to Y}}{\sigma_{X\to Y}} \quad \textit{and} \quad q^{\alpha}_{Y\to X,n} - q^{\alpha}_{X\to Y,n} = o(1) \tag{5.7}$$

**Remark 1.** *The effect-size asymmetry assumption depends on the relative magnitudes of the directional effect sizes and thus does not explicitly depend on model misspecification.*

When model misspecification (Assumption A1) is the sole source of violation, the alternatives in Equation (5.2) simplify to those in Equation (5.8).

$$\begin{aligned} &H^{1b}_{X\to Y}: X \perp \eta,\ m \notin \mathcal{M} \ \text{ and } \ H^{1c}_{Y\to X}: Y \not\perp \xi,\ l \notin \mathcal{L}, \text{if } X \to Y \text{ is true,} \\ &H^{1c}_{X\to Y}: X \not\perp \eta,\ m \notin \mathcal{M} \ \text{ and } \ H^{1b}_{Y\to X}: Y \perp \xi,\ l \notin \mathcal{L}, \text{if } Y \to X \text{ is true.} \end{aligned} \tag{5.8}$$

In this regime, the incorrect direction ($Y \to X$) exhibits two sources of deviation from its null—both $Y \not\perp \xi$ and $l \notin \mathcal{L}$—whereas the correct direction ($X \to Y$) exhibits only one, namely $m \notin \mathcal{M}$. The effect-size asymmetry assumption posits that two such deviations induce a larger departure from the null than a single deviation, and hence a larger departure relative to the corresponding rejection threshold. This disparity yields asymmetric finite-sample behavior: the incorrect direction is more readily detectable and is therefore rejected more quickly than the correct direction.

To make the connection between directional effect sizes and rejection probabilities concrete, we work under the misspecification regime in Equation (5.8) where the true causal direction is $X \to Y$, the correct-direction null is violated through model misspecification only (i.e., $H^{1b}_{X\to Y}$), and the incorrect-direction null is violated through both dependence and

model misspecification (i.e., $H^{1c}_{Y\to X}$).

In what follows, we use the independence and goodness-of-fit test of Sen and Sen (2014), which jointly assesses (i) independence between the predictor and residuals using the Hilbert–Schmidt independence criterion (HSIC) (Gretton et al., 2008) and (ii) whether the regression function lies within the assumed functional model class. We focus on linear model misspecification. Theorem C.1.1 in the Appendix shows that the corresponding test statistics $(T_{1b,X\to Y}, T_{1c,Y\to X})$ are asymptotically bivariate normal under the alternatives.[2]

We show in Theorem 5.2.1 that, under this regime, the effect-size asymmetry assumption is equivalent to the probability of correctly detecting the causal direction (i.e., directional power) exceeding the probability of incorrectly favoring the reverse direction.

**Theorem 5.2.1** (Effect-Size Asymmetry Characterization of Directional Power)**.** *Consider the misspecification regime in which the correct-direction null is violated through model misspecification only (i.e., $H^{1b}_{X\to Y}$ holds) while the incorrect-direction null is violated through both dependence and model misspecification (i.e., $H^{1c}_{Y\to X}$ holds).*

*Let $T_{1b,X\to Y}$ and $T_{1c,Y\to X}$ be the test statistics corresponding to $H^{1b}_{X\to Y}$ and $H^{1c}_{Y\to X}$. Let $c^{\alpha}_{X\to Y}$ and $c^{\alpha}_{Y\to X}$ denote the level-$\alpha$ critical values for $nT_{1b,X\to Y}$ and $nT_{1c,Y\to X}$, respectively (Sen & Sen, 2014), and define $c^{\alpha}_{X\to Y,n} := \frac{c^{\alpha}_{X\to Y}}{n}$ and $c^{\alpha}_{Y\to X,n} := \frac{c^{\alpha}_{Y\to X}}{n}$. Define the directional detectability indices*

$$I_{X\to Y,n} := \frac{\theta_{X\to Y} - c^{\alpha}_{X\to Y,n}}{\sigma_{X\to Y}}, \qquad I_{Y\to X,n} := \frac{\theta_{Y\to X} - c^{\alpha}_{Y\to X,n}}{\sigma_{Y\to X}}. \tag{5.9}$$

*Equivalently, $\sqrt{n}\, I_{X\to Y,n} = \Delta_{X\to Y,n} - q^{\alpha}_{X\to Y,n}$, $\sqrt{n}\, I_{Y\to X,n} = \Delta_{Y\to X,n} - q^{\alpha}_{Y\to X,n}$. Denote by $\eta$ the noise term in the true direction and by $\xi$ the noise term in the reverse direction, as defined in Assumption 5.2.1. Define the misspecified residuals in the true and incorrect directions as*

$$\varepsilon = m(X) - X\tilde{\beta}_0 + \eta, \qquad \delta = l(Y) - Y\tilde{\gamma}_0 + \xi,$$

[2] We present the results in Theorem 5.2.1 under an asymptotic Gaussian approximation for clarity, although the same arguments apply more generally to any pair of directional test statistics with a known joint distribution, whether asymptotic or finite-sample.

*where $\tilde{\beta}_0$ and $\tilde{\gamma}_0$ are the population linear regression coefficients in the correct and incorrect directions, respectively. Let*

$$\theta_{X\to Y} = \theta(X, \varepsilon), \qquad \theta_{Y\to X} = \theta(Y, \delta),$$

*denote the population HSIC dependence measures corresponding to $(X, \varepsilon)$ and $(Y, \delta)$, respectively, and let*

$$\sigma^2_{X\to Y} := \sigma^2_{1b,X\to Y}, \qquad \sigma^2_{Y\to X} := \sigma^2_{1c,Y\to X},$$

*denote the asymptotic variances of $T_{1b,X\to Y}$ and $T_{1c,Y\to X}$. Then*

$$\sqrt{n}(T_{1b,X\to Y} - \theta_{X\to Y}) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{X\to Y}),$$
$$\sqrt{n}(T_{1c,Y\to X} - \theta_{Y\to X}) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{Y\to X}).$$

*As a result, under the asymptotic normal approximations,*

$$\mathbb{P}\Big((\textit{reject } H^0_{Y\to X}) \wedge (\textit{fail to reject } H^0_{X\to Y}) \,\Big|\, H^{1b}_{X\to Y}, H^{1c}_{Y\to X}\Big)$$
$$> \mathbb{P}\Big((\textit{reject } H^0_{X\to Y}) \wedge (\textit{fail to reject } H^0_{Y\to X}) \,\Big|\, H^{1b}_{X\to Y}, H^{1c}_{Y\to X}\Big)$$

*if and only if*

$$I_{Y\to X,n} > I_{X\to Y,n}, \tag{5.10}$$

*or equivalently,*

$$\Delta_{Y\to X,n} - q^\alpha_{Y\to X,n} > \Delta_{X\to Y,n} - q^\alpha_{X\to Y,n}. \tag{5.11}$$

*Furthermore, since $q^\alpha_{Y\to X,n} - q^\alpha_{X\to Y,n} = o(1)$, then asymptotically Equation 5.10 is equivalent to*

$$\frac{\theta_{Y\to X}}{\sigma_{Y\to X}} > \frac{\theta_{X\to Y}}{\sigma_{X\to Y}}.$$

*Thus, directional effect sizes $\Delta_{Y\to X,n}$ and $\Delta_{X\to Y,n}$ govern detectability of the true causal direction.*

*Proof.* Under the misspecification regime $(H^{1b}_{X\to Y}, H^{1c}_{Y\to X})$, decompose the marginal rejection probabilities as

$$\begin{aligned}\mathbb{P}(\text{reject } H^0_{Y\to X} \mid H^{1b}_{X\to Y}, H^{1c}_{Y\to X}) &= \mathbb{P}\big((\text{reject } H^0_{Y\to X}) \wedge (\text{fail to reject } H^0_{X\to Y}) \,\big|\, H^{1b}_{X\to Y}, H^{1c}_{Y\to X}\big) \\ &\quad + \mathbb{P}\big((\text{reject } H^0_{Y\to X}) \wedge (\text{reject } H^0_{X\to Y}) \,\big|\, H^{1b}_{X\to Y}, H^{1c}_{Y\to X}\big),\end{aligned}$$

$$\begin{aligned}\mathbb{P}(\text{reject } H^0_{X\to Y} \mid H^{1b}_{X\to Y}, H^{1c}_{Y\to X}) &= \mathbb{P}\big((\text{reject } H^0_{X\to Y}) \wedge (\text{fail to reject } H^0_{Y\to X}) \,\big|\, H^{1b}_{X\to Y}, H^{1c}_{Y\to X}\big) \\ &\quad + \mathbb{P}\big((\text{reject } H^0_{Y\to X}) \wedge (\text{reject } H^0_{X\to Y}) \,\big|\, H^{1b}_{X\to Y}, H^{1c}_{Y\to X}\big).\end{aligned}$$

Subtracting the two identities gives

$$\begin{aligned}&\mathbb{P}\big((\text{reject } H^0_{Y\to X}) \wedge (\text{fail to reject } H^0_{X\to Y}) \,\big|\, H^{1b}_{X\to Y}, H^{1c}_{Y\to X}\big) \\ &\quad - \mathbb{P}\big((\text{reject } H^0_{X\to Y}) \wedge (\text{fail to reject } H^0_{Y\to X}) \,\big|\, H^{1b}_{X\to Y}, H^{1c}_{Y\to X}\big) \\ &= \mathbb{P}(\text{reject } H^0_{Y\to X} \mid H^{1b}_{X\to Y}, H^{1c}_{Y\to X}) - \mathbb{P}(\text{reject } H^0_{X\to Y} \mid H^{1b}_{X\to Y}, H^{1c}_{Y\to X}). \qquad (5.12)\end{aligned}$$

Because the rejection event for $H^0_{Y\to X}$ depends only on the test statistic $T_{1c,Y\to X}$ and the rejection event for $H^0_{X\to Y}$ depends only on the test statistic $T_{1b,X\to Y}$, the corresponding marginal rejection probabilities are determined by the marginal asymptotic laws of these statistics. Hence

$$\mathbb{P}(\text{reject } H^0_{Y\to X} \mid H^{1b}_{X\to Y}, H^{1c}_{Y\to X}) = \mathbb{P}(\text{reject } H^0_{Y\to X} \mid H^{1c}_{Y\to X}),$$

and

$$\mathbb{P}(\text{reject } H^0_{X\to Y} \mid H^{1b}_{X\to Y}, H^{1c}_{Y\to X}) = \mathbb{P}(\text{reject } H^0_{X\to Y} \mid H^{1b}_{X\to Y}).$$

Therefore, Equation (5.12) becomes

$$\begin{aligned}
&\mathbb{P}\Big((\text{reject } H^0_{Y\to X}) \wedge (\text{fail to reject } H^0_{X\to Y}) \,\Big|\, H^{1b}_{X\to Y}, H^{1c}_{Y\to X}\Big) \\
&\quad - \mathbb{P}\Big((\text{reject } H^0_{X\to Y}) \wedge (\text{fail to reject } H^0_{Y\to X}) \,\Big|\, H^{1b}_{X\to Y}, H^{1c}_{Y\to X}\Big) \\
&= \mathbb{P}(\text{reject } H^0_{Y\to X} \mid H^{1c}_{Y\to X}) - \mathbb{P}(\text{reject } H^0_{X\to Y} \mid H^{1b}_{X\to Y}). \qquad (5.13)
\end{aligned}$$

Under the asymptotic normal approximation we have,

$$\begin{aligned}
\mathbb{P}(\text{reject } H^0_{Y\to X} \mid H^{1c}_{Y\to X}) &= \mathbb{P}(nT_{1c,Y\to X} > c^\alpha_{Y\to X} \mid H^{1c}_{Y\to X}) \\
&= \mathbb{P}\left(\frac{\sqrt{n}(T_{1c,Y\to X} - \theta_{Y\to X})}{\sigma_{Y\to X}} > \frac{\sqrt{n}(c^\alpha_{Y\to X}/n - \theta_{Y\to X})}{\sigma_{Y\to X}} \,\middle|\, H^{1c}_{Y\to X}\right) \\
&\approx 1 - \Phi\left(\frac{\sqrt{n}(c^\alpha_{Y\to X}/n - \theta_{Y\to X})}{\sigma_{Y\to X}}\right) \\
&= \Phi\left(\frac{\sqrt{n}(\theta_{Y\to X} - c^\alpha_{Y\to X}/n)}{\sigma_{Y\to X}}\right).
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathbb{P}(\text{reject } H^0_{X\to Y} \mid H^{1b}_{X\to Y}) &= \mathbb{P}(nT_{1b,X\to Y} > c^\alpha_{X\to Y} \mid H^{1b}_{X\to Y}) \\
&= \mathbb{P}\left(\frac{\sqrt{n}(T_{1b,X\to Y} - \theta_{X\to Y})}{\sigma_{X\to Y}} > \frac{\sqrt{n}\,(c^\alpha_{X\to Y}/n - \theta_{X\to Y})}{\sigma_{X\to Y}} \,\middle|\, H^{1b}_{X\to Y}\right) \\
&\approx 1 - \Phi\left(\frac{\sqrt{n}(c^\alpha_{X\to Y}/n - \theta_{X\to Y})}{\sigma_{X\to Y}}\right) \\
&= \Phi\left(\frac{\sqrt{n}(\theta_{X\to Y} - c^\alpha_{X\to Y}/n)}{\sigma_{X\to Y}}\right).
\end{aligned}$$

Substituting these expressions into Equation (5.13) yields

$$\begin{aligned}
&\mathbb{P}\Big((\text{reject } H^0_{Y\to X}) \wedge (\text{fail to reject } H^0_{X\to Y}) \,\Big|\, H^{1b}_{X\to Y}, H^{1c}_{Y\to X}\Big) \\
&\quad - \mathbb{P}\Big((\text{reject } H^0_{X\to Y}) \wedge (\text{fail to reject } H^0_{Y\to X}) \,\Big|\, H^{1b}_{X\to Y}, H^{1c}_{Y\to X}\Big) \\
&= \Phi\left(\frac{\sqrt{n}(\theta_{Y\to X} - c^\alpha_{Y\to X}/n)}{\sigma_{Y\to X}}\right) - \Phi\left(\frac{\sqrt{n}(\theta_{X\to Y} - c^\alpha_{X\to Y}/n)}{\sigma_{X\to Y}}\right).
\end{aligned}$$

Since $\Phi$ is strictly increasing, the right-hand side is positive if and only if

$$\begin{aligned} \frac{\sqrt{n}\,(\theta_{Y\to X} - c^{\alpha}_{Y\to X}/n)}{\sigma_{Y\to X}} &> \frac{\sqrt{n}\,(\theta_{X\to Y} - c^{\alpha}_{X\to Y}/n)}{\sigma_{X\to Y}} \\ \iff \quad \Delta_{Y\to X,n} - q^{\alpha}_{Y\to X,n} &> \Delta_{X\to Y,n} - q^{\alpha}_{X\to Y,n} \\ \iff \quad I_{Y\to X,n} &> I_{X\to Y,n}. \end{aligned} \tag{5.14}$$

Since $q^{\alpha}_{Y\to X,n} - q^{\alpha}_{X\to Y,n} = o(1)$, then asymptotically Equation 5.14 is equivalent to

$$\frac{\theta_{Y\to X}}{\sigma_{Y\to X}} > \frac{\theta_{X\to Y}}{\sigma_{X\to Y}}.$$

This proves the claim. □

**Remark 2.** *The directional detectability indices $I_{Y\to X,n}$ and $I_{X\to Y,n}$, defined in Equation 5.9, provide a finite-sample operational condition for evaluating effect-size asymmetry in terms of standardized effect sizes and critical values, and are what we use in practice for CDSP. As shown in Theorem 5.2.1, for sufficiently large $n$, the condition in Equation 5.14 is equivalent to the effect-size asymmetry assumption in Equation 5.7. The latter provides a simpler, population-level characterization that aids intuition and interpretation, while the former captures the finite-sample quantities that govern the procedure.*

**Remark 3.** *CDSP is most informative under mild to moderate model misspecification, where the effect-size asymmetry is most likely to hold, as the test statistic corresponding to the correct direction tends to remain closer to the null than the test statistic corresponding to the incorrect direction. These settings correspond to regimes of greatest practical relevance, as severe model misspecification is often readily detectable using standard diagnostic tools (e.g., visual inspection), whereas mild to moderate misspecification is more subtle yet can still impact causal discovery.*

### *5.2.2 Estimating Causal Direction and Directional Support Probability via CDSP*

In Section 5.2.1, we defined notions of statistical power and effect size tailored to causal discovery and introduced the effect-size asymmetry assumption, which posits that the standardized deviation from the null is larger in the incorrect direction than in the correct direction. We showed that this asymmetry is equivalent to the probability of detecting the correct causal direction exceeding the probability of detecting the incorrect direction. In this section, we leverage effect-size asymmetry to estimate the causal direction and quantify the strength of directional evidence.

Suppose we have a bivariate dataset of size $N$ on variables $X$ and $Y$. We assume that the true data-generating process is unknown (i.e., we do not know whether $X \to Y$ or $Y \to X$, nor the functional form of $f$ in Equation 5.1). We obtain CDSP point estimate of the causal direction and an estimate of the directional support probability using Algorithm 3. Specifically, CDSP Procedure outputs (i) estimated causal direction $D_{\text{CDSP}}$: either $X \to Y$, $Y \to X$, or inconclusive, and (ii) directional support measure $P_{\text{CDSP}}$.

We estimate causal direction $D_{\text{CDSP}}$ in Algorithm 3 via effect-size asymmetry. By Theorem 5.2.1, effect-size asymmetry is equivalent to the probability of detecting the correct causal direction exceeding the probability of favoring the incorrect direction. We first estimate the required components needed to evaluate effect-size asymmetry via the directional detectibility indices—$\theta_{X\to Y}$ and $\theta_{Y\to X}$ via bootstrap means, $\sigma_{X\to Y}$ and $\sigma_{Y\to X}$ via bootstrap standard deviations, and $c^{\alpha}_{X\to Y,n}$ and $c^{\alpha}_{Y\to X,n}$ via the procedure of Sen and Sen (2014)—and then substitute these estimates into Equation (5.9). We then compare $\hat{I}_{Y\to X,n}$ and $\hat{I}_{X\to Y,n}$ to determine $\hat{D}_{\text{CDSP}}$: if $\hat{I}_{Y\to X,n} > \hat{I}_{X\to Y,n}$, we set $\hat{D}_{\text{CDSP}} = X \to Y$; if $\hat{I}_{X\to Y,n} > \hat{I}_{Y\to X,n}$, we set $\hat{D}_{\text{CDSP}} = Y \to X$.[3]

The computational complexity of computing the point estimate in Algorithm 3 is $O(BN^2)$. CDSP uses $B$ bootstrap replicates to estimate the directional detectibility indices, and each replicate requires a constant number of HSIC computations. Under the standard

[3]In the event that $\hat{I}_{Y\to X,n} = \hat{I}_{X\to Y,n}$, $\hat{D}_{\text{CDSP}}$ is inconclusive. This case is unlikely to occur in practice and is thus omitted from the procedure.

**Algorithm 3** CDSP Procedure: Estimation and Inference via the Effect-Size Asymmetry

1: **Input:** Dataset $\{(X_i, Y_i)\}_{i=1}^{N}$, significance level $\alpha$, number of bootstrap replicates $B$.

2: Using bootstrap, estimate $\hat{\theta}_{X\to Y}$ and $\hat{\theta}_{Y\to X}$ as bootstrap means, $\hat{\sigma}_{X\to Y}$ and $\hat{\sigma}_{Y\to X}$ as bootstrap standard deviations, and $\hat{c}^{\alpha}_{X\to Y,n}$ and $\hat{c}^{\alpha}_{Y\to X,n}$ via the procedure of Sen and Sen (2014).

3: Compute directional detectability index estimates

$$\hat{I}_{X\to Y,n} = \frac{\hat{\theta}_{X\to Y} - \hat{c}^{\alpha}_{X\to Y,n}}{\hat{\sigma}_{X\to Y}}, \qquad \hat{I}_{Y\to X,n} = \frac{\hat{\theta}_{Y\to X} - \hat{c}^{\alpha}_{Y\to X,n}}{\hat{\sigma}_{Y\to X}}.$$

4: **Point estimate of direction:**

$$\hat{D}_{\mathrm{CDSP}} = \begin{cases} X \to Y, & \text{if } \hat{I}_{Y\to X,n} > \hat{I}_{X\to Y,n}, \\ Y \to X, & \text{if } \hat{I}_{X\to Y,n} > \hat{I}_{Y\to X,n}. \end{cases}$$

5: **Bootstrap inference:** For $b = 1, \ldots, B$, draw a bootstrap resample and recompute $\hat{I}^{(b)}_{Y\to X,n}$ and $\hat{I}^{(b)}_{X\to Y,n}$ using Steps 2 and 3.

6: Estimate directional support probability:

$$\widehat{P}_{\mathrm{CDSP}} = \begin{cases} \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\{\hat{I}^{(b)}_{Y\to X,n} - \hat{I}^{(b)}_{X\to Y,n} > 0\}, & \text{if } \hat{D}_{\mathrm{CDSP}} = X \to Y, \\ \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\{\hat{I}^{(b)}_{X\to Y,n} - \hat{I}^{(b)}_{Y\to X,n} > 0\}, & \text{if } \hat{D}_{\mathrm{CDSP}} = Y \to X. \end{cases}$$

7: **Output:** $\hat{D}_{\mathrm{CDSP}}$ and $\widehat{P}_{\mathrm{CDSP}}$.

empirical implementation, HSIC requires $O(N^2)$ time due to pairwise kernel evaluations, yielding an overall time complexity of $O(BN^2)$. The space complexity is $O(N^2 + B)$, where $O(N^2)$ arises from storing the HSIC kernel matrices and $O(B)$ from storing the bootstrap outputs.

In addition to a point estimate of the causal direction, the CDSP procedure in Algorithm 3 outputs an inferential measure quantifying directional support. Specifically, if CDSP estimates $X \to Y$, it reports the bootstrap estimate

$$\widehat{P}_{\text{CDSP}} = \widehat{\mathbb{P}}\big(\hat{I}_{Y\to X,n} - \hat{I}_{X\to Y,n} > 0\big) = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\Big\{\hat{I}^{(b)}_{Y\to X,n} - \hat{I}^{(b)}_{X\to Y,n} > 0\Big\}, \tag{5.15}$$

where $B$ denotes the number of bootstrap samples, and $\hat{I}^{(b)}_{Y\to X,n}$ and $\hat{I}^{(b)}_{X\to Y,n}$ are the corresponding estimates computed from the $b$-th resample. An analogous quantity is reported when CDSP estimates $Y \to X$. The computational complexity of bootstrap inference in Algorithm 3 is $O(B^2N^2)$, since it computes the point-estimation procedure across $B$ bootstrap replicates, each of which has time complexity $O(BN^2)$. The space complexity remains $O(N^2 + B)$.

The population quantity $P_{\text{CDSP}}$ is the probability that the CDSP point estimate $\hat{D}_{\text{CDSP}}$ favors $X \to Y$ under sampling variability from the data-generating distribution. Thus, $\widehat{P}_{\text{CDSP}}$ estimates the probability of directional support for $\hat{D}_{\text{CDSP}}$. From a decision-theoretic perspective, when the direction favored by CDSP is the true causal direction, $\widehat{P}_{\text{CDSP}}$ can be interpreted as the success probability of the procedure under 0–1 loss—that is, the probability that the procedure selects the true causal direction (Casella & Berger, 2021).

Directional support probability $\widehat{P}_{\text{CDSP}}$ can also be interpreted as a concordance probability, that is, the probability that one random quantity exceeds another. Probabilities of this form arise in several statistical contexts, including the area under the receiver operating characteristic curve (AUC) (Hanley & McNeil, 1982). In these settings, the quantity represents the probability that a randomly drawn observation from one group exceeds a randomly drawn observation from another. In the CDSP procedure, the randomness instead arises

from the sampling distribution of the estimators $\hat{I}_{Y\to X,n}$ and $\hat{I}_{X\to Y,n}$. Accordingly, $\widehat{P}_{\text{CDSP}}$ measures the probability that the directional evidence favoring $X \to Y$ exceeds that favoring $Y \to X$ under sampling variability. Values of $\widehat{P}_{\text{CDSP}}$ near 0.5 indicate little directional separation between the two candidate directions, whereas values farther from 0.5 indicate stronger directional dominance. Table 5.1 provides a heuristic interpretation scale adapted from AUC discrimination guidelines (Hosmer et al., 2013).

| $\widehat{P}_{\text{CDSP}}$ | Interpretation |
|---|---|
| $\approx 0.5$ | Little or no directional separation |
| 0.5–0.7 | Weak directional support |
| 0.7–0.8 | Moderate directional support |
| 0.8–0.9 | Strong directional support |
| $\geq 0.9$ | Very strong directional support |

Table 5.1: Guidelines for interpretation of directional support $\widehat{P}_{\text{CDSP}}$ values.

### *5.2.3 Establishing Consistency of CDSP Direction and Support Probability Estimates*

In this section, we establish consistency of the inferred causal direction via effect-size asymmetry, and of the directional support probability $\widehat{P}_{\text{CDSP}}$ from Algorithm 3. In particular, we show that under the effect-size asymmetry assumption (Assumption 5.2.3), the inferred causal direction is consistent.

**Theorem 5.2.2** (Consistency of Directional Estimate via Effect-Size Asymmetry)**.** *Let* $c^{\alpha}_{X\to Y,n} := c^{\alpha}_{X\to Y}/n, c^{\alpha}_{Y\to X,n} := c^{\alpha}_{Y\to X}/n$, *with plug-in estimates* $\hat{c}^{\alpha}_{X\to Y,n} := \hat{c}^{\alpha}_{X\to Y}/n, \hat{c}^{\alpha}_{Y\to X,n} := \hat{c}^{\alpha}_{Y\to X}/n$. *Suppose* $\sigma_{Y\to X}$ *and* $\sigma_{X\to Y}$ *are bounded away from zero. Under standard regularity conditions, the estimators described in Algorithm 3 satisfy*

$$\hat{\sigma}_{Y\to X} \xrightarrow{p} \sigma_{Y\to X}, \quad \hat{c}^{\alpha}_{Y\to X} \xrightarrow{p} c^{\alpha}_{Y\to X}, \quad \hat{\theta}_{Y\to X} \xrightarrow{p} \theta_{Y\to X},$$

*and*

$$\hat{\sigma}_{X\to Y} \xrightarrow{p} \sigma_{X\to Y}, \quad \hat{c}^{\alpha}_{X\to Y} \xrightarrow{p} c^{\alpha}_{X\to Y}, \quad \hat{\theta}_{X\to Y} \xrightarrow{p} \theta_{X\to Y}.$$

*Furthermore, let* $I_{Y\to X} \coloneqq \frac{\theta_{Y\to X}}{\sigma_{Y\to X}}$ *and* $I_{X\to Y} \coloneqq \frac{\theta_{X\to Y}}{\sigma_{X\to Y}}$. *Then, if the asymptotic effect-size asymmetry assumption holds (Assumption 5.2.3), that is,*

$$I_{Y\to X} > I_{X\to Y},$$

*then*

$$\mathbb{P}\big(\hat{I}_{Y\to X,n} > \hat{I}_{X\to Y,n}\big) \to 1 \quad \text{as } n \to \infty.$$

*Proof.* Prior work establishes consistency of the bootstrap estimators $\hat{c}^{\alpha}_{Y\to X}, \hat{c}^{\alpha}_{X\to Y}, \hat{\theta}_{Y\to X}, \hat{\theta}_{X\to Y}, \hat{\sigma}_{Y\to X}$, and $\hat{\sigma}_{X\to Y}$ under standard regularity conditions (e.g., i.i.d. sampling and characteristic kernels) (Efron & Tibshirani, 1993; Gretton et al., 2008; Sen & Sen, 2014).

In particular, by the bootstrap consistency result of Sen and Sen (2014), the estimated critical values satisfy

$$\hat{c}^{\alpha}_{Y\to X} \xrightarrow{p} c^{\alpha}_{X}, \quad \hat{c}^{\alpha}_{X\to Y} \xrightarrow{p} c^{\alpha}_{X\to Y}. \tag{5.16}$$

Since HSIC is a V-statistic, its empirical form is consistent (Gretton et al., 2008), implying

$$\hat{\theta}_{Y\to X} = \hat{\theta}(Y,\delta) \xrightarrow{p} \theta(Y,\delta), \quad \hat{\theta}_{X\to Y} = \hat{\theta}(X,\varepsilon) \xrightarrow{p} \theta(X,\varepsilon). \tag{5.17}$$

Additionally, by standard bootstrap theory (Efron & Tibshirani, 1993),

$$\hat{\sigma}_{Y\to X} \xrightarrow{p} \sigma_{Y\to X}, \quad \hat{\sigma}_{X\to Y} \xrightarrow{p} \sigma_{X\to Y}. \tag{5.18}$$

Because $\sigma_{Y\to X}$ and $\sigma_{X\to Y}$ are bounded away from zero, the continuous mapping theorem

(van der Vaart, 1998) gives

$$\hat{I}_{Y\to X,n} = \frac{\hat{\theta}_{Y\to X} - \hat{c}^{\alpha}_{Y\to X,n}}{\hat{\sigma}_{Y\to X}} \xrightarrow{p} \frac{\theta_{Y\to X}}{\sigma_{Y\to X}} = I_{Y\to X}, \qquad \hat{I}_{X\to Y,n} = \frac{\hat{\theta}_{X\to Y} - \hat{c}^{\alpha}_{X\to Y,n}}{\hat{\sigma}_{X\to Y}} \xrightarrow{p} \frac{\theta_{X\to Y}}{\sigma_{X\to Y}} = I_{X\to Y}.$$

Here we use that

$$c^{\alpha}_{Y\to X,n} = \frac{c^{\alpha}_{Y\to X}}{n} = o(1), \qquad c^{\alpha}_{X\to Y,n} = \frac{c^{\alpha}_{X\to Y}}{n} = o(1).$$

By the asymptotic effect-size asymmetry assumption (Assumption 5.2.3), we have $I_{Y\to X} > I_{X\to Y}$. Let $\Delta := I_{Y\to X} - I_{X\to Y} > 0$ and take $\varepsilon := \Delta/4$. If $|\hat{I}_{Y\to X,n} - I_{Y\to X}| < \varepsilon$ and $|\hat{I}_{X\to Y,n} - I_{X\to Y}| < \varepsilon$, then

$$\hat{I}_{Y\to X,n} > I_{Y\to X} - \varepsilon = I_{Y\to X} - \tfrac{\Delta}{4} \;>\; I_{X\to Y} + \tfrac{\Delta}{4} = I_{X\to Y} + \varepsilon > \hat{I}_{X\to Y,n},$$

hence $\hat{I}_{Y\to X,n} > \hat{I}_{X\to Y,n}$. Therefore

$$\{\hat{I}_{Y\to X,n} \le \hat{I}_{X\to Y,n}\} \;\subseteq\; \{|\hat{I}_{Y\to X,n} - I_{Y\to X}| \ge \varepsilon\} \cup \{|\hat{I}_{X\to Y,n} - I_{X\to Y}| \ge \varepsilon\},$$

and so

$$\mathbb{P}(\hat{I}_{Y\to X,n} \le \hat{I}_{X\to Y,n}) \;\le\; \mathbb{P}(|\hat{I}_{Y\to X,n} - I_{Y\to X}| \ge \varepsilon) \;+\; \mathbb{P}(|\hat{I}_{X\to Y,n} - I_{X\to Y}| \ge \varepsilon) \;\longrightarrow\; 0.$$

Thus $\mathbb{P}(\hat{I}_{Y\to X,n} > \hat{I}_{X\to Y,n}) \to 1$. □

For inference, we use the bootstrap estimate $\widehat{P}_{\mathrm{CDSP}}$. Proposition 1 establishes consistency of $\widehat{P}_{\mathrm{CDSP}}$. In particular, when the effect-size asymmetry assumption holds for the true direction (e.g., $X \to Y$), the probability that CDSP selects $X \to Y$ converges to one as the sample size increases, while the probability of selecting $Y \to X$ converges to zero.

**Proposition 1** (Consistency of Bootstrap Directional Support Probability)**.** *Let* $\mathcal{O}_n = \{(X_i, Y_i)\}_{i=1}^{n}$ *denote an i.i.d. sample of size* $n$*, and let directional detectability indices* $I_{Y\to X,n}$ *and* $I_{X\to Y,n}$ *be defined as in Equation* (5.9)*, with* $D_n := \hat{I}_{Y\to X,n} - \hat{I}_{X\to Y,n}$*. By Theorem 5.2.2,*

$\mathbb{P}(D_n > 0) \to 1$, *so that* $D_n$ *asymptotically reflects the true causal direction. Let* $\mathcal{O}_n^*$ *denote a nonparametric i.i.d. bootstrap resample drawn from* $\mathcal{O}_n$*, and let* $D_n^*$ *denote* $D_n$ *computed from* $\mathcal{O}_n^*$*. For the b-th bootstrap resample, let* $D_n^{(b)} := \hat{I}_{Y\to X,n}^{(b)} - \hat{I}_{X\to Y,n}^{(b)}$*, and define*

$$\widehat{P}_{\mathrm{CDSP}} = \frac{1}{B_n}\sum_{b=1}^{B_n} \mathbf{1}\Big\{D_n^{(b)} > 0\Big\}, \qquad B_n \to \infty.$$

*Define the (finite-n) directional support probability*

$$P_{\mathrm{CDSP},n} := \mathbb{P}(D_n > 0).$$

*Assume the nonparametric i.i.d. bootstrap is asymptotically valid for* $D_n$ *in the sense that, conditionally on* $\mathcal{O}_n$*,*

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}(D_n^* \le t \mid \mathcal{O}_n) - \mathbb{P}(D_n \le t) \right| \xrightarrow{p} 0.$$

*Then*

$$\widehat{P}_{\mathrm{CDSP}} - P_{\mathrm{CDSP},n} \xrightarrow{p} 0, \qquad \textit{and} \qquad \widehat{P}_{\mathrm{CDSP}} \xrightarrow{p} 1.$$

*Proof.* Let

$$P_n^* := \mathbb{P}(D_n^* > 0 \mid \mathcal{O}_n).$$

Conditional on $\mathcal{O}_n$, the variables $\{\mathbf{1}(D_n^{(b)} > 0)\}_{b=1}^{B_n}$ are i.i.d. Bernoulli$(P_n^*)$. Hence, by the conditional law of large numbers,

$$\widehat{P}_{\mathrm{CDSP}} - P_n^* \xrightarrow{p} 0 \qquad (B_n \to \infty).$$

It remains to show that $P_n^* - P_{\mathrm{CDSP},n} \xrightarrow{p} 0$. By bootstrap consistency,

$$\mathbb{P}(D_n^* \le 0 \mid \mathcal{O}_n) - \mathbb{P}(D_n \le 0) \xrightarrow{p} 0,$$

which implies

$$P_n^* - P_{\mathrm{CDSP},n} = \mathbb{P}(D_n \le 0) - \mathbb{P}(D_n^* \le 0 \mid \mathcal{O}_n) \xrightarrow{p} 0.$$

Thus,

$$\widehat{P}_{\text{CDSP}} - P_{\text{CDSP},n} \xrightarrow{p} 0.$$

Finally, since $\mathbb{P}(D_n > 0) \xrightarrow{1}$, we have $P_{\text{CDSP},n} \to 1$, and therefore $\widehat{P}_{\text{CDSP}} \xrightarrow{p} 1$. □

**Remark 4.** *Under the assumptions of Theorem 5.2.2, $D_n$ is a smooth plug-in functional of the empirical distribution, as it is composed of consistent estimates with denominators bounded away from zero. Under the nonparametric i.i.d. bootstrap, such smooth functionals are bootstrap-valid under standard regularity conditions.*

### *5.3 Numerical Results*

In this section, we evaluate performance of CDSP through both simulated and real data experiments.

#### *5.3.1 Simulations*

In our simulations, we compare the accuracy of CDSP and LiNGAM under increasing levels of linear model misspecification. To evaluate CDSP, we first assess whether the population effect-size asymmetry assumption holds using the ground truth quantities $I_{Y\to X} = \frac{\theta_{Y\to X}}{\sigma_{Y\to X}}$ and $I_{X\to Y} = \frac{\theta_{X\to Y}}{\sigma_{X\to Y}}$.[4] We then examine how effect-size asymmetry is reflected in finite samples via the distributions of the directional detectability indices $I_{Y\to X,n}$ and $I_{X\to Y,n}$ across Monte Carlo replications. As misspecification increases, the overlap between the sampling distributions of $I_{Y\to X,n}$ and $I_{X\to Y,n}$ grows, leading to a gradual decline in accuracy of the CDSP directional estimate and, under severe misspecification, a breakdown of the effect-size asymmetry assumption. In contrast, LiNGAM exhibits abrupt failure to detect the causal direction even under mild misspecification. These simulation results indicate that CDSP is more robust to misspecification than LiNGAM.

[4]Using the test statistics of Sen and Sen (2014), the critical values satisfy $c^{\alpha}_{X\to Y,n} = c^{\alpha}_{X\to Y}/n$ and $c^{\alpha}_{Y\to X,n} = c^{\alpha}_{Y\to X}/n$. Consequently, the corresponding standardized critical values satisfy $q^{\alpha}_{Y\to X,n} - q^{\alpha}_{X\to Y,n} = o(1)$, so asymptotically the contribution of the critical values is negligible, and population effect-size asymmetry is determined by $I_{Y\to X}$ and $I_{X\to Y}$.

We generate bivariate datasets under the true causal direction $X \to Y$, considering settings that either satisfy the model assumptions or exhibit six increasing levels of linear model misspecification. Data are generated according to

$$Y = \text{sign}(X - a)\,|X - a|^d \beta + \eta, \tag{5.19}$$

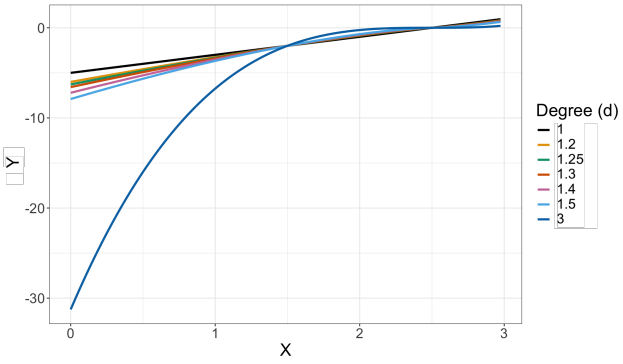



Figure 5.1: Settings of linearity

where the exponent $d$ controls the degree of nonlinearity. As illustrated in Figure 5.1, we consider $d \in \{1, 1.2, 1.25, 1.3, 1.4, 1.5, 3\}$ where $d = 1$ corresponds to the linear case and larger values represent progressively stronger departures from linearity. All other assumptions, aside from linearity, are satisfied. We sample $\eta$ from a Gaussian Mixture Model (GMM) with 3 mixtures. We sample $X$ from an exponential distribution truncated to the interval $(0, 3)$, ensuring that $Y$ is non-Gaussian. As seen in Figure 5.1, the case $d = 3$ represents a regime of severe model misspecification that lies outside the settings for which CDSP is primarily designed (Remark 3). Nonetheless, we discuss the case of $d = 3$ to highlight a regime in which the effect-size asymmetry assumption breaks down.

For each nonlinearity level, we generate $M = 100$ datasets of size $N = 3000$ from the true data-generating process in Equation (5.19). For each dataset, we obtain a point estimate of the causal direction using CDSP (Algorithm 3) with the test statistic of Sen and Sen (2014), based on the Hilbert–Schmidt Independence Criterion (HSIC). We compare the CDSP point

estimates of causal direction to those obtained from DirectLiNGAM (Shimizu et al., 2011). To ensure comparability, we use HSIC as the independence measure within DirectLiNGAM.

To evaluate the population effect-size asymmetry assumption (Assumption 5.2.3), we compute the required population quantities. We use a large Monte Carlo sample of size $10^5$ to estimate $\theta(X, \varepsilon)$ and $\theta(Y, \delta)$. We estimate $\sigma_{1b,X\to Y}$ and $\sigma_{1c,Y\to X}$ as $\sqrt{N}$ times the sample standard deviation of the test statistics $T_{1b,X\to Y}$ and $T_{1c,Y\to X}$ across the $M = 100$ Monte Carlo replications, where $N = 3000$ is the sample size of each dataset. We compute accuracy of CDSP directional estimate as the empirical fraction of Monte Carlo replications in which the estimated ordering supports the direction favored by the population effect-size asymmetry assumption. We compute accuracy of DirectLiNGAM as the proportion of replications in which DirectLiNGAM correctly detects $X \to Y$. When the effect-size asymmetry assumption holds, accuracy of CDSP directional estimate reduces to the proportion of replications in which the correct direction is selected. Further details on the simulation setup are in the Appendix. Figure 5.2 illustrates the sampling distributions of directional detectability indices $I_{Y\to X,n}$ and $I_{X\to Y,n}$ along with their corresponding population values ($I_{Y\to X}$ and $I_{X\to Y}$), while Table 5.2 reports accuracies of the CDSP and LiNGAM directional estimates across mild to moderate misspecification settings ($d \in \{1, 1.2, 1.25, 1.3, 1.4, 1.5\}$). In all cases, the effect-size asymmetry assumption holds, so both CDSP and LiNGAM accuracies are proportions of replications selecting the correct direction.

| | **Accuracy** | |
|---|---|---|
| Degree ($d$) | **CDSP** | **LiNGAM** |
| $d = 1$ | 100% | 100% |
| $d = 1.2$ | 100% | 100% |
| $d = 1.25$ | 100% | 0% |
| $d = 1.3$ | 100% | 0% |
| $d = 1.4$ | 88% | 0% |
| $d = 1.5$ | 57% | 0% |

Table 5.2: Accuracies of CDSP and LiNGAM directional estimates for polynomial degrees $d \in \{1, 1.2, 1.25, 1.3, 1.4, 1.5\}$.

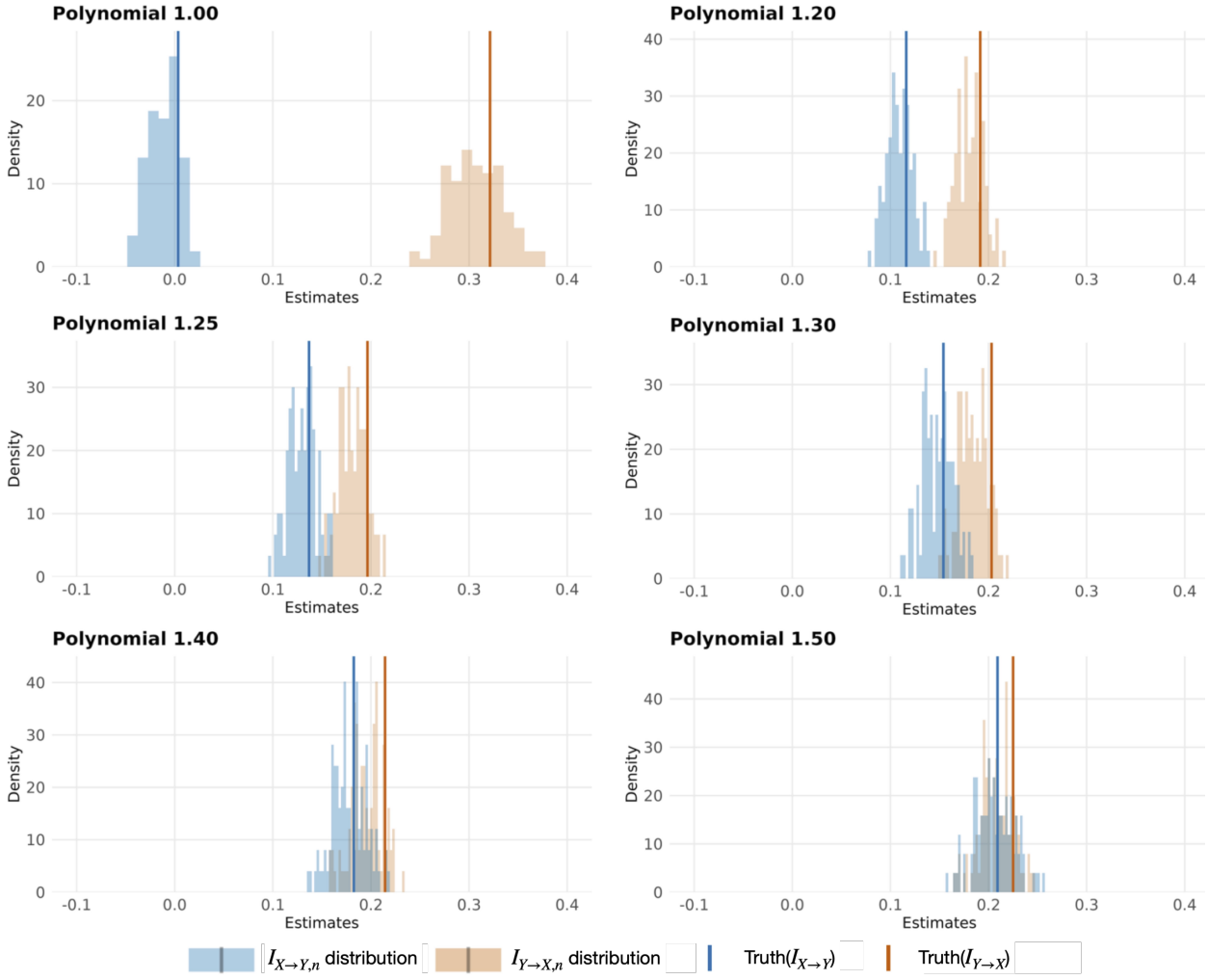



Figure 5.2: Distribution of the estimated directional detectability indices $I_{X\to Y,n}$ (blue) and $I_{Y\to X,n}$ (orange) under varying degrees of nonlinearity. The top-left panel corresponds to the linear case ($d = 1$). The remaining panels correspond to $d = 1.2$ (top-right), 1.25 (middle-left), 1.3 (middle-right), 1.4 (bottom-left), and 1.5 (bottom-right), representing progressively stronger departures from linearity. Solid blue and orange vertical lines indicate population values $I_{X\to Y}$ and $I_{Y\to X}$, respectively.

When linearity holds ($d = 1$, top left in Figure 5.2 and first row of Table 5.2), the sampling distributions of $I_{X\to Y,n}$ and $I_{Y\to X,n}$ are well separated. In this setting, both CDSP and LiNGAM also correctly identify the direction across all $M$ replications. Because $H^0_{X\to Y}$ holds, the corresponding population departure from the null, $\theta_{X\to Y}$, is zero. In contrast, $H^0_{Y\to X}$ does

not hold, so the corresponding population departure from the null, $\theta_{Y\to X}$, is strictly positive. This pronounced asymmetry induces a clear separation between the population values $I_{Y\to X}$ and $I_{X\to Y}$, which is reflected in the well-separated sampling distributions of the directional detectability indices.

As the degree of misspecification increases from $d = 1$ to $d = 1.5$, the separation between the population values $I_{Y\to X}$ and $I_{X\to Y}$ decreases, and the sampling distributions of the directional detectability indices $I_{Y\to X,n}$ and $I_{X\to Y,n}$ move closer together. Beginning at $d = 1.25$, the two distributions begin to overlap, and LiNGAM incorrectly identifies the direction across all replications, reflecting a breakdown of the traditional asymmetry assumption (Assumption 5.2.2). Despite the slight overlap for $d \in \{1.25, 1.3\}$, CDSP accuracy remains at 100%, indicating that the ordering $I_{Y\to X,n} > I_{X\to Y,n}$ is preserved across Monte Carlo replications. This demonstrates that, under mild model misspecification, CDSP remains substantially more robust than LiNGAM.

At higher levels of misspecification ($d \in \{1.4, 1.5\}$), the separation between the population values $I_{Y\to X}$ and $I_{X\to Y}$ decreases further. The sampling distributions of the directional detectability indices $I_{Y\to X,n}$ and $I_{X\to Y,n}$ exhibit substantial overlap across Monte Carlo replications, and accuracy of CDSP directional estimate decreases correspondingly. In these settings, sampling variability becomes large relative to the population gap between $I_{Y\to X}$ and $I_{X\to Y}$, so the ordering $I_{Y\to X,n} > I_{X\to Y,n}$ is no longer preserved uniformly across datasets. Nevertheless, CDSP continues to exhibit stronger directional performance than LiNGAM, which incorrectly concludes $Y \to X$ in all $M$ replications.

For the severe model misspecification case of $d = 3$, we observed that, unlike the previous settings ($d \in \{1, 1.2, 1.25, 1.3, 1.4, 1.5\}$), the effect-size asymmetry assumption no longer holds. In this setting, CDSP continues to favor the direction implied by effect-size asymmetry; however, because this asymmetry is misaligned with the true causal direction, it results in incorrect inference relative to the ground truth $X \to Y$. As a result, both CDSP and LiNGAM favor the incorrect direction ($Y \to X$) across all Monte Carlo replications. We provide additional details for the $d = 3$ case in the Appendix.

As discussed in Section 5.2, $\widehat{P}_{\text{CDSP}}$ measures the directional support for the estimated causal direction, defined as the proportion of bootstrap samples favoring the directional estimate. Similarly, LiNGAM bootstrap rates (Thamvitayakul et al., 2012b) quantify the proportion of bootstrap resamples favoring the estimated direction. The accuracies of the CDSP and LiNGAM directional estimates reported in Table 5.2 are directly related to $\widehat{P}_{\text{CDSP}}$ and LiNGAM bootstrap rates, respectively: when the estimated direction is the true causal direction, $\widehat{P}_{\text{CDSP}}$ and the LiNGAM bootstrap rate correspond to accuracy, whereas when the estimated direction is incorrect, one minus the corresponding value reflects accuracy. For CDSP and LiNGAM, the effect-size asymmetry assumption and the traditional asymmetry assumption, respectively, characterize the theoretical conditions under which the estimated direction is expected to align with the true causal direction.

To assess the role of sample size, we increase $N$ from 3000 to 6000 and then to 9000 in the $d \in \{1.4, 1.5\}$ settings, where accuracy of CDSP directional estimate is below 100%. LiNGAM is excluded from this study because its accuracy is consistently 0% in these settings, indicating failure under this level of misspecification; as a result, increasing the sample size would not provide any additional insights into its performance. As shown in Table 5.3, for $d = 1.4$, the accuracy increases from 88% to 91% and then to 94%, while for $d = 1.5$ it increases from 57% to 62% and then to 65%. This behavior is consistent with Proposition 1: as the sample size increases, the estimates of the directional detectability indices $I_{X\to Y,n}$ and $I_{Y\to X,n}$ concentrate more tightly around their population values $I_{X\to Y}$ and $I_{Y\to X}$, reducing sampling variability and stabilizing their ordering across Monte Carlo replications. Consequently, when the effect-size asymmetry assumption holds, increasing the sample size can strengthen directional stability and improve accuracy of CDSP estimate. In settings where the effect-size asymmetry assumption holds but the directional signal is weak (e.g., $d = 1.5$), this consistency may manifest in finite samples as an increasing trend in $\widehat{P}_{\text{CDSP}}$ with sample size, as the ordering between $I_{X\to Y,n}$ and $I_{Y\to X,n}$ stabilizes, thereby providing further evidence in favor of the correct direction.

In summary, our simulations demonstrate CDSP's robustness to mild to moderate model

| Degree ($d$) | $N$ = 3000 | $N$ = 6000 | $N$ = 9000 |
|---|---|---|---|
| $d$ = 1.4 | 88% | 91% | 94% |
| $d$ = 1.5 | 57% | 62% | 65% |

Table 5.3: Accuracy of CDSP directional estimate for increasing sample size $N$ for $d = \{1.4, 1.5\}$.

misspecification and show that accuracy of CDSP directional estimate slightly declines as misspecification increases from mild to moderate. We additionally illustrate that CDSP outperforms LiNGAM across settings with varying degrees of nonlinearity. Importantly, the decline in CDSP accuracy reflects increased overlap in the sampling distributions of directional detectability indices $I_{X\to Y,n}$ and $I_{Y\to X,n}$, rather than a structural failure of the method. We also demonstrate that increasing the sample size mitigates this effect: larger samples can strengthen directional stability when the effect-size asymmetry assumption holds. In contrast, LiNGAM fails to infer the correct direction even under mild misspecification, with accuracy dropping abruptly to 0%.

### *5.3.2 Real Data*

We apply CDSP as described in Algorithm 3 to the Tübingen Cause–Effect Pairs benchmark of 100 real-world univariate cause–effect pairs (Mooij et al., 2016b), and compare the resulting direction estimates and associated uncertainty to those obtained from LiNGAM. For CDSP, we quantify uncertainty via the directional strength probability $\widehat{P}_{\text{CDSP}}$. For LiNGAM, we use bootstrap selection rates as suggested by (Thamvitayakul et al., 2012b) and commonly used in practice (e.g.,Motokawa et al., 2020; Rosenström et al., 2012).

As noted by the dataset's authors, the Tübingen pairs are inherently heterogeneous: although each is believed to reflect a genuine causal relationship, the presumed ground-truth directions have not been established through intervention and hidden confounding or selection bias may be present (Mooij et al., 2016b). This heterogeneity is evident in the pairs'

scatterplots that show a range of functional forms—from approximately linear relationships to strongly nonlinear ones.

We consider two evaluation settings across pairs: the full set of 100 pairs and a subset of approximately linear pairs (i.e., pairs that do not exhibit clear violations of the linearity assumption). Given the substantial heterogeneity across all pairs, both LiNGAM and CDSP may struggle to reliably infer directionality on the full dataset. We expect better performance by both LiGNAM and CDSP on the subset of approximately linear pairs which reflects settings where both methods might reasonably be applied in practice. In addition, we further examine three pairs to better understand the comparative performance of the two procedures across different regimes: a pair where both CDSP and LiNGAM correctly identify the direction; a pair where LiNGAM selects the incorrect direction while CDSP correctly infers the direction with strong directional strength $\widehat{P}_{\text{CDSP}}$; and a setting in which LiNGAM again selects the incorrect direction while CDSP correctly infers the direction but with weak directional strength.

To identify approximately linear relationships, we compare a linear regression model to a flexible spline-based generalized additive model (GAM) following standard practice in the GAM literature (Ellison, 2022; Wood, 2017). For each pair $(X, Y)$ we fit (i) a linear regression model, $Y = \beta_0 + \beta_1 X + \varepsilon$ and (ii) a semiparametric GAM, $Y = \beta_0 + f(X) + \varepsilon$, where $f$ is a smooth function estimated using penalized regression splines with basis dimension $k = 10$ (Wood, 2017, 2003). GAM Smoothing parameters are estimated using maximum likelihood to allow likelihood-based comparison across models (Wood, 2017). We compare the two models using the Bayesian Information Criterion (BIC) (Schwarz, 1978) and classify a pair as approximately linear when the linear model achieves lower BIC (Kutner et al., 2005).

We assess performance of LiNGAM and CDSP on the benchmark pairs using two metrics: the *true discovery rate (TDR)* and *false discovery rate (FDR)*. Following standard usage in statistical model evaluation (e.g., Fawcett, 2006; Powers, 2011), we define

$$\text{TDR} = \frac{\text{number of correct decisions}}{\text{number of all decisions}}, \qquad \text{FDR} = \frac{\text{number of incorrect decisions}}{\text{number of all decisions}}.$$

Table 5.4: Comparison of CDSP and LiNGAM on all 100 benchmark pairs.

(a) Performance metrics across all 100 pairs.

| Metric | CDSP | LiNGAM |
|---|---|---|
| True Discovery Rate | $62/100 = 62\%$ | $44/100 = 44\%$ |
| False Discovery Rate | $38/100 = 38\%$ | $56/100 = 56\%$ |

(b) Contingency table comparing causal discovery decisions across all 100 pairs.

| | LiNGAM Right | LiNGAM Wrong | Total |
|---|---|---|---|
| CDSP Right | 43 | 19 | 62 |
| CDSP Wrong | 1 | 37 | 38 |
| Total | 44 | 56 | 100 |

(c) False discovery rates by $P_{\text{CDSP}}$-defined support categories across all 100 pairs.

| Support Category | CDSP FDR | LiNGAM FDR |
|---|---|---|
| No/little | $2/3 = 66.7\%$ | $2/8 = 25\%$ |
| Weak | $18/45 = 40\%$ | $13/15 = 86.7\%$ |
| Moderate | $4/8 = 50\%$ | $7/10 = 70\%$ |
| Strong | $8/17 = 47.1\%$ | $6/10 = 60\%$ |
| Very strong | $6/27 = 22.2\%$ | $28/57 = 49.1\%$ |
| Total | $38/100 = 38\%$ | $56/100 = 56\%$ |

We present results across all 100 benchmark pairs in Table 5.4. Across the 100 real-data examples, CDSP achieves a substantially lower false discovery rate (38% vs. 56%) and a higher true discovery rate (62% vs. 44%). When the two methods disagree, CDSP is correct far more often: it is correct in 19 cases where LiNGAM is wrong, whereas LiNGAM is correct in only one case where CDSP is wrong.

In Table 5.4c, we stratify both $P_{\text{CDSP}}$ and LiNGAM bootstrap rates into the directional support categories defined in Table 5.1. We use the $P_{\text{CDSP}}$-defined bins since no established interpretive scale exists for LiNGAM bootstrap rates. For a reliable method, we would expect false discovery rates to decrease as directional support increases, with low-support categories exhibiting higher error rates and high-support categories exhibiting lower error rates. We

Table 5.5: Comparison of CDSP and LiNGAM on 34 approximately linear pairs.

(a) Performance metrics across a subset of 34 approximately linear pairs.

| Metric | CDSP | LiNGAM |
|---|---|---|
| True Discovery Rate | 27/34 = 79.4% | 21/34 = 61.8% |
| False Discovery Rate | 7/34 = 20.6% | 13/34 = 38.2% |

(b) Contingency table comparing causal discovery decisions across 34 approximately linear pairs.

| | LiNGAM Right | LiNGAM Wrong | Total |
|---|---|---|---|
| CDSP Right | 21 | 6 | 27 |
| CDSP Wrong | 0 | 7 | 7 |
| Total | 21 | 13 | 34 |

(c) False discovery rates by $P_{\text{CDSP}}$-defined support categories across 34 approximately linear pairs.

| Support Category | CDSP FDR | LiNGAM FDR |
|---|---|---|
| No/little | 1/1 = 100.0% | 1/1 = 100.0% |
| Weak | 4/9 = 44.4% | 7/10 = 70.0% |
| Moderate | 2/5 = 40.0% | 0/2 = 0.0% |
| Strong | 0/7 = 0.0% | 2/4 = 50.0% |
| Very strong | 0/12 = 0.0% | 3/17 = 17.6% |
| Total | 7/34 = 20.6% | 13/34 = 38.2% |

observe that LiNGAM-based bootstap rates provide 'very strong' directional support to more than half of the pairs, yet LiNGAM exhibits a high false discovery rate within this category (49.1%, compared to 22.2% for CDSP). In contrast, CDSP exhibits higher false discovery rates in lower-support categories and substantially lower false discovery rates in higher-support categories. Thus, results in Table 5.4c indicate that $P_{\text{CDSP}}$ provides a more meaningful measure of uncertainty quantification than LiNGAM's bootstrap rate.

Table 5.5 presents results for the approximately linear pairs. On this subset, as expected, classification performance improves for both methods; however, CDSP maintains a consistent advantage. CDSP again exhibits a lower false discovery rate (20.6% vs. 38.2%) and a higher true discovery rate (79.4% vs. 61.8%). In cases of disagreement, CDSP is correct 6 times, while LiNGAM is never correct when CDSP is wrong. In Table 5.5c, we again stratify both $P_{\text{CDSP}}$ and LiNGAM bootstrap rates into the directional support categories defined in Table 5.1. We observe that false discovery rates are lower for both CDSP and LiNGAM in this subset compared to the full dataset. However, CDSP continues to show lower false discovery rates than LiNGAM, particularly in the higher-support categories, providing us with uncertainty quantification that is more meaningful than LiNGAM's bootstrap rates.

In Figure 5.3, we examine three approximately linear pairs, selected to highlight distinct scenarios in which CDSP and LiNGAM differ in performance. The first pair relates population growth to food consumption and contains 347 observations from 174 countries or regions, covering two time periods: 1990–1992 to 1995–1997 and 1995–1997 to 2000–2002 (with one missing entry). The data, collected by the Food and Agriculture Organization of the United Nations, record the average annual rate of change in population ($X$) and total dietary consumption ($Y$). Following Mooij et al. (2016b), we treat the causal direction as $X \to Y$. The scatterplot in Figure 5.3 suggests an approximately linear relationship between the two variables. In this well-behaved setting, both LiNGAM and CDSP recover the correct direction. CDSP yields $\widehat{P}_{\text{CDSP}} = 0.80$, while LiNGAM yields a bootstrap rate of 1. As indicated in Table 5.1, this value of $\widehat{P}_{\text{CDSP}}$ corresponds to strong directional support for $X \to Y$.

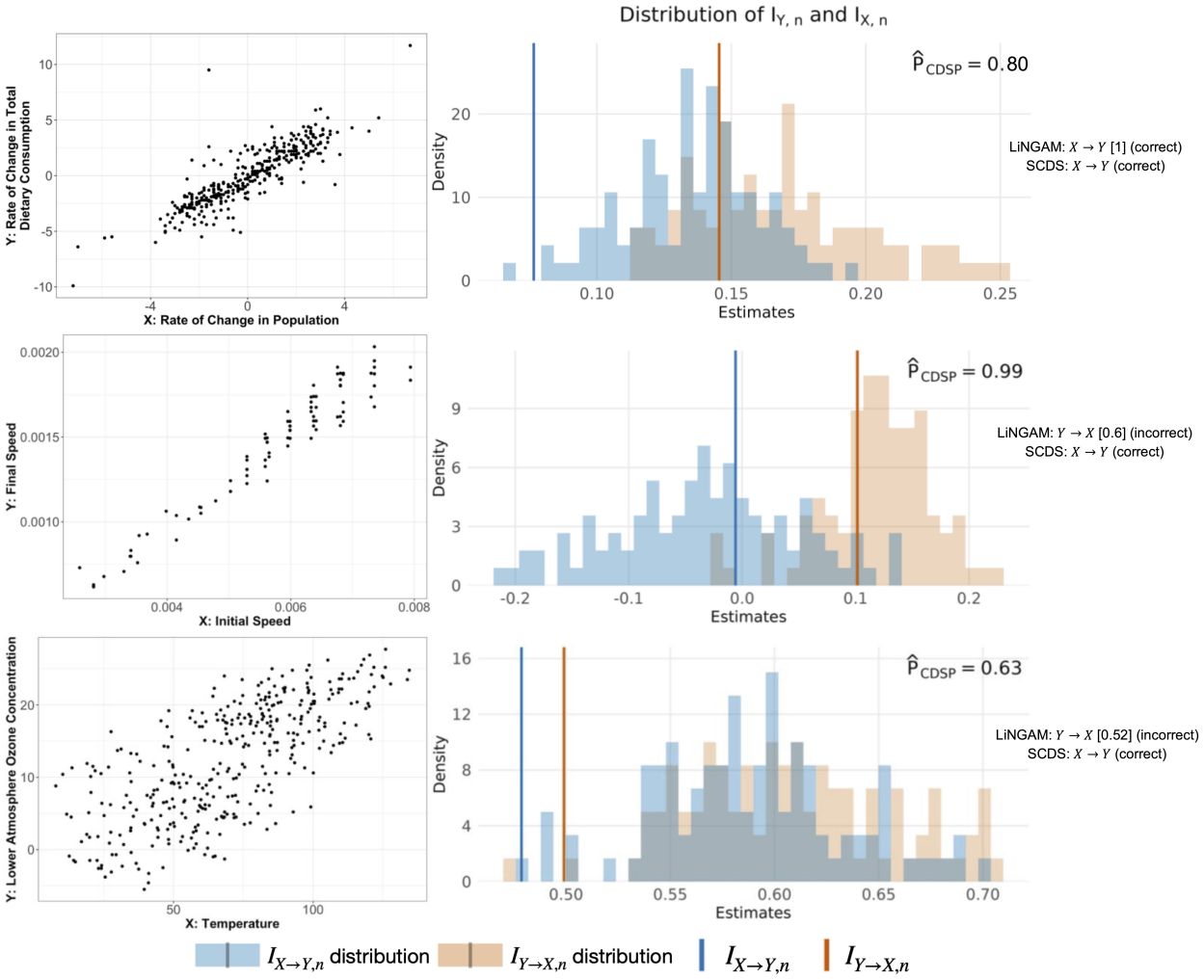



Figure 5.3: Rows 1–3 correspond to the Population and Dietary Consumption, Balltrack, and Ozone–Temperature datasets, respectively. The first column shows scatterplots of the observed data. The second column shows histograms of the bootstrap distributions of $I_{X\to Y,n}$ (blue) and $I_{Y\to X,n}$ (orange); the solid vertical lines mark the corresponding estimates computed on the full dataset. Each histogram also reports $\widehat{P}_{\text{CDSP}}$. To the right of each histogram, we report the overall LiNGAM and CDSP results; LiNGAM bootstrap rates for the inferred direction are shown in brackets.

The second pair relates the initial and final speeds of a rolling ball and consists of 94 measurements collected for the Tübingen Cause–Effect Pairs benchmark suite (Mooij et al., 2016b). The data were obtained using a physical ball track equipped with two light barrier sensors, where $X$ records the initial speed and $Y$ records the speed at a later position along the track. The causal direction is taken to be known, with initial speed causing final speed. The scatterplot (Figure 5.3) suggests an approximately linear relationship. Despite this apparent linearity, LiNGAM infers the incorrect direction, with a bootstrap rate of 0.6. In contrast, CDSP correctly identifies the direction, with ($\widehat{P}_{\text{CDSP}} = 0.99$), corresponding to very strong support (Table 5.1). Here, we observe a notable difference in conditional variability between the two directions, which may affect the variability of the corresponding test statistics. CDSP accounts for such variability in the effect-size asymmetry assumption. In contrast, LiNGAM does not explicitly account for this in the traditional asymmetry assumption, which may contribute to the discrepancy in performance, with CDSP correctly identifying the direction while LiNGAM does not.

The third pair relates ozone concentration to temperature and consists of 365 daily measurements of mean temperature and lower-atmosphere ozone concentration recorded in Lausanne-César-Roux, Switzerland in 2009 (Federal Office for the Environment (FOEN), 2009). Following Mooij et al. (2016b), we treat the causal direction as known, with changes in temperature ($X$) driving changes in ozone concentration ($Y$). The scatterplot (Figure 5.3) suggests an approximately linear relationship, making this a setting in which LiNGAM would ordinarily be expected to perform well. Despite this, LiNGAM infers the incorrect direction, with a bootstrap rate of 0.52. In contrast, CDSP correctly identifies the direction, with $\widehat{P}_{\text{CDSP}} = 0.63$, indicating weak directional support (Table 5.1). The weak support may in part be due to high variability in $X$, which can reduce power in the correct direction and diminish the separation between $\hat{I}_{X\to Y,n}$ and $\hat{I}_{Y\to X,n}$. As suggested by Proposition 1, if the effect-size asymmetry assumption holds, increasing the sample size would be expected to strengthen the directional evidence and drive $\widehat{P}_{\text{CDSP}}$ closer to one. Consistent with the patterns observed in Figure 5.3, the bootstrap distributions of the directional detectability index

estimates $\hat{I}_{X\to Y,n}$ and $\hat{I}_{Y\to X,n}$ are left-skewed, with the original-sample estimates appearing toward the lower tail. This suggests substantial variability in the sampling distributions, which can reduce the separation between $\hat{I}_{X\to Y,n}$ and $\hat{I}_{Y\to X,n}$ and thereby contribute to the weak directional support. Another possible contributing factor is that the nonparametric bootstrap relies on an i.i.d. sampling assumption (Efron & Tibshirani, 1993), which may be imperfectly satisfied in this dataset, potentially affecting stability of the bootstrap estimates.

In sum, CDSP yields materially improved performance across both the full set of 100 benchmark pairs and the subset of 34 approximately linear pairs, reducing the false discovery rate by approximately 18% in both settings relative to LiNGAM. We observe that LiNGAM bootstrap assigns very strong support to a large fraction of pairs, yet exhibits relatively high false discovery rates even within this category, particularly across all 100 pairs where assumption violations are more likely. In contrast, $P_{\text{CDSP}}$ spans a broader range of values across support categories, with higher false discovery rates in lower-support categories and substantially lower false discovery rates in higher-support categories. These patterns indicate that larger values of $\widehat{P}_{\text{CDSP}}$ are associated with lower error rates, suggesting that $\widehat{P}_{\text{CDSP}}$ provides uncertainty quantification, especially under model misspecification. In addition, the three representative pairs illustrate how CDSP distinguishes between strong and weak evidence across settings, correctly inferring the causal direction with high support in settings where model assumptions are approximately satisfied and weaker support when the evidence is less clear.

### *5.4 Discussion*

In this work, we introduce Causal Discovery via Statistical Power (CDSP), a framework that provides a statistically principled basis for causal direction estimation. Existing methods often focus on identifiability under idealized modeling assumptions and pay limited attention to uncertainty quantification and robustness under model misspecification. In contrast, CDSP places these aspects at the center of the framework by characterizing when the data contain sufficient information to favor one causal direction over the other. CDSP reframes

causal discovery as a problem of directional detectability through notions of power and effect size within a hypothesis-testing–based framework.

Through simulation studies involving linear model misspecification, we demonstrate that statistical inference via CDSP yields robust direction estimates and meaningful uncertainty quantification, while improving recovery of true causal direction relative to the commonly used LiNGAM. In addition, CDSP provides diagnostic insight into potential assumption violations through the underlying directional hypothesis test outcomes (Equation 5.2). In particular, cases in which both directional hypotheses are either rejected or not rejected may indicate potential model misspecification or identifiability issues, respectively.

The most important practical consideration for using CDSP is that it relies on the effect-size asymmetry assumption (Assumption 5.2.3), which represents a first step toward elucidating how effect size influences causal direction detection. This assumption cannot be directly verified in practice, since the true data-generating mechanism is unknown. However, it is weaker than the traditional asymmetry assumption (Assumption 5.2.2) implicitly relied upon in functional causal discovery methods, which likewise cannot be verified in practice. The strength of the effect-size asymmetry depends on the choice of a test statistic, as different measures of dependence can capture deviations from the null to varying degrees. In this work, we demonstrate CDSP using the test statistic of Sen and Sen (2014), which is based on the Hilbert–Schmidt independence criterion (Gretton et al., 2008). In our simulation studies, we observe that the effect-size asymmetry may fail under severe model misspecification but tends to hold under mild to moderate deviations. In practice, we therefore recommend applying standard model checking procedures appropriate to the assumed model class—e.g., linearity diagnostics for linear models (Kutner et al., 2005; Weisberg, 2005)—to avoid settings with severe misspecification, where the effect-size asymmetry assumption may not hold. Further evaluation of the effect-size asymmetry assumption—e.g., across different test statistics and more extensive simulation settings—remains an important direction for future work.

Our practical recommendation to ensure appropriateness of the assumed model class is confirmed by the real data analyses of the Tübingen cause–effect pairs (Mooij et al., 2016a).

Thus, identifying 34 approximately linear pairs led to substantial performance improvements for both CDSP and LiNGAM. In particular, CDSP's false discovery rate decreased by 17.4% relative to its performance on the full set of 100 pairs. In addition, CDSP performed uniformly better than LiNGAM on real data, with its false discovery rate being approximately 18% lower than LiNGAM's for both the full set of 100 cause-effect benchmark pairs and for the subset of 34 approximately linear pairs. We should note that our numerical findings on the Tübingen cause–effect pairs come with the caveat that, although widely treated as known for demonstration purposes in the causal discovery literature (e.g., (Compton et al., 2020; Kocaoglu et al., 2020; Marx & Vreeken, 2017; Peters et al., 2017)), the ground-truth causal directions are not known with complete certainty.

Another practical consideration in CDSP is the choice of resampling method. We use the nonparametric bootstrap to estimate the quantities required to infer causal directionality via effect-size asymmetry and $P_{\text{CDSP}}$, and establish consistency of these estimates in Section 5.2.3. Prior work has identified limitations of the nonparametric bootstrap in specific settings, such as when particular test statistics or model selection criteria are evaluated on bootstrap samples, or when combined with causal discovery followed by downstream causal effect estimation (Chang et al., 2026; Janitza et al., 2016). Even though these existing results pertain to different inferential targets, investigating how alternative resampling approaches affect CDSP performance and theoretical guarantees could be an important direction for future work.

We note that CDSP's notion of "directional power"—defined as the probability of correctly detecting the causal direction—is conceptually related to classical power in hypothesis testing, but is used here to characterize directional detectability rather than to perform classical power analysis (Casella & Berger, 2021; Cohen, 1988; Lehmann & Romano, 2005). Our framework does not naturally lend itself to classical power analysis for causal discovery because such analyses would require specifying directional effect sizes, thereby assuming the causal direction—the very quantity the procedure seeks to infer.

In this chapter, we illustrate CDSP in the context of linear model misspecification for

bivariate causal discovery. However, CDSP applies more broadly to functional causal discovery methods, including ANMs and PNL models. In addition, beyond model misspecification, violations of other assumptions—such as unobserved confounding, cyclicity, or non-i.i.d. data—may also affect causal direction estimation. Understanding how such violations influence the effect-size asymmetry underlying CDSP remains an important direction for future work. In particular, studying how different types of assumption violations—individually or in combination—affect the effect-size asymmetry, and therefore the ability of CDSP to correctly identify the causal direction, is an open problem. Finally, in this work, we study CDSP in the bivariate setting. Extending the framework to higher-dimensional settings and developing scalable approaches for such applications are important directions for future research.

Overall, this chapter takes a first step toward a power-motivated statistical framework for causal discovery. We encourage researchers use this method to learn when the data provide sufficient evidence to favor a causal direction, even in the presence of model misspecification.

# Chapter 6

# DISCUSSION

## *6.1 Contributions*

In this dissertation, we develop statistical methodologies for causal decomposition and causal discovery, motivated by the challenges of applying causal methods reliably in complex, real-world settings where underlying assumptions may not align with the data or may be violated in practice. We treat both causal decomposition and causal discovery as fundamentally statistical methodologies, whose application requires careful attention to uncertainty, model assumptions, and real-world complexities. Across the chapters, the focus is on developing methods that remain informative and interpretable in realistic settings, particularly in the presence of model misspecification and practical constraints. The three lines of work in this dissertation address complementary aspects of this objective, as summarized below.

In Chapter 3, we extended the causal decomposition framework to study Black–white disparities in NIH R01 peer review, focusing on selection into panel discussion. We used causal decomposition to evaluate what factors causally contribute to these disparities and whether hypothetical interventions on those factors could reduce them. We also conducted a thought experiment exploring the potential impact of the NIH's transition to binary scoring for the Investigator and Environment criteria under the Simplified Peer Review Framework. Our analysis demonstrated how causal decomposition can be used to evaluate and pilot interventions in settings with features such as interference, complex institutional structure and constraints, and highlighted which types of interventions have the greatest potential to reduce Black–white disparities in selection into panel discussion.

In Chapter 4, we addressed a key gap in functional causal discovery by developing a hypothesis testing-based framework that enables formal statistical inference on causal direction

estimation. Existing functional causal discovery approaches use structural asymmetries to identify causal directionality but rely on strong modeling assumptions and provide limited support for uncertainty quantification. To address these limitations, we developed a hypothesis testing-based framework that enables uncertainty quantification, characterizes possible causal discovery outcomes, and provides diagnostic insight into potential assumption violations. This approach allows practitioners to report evidence for a causal direction, quantify uncertainty in that conclusion, and obtain diagnostic insight into potential assumption violations. Importantly, this framework applies broadly across major functional causal model classes, providing a unified inferential perspective for functional causal discovery.

In Chapter 5, we built upon this foundational hypothesis testing-based framework and introduced Causal Discovery via Statistical Power (CDSP), a robust framework for direction detection that connects causal direction estimation with statistical power. By introducing notions of power and effect size tailored to causal discovery, we characterized when one direction is more detectable than the other through effect-size asymmetry. This framework provides both direction estimates and a measure of directional support, enabling inference on the strength of evidence for a given direction. Simulation studies under varying degrees of model misspecification showed that CDSP is robust to mild and moderate violations and performs favorably relative to existing methods. Real data analyses on 100 cause–effect benchmark pairs further demonstrated improved performance in practice, reducing the false discovery rate by approximately 18% relative to LiNGAM. These results demonstrate that CDSP provides a statistically grounded and practically effective approach to causal direction estimation.

Together, the work in Chapters 4 and 5 moves causal discovery methods beyond reliance on idealized assumptions toward approaches that explicitly account for uncertainty, robustness, and the challenges of observational data.

## *6.2 Discussion and Future Work*

A unifying theme across these contributions is the emphasis on developing causal methods that remain informative and useful for decision-making in realistic settings. In practice, causal analyses are often conducted in environments with complex data structures and uncertainty about underlying mechanisms. The methods developed in this dissertation address these challenges by providing interpretable outputs that clarify what can be learned from the available data and which conclusions are supported under the assumed framework. In the NIH application, this perspective enables the evaluation of interventions relevant to NIH peer review and clarifies which types of changes have the potential to meaningfully reduce disparities. In the causal discovery setting, this perspective motivates frameworks that quantify uncertainty, provide diagnostic insight, and characterize when the data contain sufficient information to support a directional conclusion. Several directions for future research emerge from this work.

**Causal Decomposition:** For causal decomposition, one direction is to incorporate additional applicant- and proposal-level information that may improve the precision and robustness of the estimated effects. In the NIH application, the absence of variables such as bibliometric profiles and mentorship networks may limit the precision of the estimates, although the included variables are sufficient for identification of the causal quantities of interest.

A second direction is to extend the causal decomposition framework to evaluate interventions at other stages of the peer review process. While the analysis in this dissertation focuses on selection into panel discussion, the same framework could be applied to study hypothetical interventions in downstream outcomes, such as funding decisions, provided that the required identification assumptions remain reasonable. More broadly, the framework can be used to pilot and evaluate other proposed policy changes by formally specifying the corresponding estimands and assumptions and assessing their implications using observed data.

This highlights the potential for causal decomposition to provide insight into policy-relevant questions in complex institutional settings.

**Causal Discovery:** For the causal discovery frameworks developed in Chapters 4 and 5, several directions for future work remain. First, an important open problem is understanding how a wider range of assumption violations—such as unobserved confounding, cyclicity, and non-i.i.d. data—affect the behavior of the hypothesis tests underlying the test-based framework and the resulting causal discovery outcomes, as well as how such violations influence the effect-size asymmetry underlying CDSP. While the test-based framework provides diagnostic insight into potential model misspecification and lack of identifiability, and CDSP is robust to mild and moderate model misspecification, systematically characterizing the effects of a broader range of assumption violations would further enhance their diagnostic scope and practical utility.

Second, extending these frameworks to multivariate settings remains a central challenge. While both approaches can be applied in simple settings where the causal structure is known except for a pair of variables of interest, extending them to general multivariate settings—where the causal structure is unknown—introduces substantial challenges. The number of hypothesis tests and the resulting space of possible causal outcomes grows combinatorially with the number of variables, making coherent inference and uncertainty summarization challenging. Moreover, the presence of confounders complicates estimation, as it necessitates modeling and removing nuisance components across multiple variables. Developing scalable procedures and principled methods for summarizing uncertainty in these settings remains an important avenue for future research.

Third, investigating the role of resampling methods for inference in the test-based framework and CDSP is an important direction. While we use the nonparametric bootstrap in this work, understanding how alternative resampling procedures affect inference and the associated theoretical guarantees remains an important direction for future research.

**Connections and Future Directions:** An additional direction for future work is to integrate the causal discovery methods developed in Chapters 4 and 5 with the causal decomposition framework used in the NIH application in Chapter 3. An important unknown in the NIH setting is the relationship between preliminary overall impact scores and criterion scores, which is not directly observed and may vary across plausible data-generating structures (e.g., the Normative Backwards and Documentation models in Chapter 3). This uncertainty is particularly relevant for evaluating policy changes such as the Simplified Peer Review Framework, which modifies the role of criterion scores.

A key challenge in pursuing this direction is that the NIH peer review process constitutes a complex, multivariate system, whereas the causal discovery methods developed in this dissertation focus on the foundational bivariate setting, with extensions to multivariate settings remaining an important avenue for future research. In particular, applying these methods would require restricting attention to specific pairs of variables while making assumptions about the broader causal structure. More generally, it would require careful consideration of the underlying assumptions of the methods and how they interact with the complex institutional features of the NIH peer review process.

Within these constraints, the causal discovery tools developed in this dissertation could be used to assess which relationships are supported by the data and to quantify uncertainty in those conclusions. In particular, they could help determine where the data lie among a set of plausible causal structures relating preliminary overall impact and criterion scores. While such analyses may not substantially change the primary conclusions regarding disparities in selection into discussion, they could improve understanding of the underlying review process and inform the design and evaluation of additional interventions.

Finally, an important direction that spans all chapters is the application of these methods to new domains. The causal decomposition framework can be used to evaluate interventions and disparities in a variety of settings beyond NIH peer review. Similarly, the inferential perspective on causal discovery developed in this dissertation is broadly applicable to scientific problems where understanding both causal direction and the strength of evidence is

critical. Applying these methods in new domains will be important for understanding their practical performance and guiding their continued development.

Overall, this dissertation develops statistical frameworks for causal decomposition and causal discovery that support reliable application in complex, real-world settings. The causal decomposition framework enables the evaluation of interventions in settings with complex decision processes, while the causal discovery frameworks introduce a formal inferential perspective and a power-based characterization of when causal direction can be determined from the data. Together, these contributions emphasize the importance of grounding causal decomposition and discovery in formal statistical reasoning in practical applications.

# BIBLIOGRAPHY


Atkinson, A. C. (1985). *Plots, transformations, and regression: An introduction to graphical methods of diagnostic regression analysis*. Clarendon Press.

Bareinboim, E., & Pearl, J. (2012). Controlling selection bias in causal inference. *Artificial Intelligence and Statistics*, 100–108.

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT Press.

Bonetta, L. (2009). Study section overview [Accessed: 2024-10-31].

Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E.-K., Swayne, D. F., & Wickham, H. (2009). Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society A*, *367*(1906), 4361–4383.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The review of economics and statistics*, *90*(3), 414–427.

Casella, G., & Berger, R. L. (2021). *Statistical inference* (3rd). Cengage Learning.

Chang, T.-H., Guo, Z., & Malinsky, D. (2026). Post-selection inference for causal effects after causal discovery. *Biometrika*, *113*(1), asaf073.

Chiappa, S. (2019). Path-specific counterfactual fairness. *Proceedings of the AAAI conference on artificial intelligence*, *33*(01), 7801–7808.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd). Routledge.

Compton, S., Kocaoglu, M., Greenewald, K., & Katz, D. (2020). Entropic causal inference: Identifiability and finite sample results. *Advances in Neural Information Processing Systems*, *33*, 14772–14782.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, *19*(1), 15–18.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall/CRC.

Ellison, A. M. (2022). Model selection for generalized additive models using bic. *Demographic Research*, *47*, 561–594.

Erosheva, E. A., Grant, S., Chen, M.-C., Lindner, M. D., Nakamura, R. K., & Lee, C. J. (2020a). NIH peer review: Criterion scores completely account for racial disparities in overall impact scores. *Science Advances*, *6*(23), eaaz4868.

Erosheva, E. A., Grant, S., Chen, M.-C., Lindner, M. D., Nakamura, R. K., & Lee, C. J. (2020b). NIH-public-data_Erosheva-et-al.csv [Dataset associated with Erosheva et al. (2020), *Science Advances*].

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874.

Federal Office for the Environment (FOEN). (2009). Air: Data, Indicators and Maps [Accessed: 2025-10-25].

Forscher, P. S., Cox, W. T., Brauer, M., & Devine, P. G. (2019). Little race or gender bias in an experiment of initial review of NIH R01 grant proposals. *Nature human behaviour*, *3*(3), 257–264.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization in Bayesian Workflow. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *182*(2), 389–402.

Ginther, D. K., Basner, J., Jensen, U., Schnell, J., Kington, R., & Schaffer, W. T. (2018). Publications as predictors of racial and ethnic differences in NIH research awards. *PloS one*, *13*(11), e0205929.

Ginther, D. K., Schaffer, W. T., Schnell, J., Masimore, B., Liu, F., Haak, L. L., & Kington, R. (2011). Race, ethnicity, and NIH research awards. *Science*, *333*(6045), 1015–1019.

Glymour, C., Zhang, K., & Spirtes, P. (2019). Review of Causal Discovery Methods Based on Graphical Models. *Frontiers in Genetics*, *10*.

Goldberg, P. (1968). Are women prejudiced against women? *Trans-action*, *5*(5), 28–30.

Good, P. I. (2005). *Permutation, parametric, and bootstrap tests of hypotheses* (3rd). Springer.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., & Smola, A. J. (2008). A Kernel Statistical Test of Independence. In J. C. Platt, D. Koller, Y. Singer, & S. T. Roweis (Eds.), *Advances in Neural Information Processing Systems 20* (pp. 585–592). Curran Associates, Inc.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, *143*(1), 29–36.

Hill, J., & Stuart, E. A. (2015). Causal inference: Overview. *International Encyclopedia of the Social & Behavioral Sciences: Second Edition*, 255–260.

Hoppe, T. A., Litovitz, A., Willis, K. A., Meseroll, R. A., Perkins, M. J., Hutchins, B. I., Davis, A. F., Lauer, M. S., Valantine, H. A., Anderson, J. M., et al. (2019). Topic choice contributes to the lower rate of NIH awards to African-American/black scientists. *Science advances*, *5*(10), eaaw7238.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., & Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in Neural Information Processing Systems 21* (pp. 689–696). Curran Associates, Inc.

Hu, L. (2023). What is "race" in algorithmic discrimination on the basis of race? *Journal of Moral Philosophy*, *1*(aop), 1–26.

Hu, P., Jiao, R., Jin, L., & Xiong, M. (2018). Application of causal inference to genomic analysis: Advances in methodology. *Frontiers in Genetics*, *9*, 238.

Hudgens, M. G., & Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the american statistical association*, *103*(482), 832–842.

Ikeuchi, T., Ide, M., Zeng, Y., Maeda, T. N., & Shimizu, S. (2023). Python package for causal discovery based on lingam. *Journal of Machine Learning Research*, *24*(14), 1–8.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press.

Jackson, J. W., & VanderWeele, T. J. (2018). Decomposition analysis to identify intervention targets for reducing disparities. *Epidemiology (Cambridge, Mass.)*, *29*(6), 825.

Janitza, S., Binder, H., & Boulesteix, A.-L. (2016). Pitfalls of hypothesis tests and model selection on bootstrap samples: Causes and consequences in biometrical applications. *Biometrical Journal*, *58*(3), 447–473.

Jiang, Z., Chen, C., Xu, Z., Wang, X., Zhang, M., & Zhang, D. (2023). Signet: Transcriptome-wide causal inference for gene regulatory networks. *Scientific Reports*, *13*(1), 19371.

Jiao, R., Lin, N., Hu, Z., Bennett, D. A., Jin, L., & Xiong, M. (2018). Bivariate causal discovery and its applications to gene expression and imaging data analysis. *Frontiers in genetics*, *9*, 347.

Kalainathan, D., Goudet, O., & Dutta, R. (2020). Causal discovery toolbox: Uncover causal relationships in python [arXiv:1903.02278]. *Journal of Machine Learning Research*, *21*(37), 1–5.

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., & Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, *47*(11), 1–26.

Kocaoglu, M., Shakkottai, S., Dimakis, A. G., Caramanis, C., & Vishwanath, S. (2020). Applications of common entropy for causal inference. *Advances in neural information processing systems*, *33*, 17514–17525.

Komatsu, Y., Shimizu, S., & Shimodaira, H. (2010a). Assessing statistical reliability of LiNGAM via multiscale bootstrap. *International Conference on Artificial Neural Networks*, 309–314.

Komatsu, Y., Shimizu, S., & Shimodaira, H. (2010b). Computing p-values of lingam outputs via multiscale bootstrap [preprint arXiv 0909.2904]. *Proceedings of the 20th International Conference on Artificial Neural Networks (ICANN 2010)*.

Kotoku, J., Oyama, A., Kitazumi, K., Toki, H., Haga, A., Yamamoto, R., Shinzawa, M., Yamakawa, M., Fukui, S., Yamamoto, K., et al. (2020). Causal relations of health indices inferred statistically using the directlingam algorithm from big data of osaka prefecture health checkups. *Plos one*, *15*(12), e0243229.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). *Counterfactual fairness*.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). McGraw-Hill/Irwin.

Lauer, M., & Roychowdhury, D. (2023). Analyses of Demographic-Specific Funding Rates for Type 1 Research Project Grant and R01-Equivalent Applications [[Accessed 01-13-2024]].

Lauer, M., & Roychowdhury, D. (2024, July). Analyses of Demographic-Specific Funding Rates for Type 1 Research Project Grant and R01-Equivalent Applications [Accessed: 2026-04-23].

Lauer, M. S., Doyle, J., Wang, J., & Roychowdhury, D. (2021). Associations of topic-specific peer review outcomes and institute and center award rates with funding disparities at the national institutes of health. *Elife*, *10*, e67173.

Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses* (3rd). Springer.

Marx, A., & Vreeken, J. (2017). Telling cause from effect using mdl-based local and global regression. *2017 IEEE international conference on data mining (ICDM)*, 307–316.

Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., & Schölkopf, B. (2016a). Distinguishing Cause from Effect Using Observational Data: Methods and Benchmarks. *Journal of Machine Learning Research*, *17*(32), 1–102.

Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., & Schölkopf, B. (2016b). Distinguishing cause from effect using observational data: Methods and benchmarks. *The Journal of Machine Learning Research*, *17*(1), 1103–1204.

Motokawa, T., Watanabe, T., Moriyama, K., Takada, Y., Kitamura, A., Kato, T., Kokubo, Y., Miyamoto, Y., & Okamura, T. (2020). Causal relationships among health checkup variables with the use of lingam. *PLOS ONE*, *15*(12), e0243432.

Nabi, R., & Shpitser, I. (2018). Fair inference on outcomes. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1).

Nakamura, R. K., Mann, L. S., Lindner, M. D., Braithwaite, J., Chen, M.-C., Vancea, A., Byrnes, N., Durrant, V., & Reed, B. (2021). An experimental test of the effects of redacting grant applicant identifiers on peer review outcomes. *Elife*, *10*, e71368.

National Institute of Biomedical Imaging and Bioengineering. (2026). Nih research project grant program (r01) [Accessed: 2026-04-21].

National Institutes of Health. (2026). Grants & funding [Accessed: 2026-04-21].

National Institutes of Health. (n.d.). Research Project Grant (R01) [Accessed: 2026-03-20].

NIH Center for Scientific Review. (2016). Definitions of Criteria and Considerations for Research Project Grant (RPG/R01/R03/R15/R21/R34) Critiques [[Accessed 01-27-2024]].

NIH Center for Scientific Review. (2009, January). *Enhancing peer review: The nih announces enhanced review criteria for evaluation of research applications received for potential fy2010 funding and beyond* [Notice Number: NOT-OD-09-025]. National Institutes of Health. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-09-025.html

NIH Center for Scientific Review. (2024a). Frequently asked questions: Simplifying peer review [Accessed 15-Aug-2024; page current as of 2024].

NIH Center for Scientific Review. (2023a). Peer review [[Accessed 2023-10-19]].

NIH Center for Scientific Review. (2018). Scientific peer review regional meeting 2018 [Accessed: 2024-09-29].

NIH Center for Scientific Review. (2023b). Simplified peer review framework [Announced October 19, 2023; page updated August 19, 2024; accessed 21-Sept-2024].

NIH Center for Scientific Review. (2024b). Simplified scoring system [Accessed: September 21, 2024].

NIH NIAD. (2023, May). First-Level Peer Review — NIAID: National Institute of Allergy and Infectious Diseases.

Ogburn, E. L., & VanderWeele, T. J. (2014). Causal diagrams for interference. *Statistical Science*, *29*(4), 559–578.

Park, S., Qin, X., & Lee, C. (2020). Estimation and sensitivity analysis for causal decomposition in health disparity research. *Sociological Methods & Research*, 00491241211067516.

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms*. MIT Press.

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

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63.

Rosenström, T., Jokela, M., Puttonen, S., Hintsanen, M., Pulkki-Råback, L., Viikari, J. S., Raitakari, O. T., & Keltikangas-Järvinen, L. (2012). Pairwise measures of causal direction in the epidemiology of sleep problems and depression. *PloS one*, *7*(11), e50841.

Rubin, D. B. (1981). The bayesian bootstrap. *The annals of statistics*, 130–134.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688.

Saito, T., & Fujiwara, K. (2023). Causal analysis of nitrogen oxides emissions process in coal-fired power plant with lingam. *Frontiers in Analytical Science*, *3*, 1045324.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Sen, A., & Sen, B. (2014). Testing independence and goodness-of-fit in linear models [Publisher: Oxford Academic]. *Biometrika*, *101*(4), 927–942.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., & Kerminen, A. (2006). A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, *7*(Oct), 2003–2030.

Shimizu, S., & Kano, Y. (2008). Use of non-normality in structural equation modeling: Application to direction of causation. *Journal of Statistical Planning and Inference*, *138*(11), 3483–3491

Special Issue in Honor of Junjiro Ogawa (1915 - 2000): Design of Experiments, Multivariate Analysis and Statistical Inference.

Shimizu, S., Inazumi, T., Sogawa, Y., Hyvarinen, A., Kawahara, Y., Washio, T., Hoyer, P. O., Bollen, K., & Hoyer, P. (2011). Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research-JMLR*, *12*(Apr), 1225–1248.

Song, J., Oyama, S., & Kurihara, M. (2017). Tell cause from effect: Models and evaluation. *International Journal of Data Science and Analytics*, *4*, 99–112.

Spirtes, P., Glymour, C., Scheines, R., Kauffman, S., Aimale, V., & Wimberly, F. (2000). Constructing bayesian network models of gene expression networks from microarray data.

Thamvitayakul, K., Shimizu, S., Ueno, T., Washio, T., & Tashiro, T. (2012a). Bootstrap confidence intervals in directlingam. *2012 IEEE 12th International Conference on Data Mining Workshops (ICDMW)*.

Thamvitayakul, K., Shimizu, S., Ueno, T., Washio, T., & Tashiro, T. (2012b). Bootstrap confidence intervals in DirectLiNGAM. *2012 IEEE 12th International Conference on Data Mining Workshops*, 659–668.

van der Vaart, A. W. (1998). *Asymptotic statistics* [Continuous Mapping Theorem, Section 2.3]. Cambridge University Press.

VanderWeele, T. J., & Robinson, W. R. (2014). On causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology (Cambridge, Mass.)*, *25*(4), 473.

Wang, Y. S., Kolar, M., & Drton, M. (2025). Confidence sets for causal orderings. *Journal of the American Statistical Association*, (just-accepted), 1–25.

Warner, E. T., Carapinha, R., Weber, G. M., Hill, E. V., & Reede, J. Y. (2017). Gender Differences in Receipt of National Institutes of Health R01 Grants Among Junior Faculty at an Academic Medical Center: The Role of Connectivity, Rank, and Research Productivity. *Journal of Women's Health*, *26*(10), 1086–1093.

Wasserman, L. (2010). *All of statistics: A concise course in statistical inference*. Springer Publishing Company, Incorporated.

Wasserstein, R. L., & Lazar, N. A. (2016). The asa's statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p ¡ 0.05". *The American Statistician*, *73*(sup1), 1–19.

Weisberg, S. (2005). *Applied linear regression* (3rd ed.). Wiley.

Wood, S. N. (2017). *Generalized additive models: An introduction with r* (2nd ed.). Chapman; Hall/CRC.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *65*(1), 95–114.

Zhang, K., & Chan, L.-W. (2006). Extensions of ica for causality discovery in the hong kong stock market. *International Conference on Neural Information Processing*, 400–409.

Zhang, K., & Hyvarinen, A. (2012). On the identifiability of the post-nonlinear causal model. *arXiv preprint arXiv:1205.2599*.

Zhang, X., Zhao, X.-M., He, K., Lu, L., Cao, Y., Liu, J., Hao, J.-K., Liu, Z.-P., & Chen, L. (2012). Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics*, *28*(1), 98–104.

Zheng, Y., Huang, B., Chen, W., Ramsey, J., Gong, M., Cai, R., Shimizu, S., Spirtes, P., & Zhang, K. (2024). Causal-learn: Causal discovery in python. *Journal of Machine Learning Research*, *25*(60), 1–8.

# Appendix A

# APPENDIX TO CHAPTER 3

The appendix for this chapter is organized as follows. In Section A.1, we present an exploratory analysis of the distributions of preliminary scores, proposal characteristics, PI characteristics, structural factors, and selection into discussion, focusing on how these distributions differ between Black and white PIs. In Section A.2, we analyze whether selection into panel discussion is a fully deterministic process with respect to Preliminary Overall Impact Scores. In Section A.3, we further illustrate the concept of partial interference. In Section A.4, we present our simulation setup and results, which assess the statistical bias of the shifted estimator and evaluate the impact of interference by comparing estimates with and without accounting for it. In Section A.5, we provide identification proofs for the Disparity Reduction, Disparity Remaining, and Thresholded Disparity estimands under no selection bias. In Section A.6, we discuss the effects of self-selection on the identification of these estimands. Finally, in Section A.7, we present the full Thresholded Disparity results for all binarization thresholds.

## *A.1 Exploratory Data Analysis*

In what follows, we present an exploratory analysis of the distributions of preliminary scores, proposal characteristics, PI characteristics, structural factors, and selection into discussion – focusing on how these distributions differ between Black and white PIs.

**Summary Statistics for Preliminary Overall Impact scores** Table A.1 presents summary statistics related to Preliminary Overall Impact Scores and Investigator and Environment criterion scores.

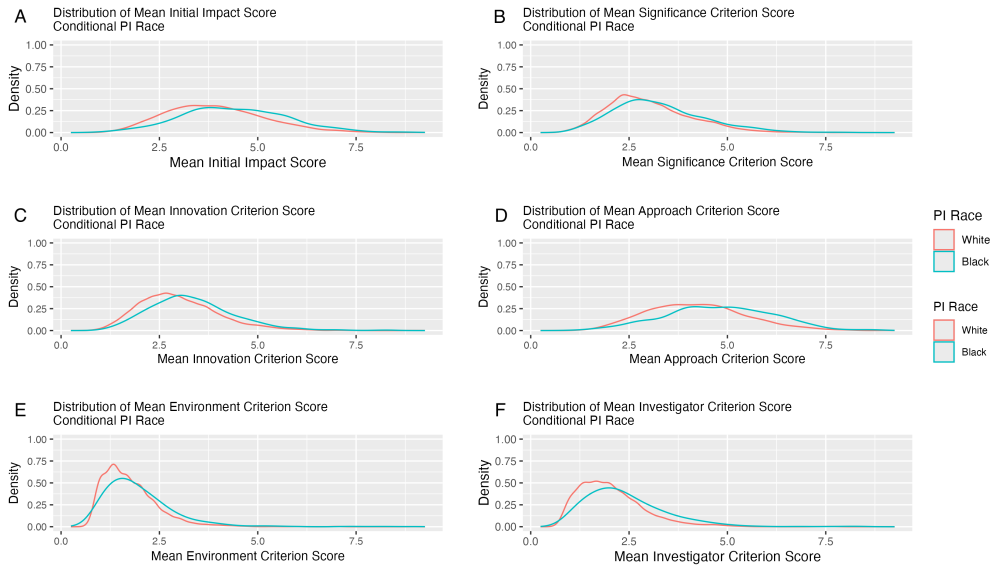



Figure A.1: Distributions of Mean Preliminary Overall Impact Scores Conditioned on PI race.

**Summary Statistics for Proposal and PI Characteristics** Table A.2 presents summary statistics for proposal and PI characteristics.

**Distribution of Preliminary Overall Impact scores** Figure A.1 shows the empirical distributions of Preliminary Overall Impact Scores conditioned on PI race.

**Distribution of Structural Characteristics** Figure A.2 shows the empirical distributions of IRG and Administering Organization conditioned on PI race.

**Summary Statistics for Discussion and PI race** Table A.3 presents summary statistics for discussion and PI race.

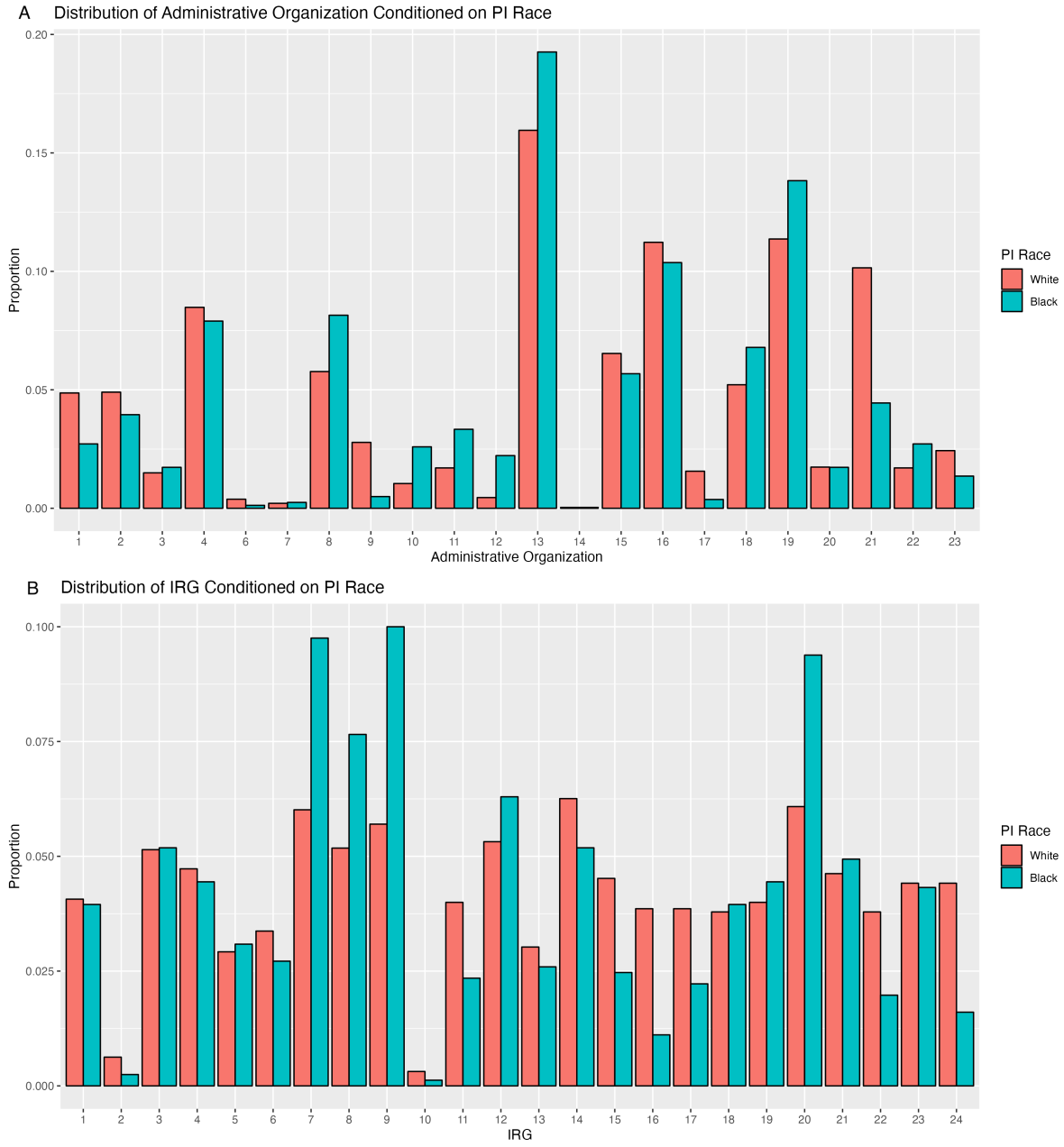



Figure A.2: Distributions of IRG and Administering Organization Conditioned on PI race.

| Criterion Score | % White PI | % Black PI | % Discuss | % Discuss for White PIs | % Discuss for Black PIs | % Disparity in Discussion |
|---|---|---|---|---|---|---|
| **Preliminary Overall Impact scores** | | | | | | |
| 1 | 0.7% | 0.9% | 96.4% | 95.2% | 100% | 4.8% |
| 2 | 11.2% | 5.9% | 97% | 97.2% | 95.8% | 1.4% |
| 3 | 28.3% | 19% | 93.8% | 94% | 92.9% | 1.1% |
| 4 | 29.4% | 28% | 53.6% | 52.9% | 56.4% | -3.5% |
| 5 | 19.2% | 24.6% | 13.7% | 13.1% | 15.6% | -2.5% |
| 6 | 8.4% | 14.9% | 0.6% | 0.8% | 0% | 0.8% |
| 7 | 2.3% | 5.3% | 0% | 0% | 0% | 0% |
| 8 | 0.5% | 1.0% | 0% | 0% | 0% | 0% |
| 9 | 0.07% | 0.4% | 20% | 0% | 33.3% | -33.3% |
| **Investigator Criterion Scores** | | | | | | |
| 1 | 29.1% | 17.7% | 79.2% | 79.8% | 75.5% | 4.3% |
| 2 | 46.8% | 44.3% | 55.3% | 56.8% | 49.6% | 7.3% |
| 3 | 18.3% | 24.9% | 31.3% | 32.5% | 28.2% | 4.3% |
| 4 | 4.1% | 9.4% | 12.9% | 11.9% | 14.5% | -2.6% |
| 5 | 1.5% | 2.6% | 4.8% | 4.8% | 4.8% | 0% |
| 6 | 0.07% | 0.4% | 0% | 0% | 0% | 0% |
| 7 | 0.07% | 0.2% | 0% | 0% | 0% | 0% |
| 8 | 0.1% | 0.5% | 14.3% | 0% | 25% | -25% |
| 9 | 0.03% | 0% | 0% | 0% | 0% | 0% |
| **Environment Criterion Scores** | | | | | | |
| 1 | 43.4% | 32.8% | 71.3% | 72.7% | 64.7% | 8% |
| 2 | 44.6% | 47.5% | 48.2% | 50.2% | 41.6% | 8.6% |
| 3 | 9.5% | 14.1% | 21.4% | 23% | 17.5% | 5.4% |
| 4 | 1.9% | 3.8% | 9.3% | 10.9% | 6.5% | 4.5% |
| 5 | 0.3% | 0.9% | 0% | 0% | 0% | 0% |
| 6 | 0.1% | 0.2% | 20% | 0% | 50% | -50% |
| 7 | 0.1% | 0.2% | 0% | 0% | 0% | 0% |
| 8 | 0% | 0.2% | 50% | 0% | 50% | -50% |
| 9 | 0.03% | 0.1% | 0% | 0% | 0% | 0% |

Table A.1: Summary Statistics for Investigator and Environment Criterion Scores.

| Quantity | % | % Black PI | % White PI | % Difference | % Disparity in Discussion |
|---|---|---|---|---|---|
| **Degree Category** | | | | | |
| PhD | 71.8% | 65.7% | 73.6% | **7.9%** (3.6%, 13.2%) | **14.4%** (8.2%, 20.2%) |
| MD | 16.5% | 19.9% | 15.6% | -4.3% (-8.5%, 0.1%) | 7.0% (-2.2%, 16.2%) |
| MD/Phd | 9.5% | 9.6% | 9.5% | -0.2% (-3.2%, 2.6%) | **9.9%** (2.6%, 23.4%) |
| Others | 2.2% | 4.8% | 1.4% | **-3.4%** (-5.8%, -1.5%) | 5.6% (-13.0%, 26.9%) |
| **Institution Funding Bin** | | | | | |
| 1 | 19.9% | 19.4% | 20.1% | 0.7% (-3.5%, 5.1%) | **18.7%** (9.8%, 27.8%) |
| 2 | 22.6% | 19.9% | 23.4% | 3.6% (-1.0%, 7.8%) | 2.1% (-7.1%, 11.6%) |
| 3 | 21.1% | 21.4% | 21.0% | -0.4% (-4.6%, 4.2%) | **12.8%** (2.9%, 22.3%) |
| 4 | 18.6% | 18.1% | 18.7% | 0.5% (-3.7%, 4.7%) | **11.5%** (1.5%, 20.8%) |
| 5 | 17.8% | 21.2% | 16.9% | -4.4% (-8.8%, 0.4%) | **13.7%** (4.3%, 22.3%) |
| **Career Stage** | | | | | |
| ESI | 20.1% | 27.3% | 18.1% | **-9.17%** (-13.8%, -4.5%) | **12.0%** (4.1%, 19.9%) |
| Non-ESI | 15.3% | 20.4% | 13.9% | **-6.5%** (-10.6%, -2.1%) | **22.0%** (13.1%, 30.6%) |
| Experienced | 64.6% | 52.3% | 68.0% | **15.7%** (10.4%,21.2%) | **9.8%** (2.8%,16.2%) |
| **Application Type** | | | | | |
| New | 88.4% | 92.4% | 87.2% | **-5.1%** (-8.0%,-1.8%) | **12.0%** (7.0%,17.5%) |
| Renewal | 11.6% | 7.7% | 12.8% | **5.1%** (2.0%,8.2%) | 8.4% (-4.8%,22.8%) |
| **Application Amended** | | | | | |
| True | 26.7% | 25.1% | 27.2% | 2.1% (-6.9%,2.7%) | **9.5%** (1.0%,17.7%) |
| False | 73.3% | 74.9% | 72.8% | -2.1% (-2.7%, 7.0%) | **12.7%** (7.0%, 18.2%) |

Table A.2: Summary Statistics for Proposal and PI Characteristics. Substantial differences and disparities are bolded.

| Quantity | % | % Black PI | % White PI | % Difference | % Accepted | % Rejected | % Difference |
|---|---|---|---|---|---|---|---|
| **PI race** | | | | | | | |
| Black | 22.0% | - | - | - | 18.0% | 26.5% | -8.5% |
| White | 78.0% | - | - | - | 82.0% | 73.5% | 8.5% |
| **Discussion** | | | | | | | |
| Accepted | 53.6% | 44.0% | 56.3% | -12.3% | - | - | - |
| Rejected | 46.4% | 56.0% | 43.7% | 12.3% | - | - | - |

Table A.3: Summary Statistics for Discussion and PI race.

### *A.2 Selection into Discussion and Preliminary Overall Impact Scores*

The analyses in this section indicate that the selection of proposals for panel discussion is not a fully deterministic process with respect to Preliminary Overall Impact Scores. Although our data lack indicators for which applications within each SRG were reviewed together in the same panel meeting, we can still support this claim through two key observations. First, we fit a logistic regression model for selection into discussion (Y), including PI race, ethnicity, gender, average Preliminary Overall Impact Scores, and other applicant, proposal, and structural attributes. Second, as noted in our preliminary analysis (Section 3.2), we observe substantial variability in discussion rates and disparities across IRGs and SRGs. Together, these analyses suggest that while these Preliminary Overall Impact scores play a central role, there may be other additional factors that influence which applications are ultimately selected for discussion.

We model selection into discussion ($Y$) using logistic regression with PI race, ethnicity, gender, average Preliminary Overall Impact Scores, and other covariates as fixed effects, and SRG as a random effect. We compute 95% credible intervals for each coefficient. The analysis reveals that Preliminary Overall Impact Scores have a substantial effect, along with career stage and some individual IRGs (estimates reported in Table A.4). This suggests additional contributors to the selection process beyond Preliminary Overall Impact Scores. While the regression analysis suggests that other variables besides Preliminary Overall Impact Score are associated with selection into discussion, it does not allow us to assess their causal contribution to observed Black-white disparities in selection into discussion. To address this, we apply a causal decomposition framework for observational data, extending the approach introduced by VanderWeele and Robinson (2014).

Additionally, our preliminary analysis in Section 3.2 demonstrated substantial variability in discussion rates and disparities across IRGs and SRGs. This variability provides further evidence that, although Preliminary Overall Impact Scores are the primary driver of selection into discussion, the process does not yield uniform discussion rates across SRGs – suggesting

| Variable | Estimate | 95% Credible Interval |
|---|---|---|
| Intercept | 12.30 | (10.91, 13.71) |
| Average Preliminary Overall Impact Score | -3.16 | (-3.39, -2.94) |
| Hispanic/Latino | 0.63 | (0.04, 1.24) |
| IRG 9 | -1.29 | (-2.45, -0.13) |
| IRG 21 | -1.31 | (-2.55, -0.08) |
| ESI | 1.84 | (1.52, 2.17) |
| Non-ESI | 1.76 | (1.41, 2.13) |

Table A.4: We fit a logistic regression model for selection into discussion as the outcome, with PI race, ethnicity, gender, average Preliminary Overall Impact Scores, and other covariates as fixed effects, and SRG included as a random effect. For brevity, we report only those estimates whose 95% credible intervals do not include zero.

some flexibility at the level of SRGs rather than strict, score-based determination.

Overall, our analysis suggests that selection into panel discussion is primarily driven by Preliminary Overall Impact Scores but that selection is not entirely deterministic.

### *A.3 Discussion of Partial Interference Assumption*

In Section 3.4, we summarize the causal decomposition framework applied to peer review. We introduce our estimands of interest: the Observed Disparity, Disparity Reduction, Disparity Remaining, and Thresholded Disparity. To identify these estimands, we account for the interference introduced by the streamlining procedure. Because proposals are reviewed in parallel and relative to one another within study sections, interventions on one proposal may influence the outcomes of other proposals within the same study section, introducing interference. However, it is reasonable to assume partial interference, meaning that interference occurs only within each SRG (Hudgens & Halloran, 2008).

To illustrate the concept of partial interference, Ogburn and VanderWeele describe a scenario involving two individuals living in the same household, each randomized to an intervention aimed at lowering cholesterol. The intervention includes cooking classes, nutritional counseling, and coupons for fresh produce. If only one household member receives the intervention, they may still bring fresh produce into the home, prepare healthier meals, or share

advice from the nutritionist—indirectly improving the untreated individual's diet. The untreated individual may still experience improvements in blood cholesterol, not through direct intervention, but because exposure to the treated household environment can influence dietary habits and affect cholesterol levels. As this interference occurs only within households, it constitutes an instance of partial interference.

We encounter a parallel situation in NIH peer review. Under the streamlining procedure, changes to the score of one proposal can influence which other proposals in the same study section are selected for discussion, resulting in interference. As in the cholesterol example, this interference is localized (i.e., partial interference): just as the household limits who is affected, in our setting, interference occurs only within SRGs. This allows us to formalize our framework using an approach similar to that proposed by Ogburn and VanderWeele (Ogburn & VanderWeele, 2014).

We represent partial interference using the simplified structure depicted in the DAG in Figure 3.6, which closely follows the direct interference example shown in Figure 4 of (Ogburn & VanderWeele, 2014). This DAG illustrates a basic setting involving only two proposals, $i$ and $j$, within a single SRG $k$. While no SRG in practice evaluates only two proposals, this simplified setting is sufficient to illustrate the type of interference under consideration and the variables that must be adjusted for to control confounding. To simplify the representation, all variables that confound the relationships between $M_i$ and $Y_i$ are grouped into a single node $C_i \equiv (W_i, A_i, Z_i, L_i)$. The DAG implies that adjusting for $C_i$ suffices to block all backdoor paths from both $M_i$ and $M_j$ to $Y_i$, and thus $Y_i(m_i, m_j) \perp\!\!\!\perp (M_i, M_j) \mid C_i$. This permits identification of the causal effect of the score vector $\mathbf{M}$ on $Y_i$ under partial interference.

### *A.4 Simulation Study*

The goal of our simulation study is to assess the statistical bias of the shifted estimator under the true data-generating process and to evaluate the impact of interference. To do this, we simulate data under a set of simplifying assumptions.

To simplify the hierarchical structure of the NIH review process, we assume that SRGs,

IRGs, and Administering Organizations are fully nested (i.e., $SRG \rightarrow IRG \rightarrow Admin$). Each SRG is treated as an independent unit and corresponds to a single study section meeting, with 100 proposals per SRG—consistent with typical NIH study section sizes, which range from 60 to 100 proposals (Bonetta, 2009).

For each SRG, we simulate proposal-level data based on a structural equation model (SEM) that encodes the assumptions of our causal diagrams. Specifically, we first sample PI ethnicity and gender ($W$) and PI race ($A$) independently with replacement from their empirical marginal distributions for the matched SRG. Then, conditional on $W$ and $A$, we draw additional application-level attributes—Application Type, Amended Status, Career Stage, Degree Category, and Funding Bin (jointly denoted as $Z$)—from their empirical conditional distribution. Preliminary criterion scores are then generated using the SEM based on these covariates. This framework assumes that proposals are independent across SRGs. Within each SRG, proposals are generated independently conditional on covariates, but their outcomes become dependent through the ranking procedure that determines selection into discussion.

We estimate SEM parameters using the matched NIH dataset, which exhibits fewer extreme class imbalances. This improved balance enables more robust estimation and reduces sensitivity to model misspecification (Erosheva et al., 2020a).

This setup allows us to ground our simulations in real NIH data while maintaining a fully known data-generating process. It allows us to assess the statistical bias of our estimates, including bias introduced by interference, under specific interventions, while keeping the simulated data aligned with the structure of the observed data. For simplicity, we assume that selection into discussion is deterministic: within each SRG, proposals are ranked by Preliminary Overall Impact Scores, and the top 50% are selected for discussion.

We present our simulations for the Thresholded Disparity estimand. Because the shifted estimator operates analogously across all estimands, evaluating its performance for the Thresholded Disparity provides insight into its behavior more generally. Through our simulations, we aim to quantify the overall bias in our estimates, which includes the bias arising

from interference. To do so, we must first define the *truth*. In our simulation setting, the truth is defined as the disparity in selection into discussion under the intervention that jointly binarizes Investigator and Environment scores and applies the deterministic ranking rule within each SRG.

We use linear mixed models with SRG as a random effect to learn the structural equation model parameters from the NIH matched dataset:

- $M_e|A, W, L, Z$
- $M_i|A, W, L, Z$
- $M_s|A, W, L, Z, M_e, M_i$
- $M_{inn}|A, W, L, Z, M_e, M_i$
- $M|A, W, L, Z, M_e, M_i, M_s, M_{inn}$
- $M_a|A, W, L, Z, M_e, M_i, M$

Using these models, we simulate data following Algorithm 4. For each Preliminary Overall Impact Score, we introduce noise based on within-SRG variability estimated from the linear mixed models. We then compute the true Thresholded Disparity using the known data-generating process (Algorithm 5).

We use Monte Carlo to generate $D = 100$ datasets from the same data-generating process and calculate the bias under our shifted estimator, with and without accounting for interference. We set the Investigator and Environment criterion score thresholds at 2.

Table A.5 reports the average estimator bias in the Thresholded Disparity, both with and without accounting for interference. In both cases, the bias is small, and the shifted estimator is approximately unbiased when interference is accounted for. Overall, the simulation results suggest that, under the specified data-generating process, accounting for interference has a limited impact on the estimator's bias.

**Algorithm 4** Simulating Data

**for** admin org **in the** Admin Orgs **do**
  **for** irg **in** admin org **do**
    **for** srg **in** srg **do**

1. Compute $P(W|SRG = srg)$ and $P(A|SRG = srg)$ using NIH matched data
2. Simulate 100 proposals for this SRG:
   - $SRG = srg, IRG = irg$, Admin Org = admin org
   - Sample $W$ with replacement using $P(W|SRG = srg)$
   - Sample $A$ with replacement using $P(A|SRG = srg)$
   - Sample $Z$ with replacement using $P(Z|SRG = srg, A = a, W = w)$
   - Using linear mixed models, predict:
     - $M_{\text{environment}}$ adding noise $\eta_{m_e} \sim N(0, \sigma^2_{m_e})$, where $\sigma^2_{m_e}$ is the estimated within-SRG variability
     - $M_{\text{investigator}}$ adding noise $\eta_{m_i} \sim N(0, \sigma^2_{m_i})$, where $\sigma^2_{m_i}$ is the estimated within-SRG variability
     - $M_{\text{significance}}$ adding noise $\eta_{m_s} \sim N(0, \sigma^2_{m_s})$, where $\sigma^2_{m_s}$ is the estimated within-SRG variability
     - $M_{\text{innovation}}$ adding noise $\eta_{m_{inn}} \sim N(0, \sigma^2_{m_{inn}})$, where $\sigma^2_{m_{inn}}$ is the estimated within-SRG variability
     - $M$ adding noise $\eta_m \sim N(0, \sigma^2_m)$, where $\sigma^2_m$ is the estimated within-SRG variability
     - $M_{\text{approach}}$ adding noise $\eta_{m_a} \sim N(0, \sigma^2_{m_a})$, where $\sigma^2_{m_a}$ is the estimated within-SRG variability
   - Rank proposals based on $M$ and set selection into discussion $Y$ as 1 for the top 50% of proposals and 0 otherwise

    **end for**
  **end for**
**end for**

| **Estimator** | **Mean Bias (%)** | **2.5%** | **97.5%** |
|---|---|---|---|
| Shifted (Adjusted for Interference) | 0.0 | -3.4 | 3.3 |
| Shifted (Not Adjusted for Interference) | -0.3 | -3.6 | 2.9 |

Table A.5: Estimator bias based on simulations. We use Monte Carlo to generate $D$ = 100 datasets from the same data-generating process and calculate the mean bias under the shifted estimator, with and without accounting for interference. We also report 95% credible intervals. Investigator and Environment criterion score thresholds are set to 2.

**Algorithm 5** Using simulation to get true thresholded disparity.

**for** admin org **in** Admin Orgs **do**
  **for** irg **in** admin org **do**
    **for** srg **in** srg **do**

1. Get the simulated data for $SRG = srg$
2. For every proposal in this SRG, threshold $M_{i,e}$ using $M^t_{i,e} = \begin{cases} 1, M^{new}_{i,e} \leq \begin{bmatrix} 2 \\ 2 \end{bmatrix} \\ 5, M^{new}_{i,e} > \begin{bmatrix} 2 \\ 2 \end{bmatrix} \end{cases}$
3. Using linear mixed models, obtain:
   - $M_s(M^t_{i,e})|A, W, L, Z, M^t_{i,e}$, adding noise $\eta_{m_s} \sim N(0, \sigma^2_{m_s})$, where $\sigma^2_{m_s}$ is the estimated within-SRG variability
   - $M_{inn}(M^t_{i,e})|A, W, L, Z, M^t_{i,e}$, adding noise $\eta_{m_{inn}} \sim N(0, \sigma^2_{m_{inn}})$, where $\sigma^2_{m_{inn}}$ is the estimated within-SRG variability
   - $M(M^t_{i,e})|A, W, L, Z, M^t_{i,e}, M_s(M^t_{i,e}), M_{inn}(M^t_e)$, adding noise $\eta_m \sim N(0, \sigma^2_m)$, where $\sigma^2_m$ is the estimated within-SRG variability
4. Rank proposals based on $M(M^t_{i,e})$ and set selection into discussion after the intervention, $Y(M^t_{i,e})$, as 1 for the top 50% of proposals and 0 otherwise

    **end for**
  **end for**
**end for**

Use empirical mean to estimate $E[Y(M^t_e) = 1|A, M_e = m^t_e, Z, W, L]$ and estimate Thresholded Disparity

Table A.6: Summary of the interventions of interest, corresponding estimands, identification assumptions, and identification forms for the interventions considered in Section 3.6.1. In addition to the notation introduced in Section 3.4, we denote PI attributes (career stage, degree category, and institution funding bin, a proxy for institution prestige) by $Z_{pi}$ and proposal attributes (application type and amended status) by $Z_{prop}$.

| Intervention | Estimands of Interest | Identification Assumptions | Identification Form of Counterfactual Quantity (e.g. $E[\mathbf{Y}(\mathbf{M}^*)\|\mathbf{A}=1]$) |
|---|---|---|---|
| Preliminary Impact Scores | $E[\mathbf{Y}(\mathbf{M}^*)\|\mathbf{A}=1] - E[\mathbf{Y}\|\mathbf{A}=1]$<br>$E[\mathbf{Y}\|\mathbf{A}=0] - E[\mathbf{Y}(\mathbf{M}^*)\|\mathbf{A}=1]$ | $Y_{ki}(\mathbf{m}_k) \perp \mathbf{M}_k \mid \mathbf{W}_k, \mathbf{A}_k, \mathbf{Z}_k, \mathbf{L}_k$<br>$P(\mathbf{A}_k \mid \mathbf{W}_k, \mathbf{L}_k) > 0$ a.e.<br>$P(\mathbf{M}_k \mid \mathbf{A}_k, \mathbf{W}_k, \mathbf{L}_k, \mathbf{Z}_k) > 0$ a.e. | $\frac{1}{M}\sum_{k=1}^{N}\sum_{i=1}^{n_k}\sum_{\mathbf{m}_k,\mathbf{z}_k,\mathbf{w}_k,\mathbf{l}_k} P(\mathbf{w}_k,\mathbf{l}_k \mid A_{ki}=1)$<br>$\times E[Y_{ki} \mid A_{ki}=1, \mathbf{m}_k, \mathbf{z}_k, \mathbf{w}_k, \mathbf{l}_k]$<br>$\times P(\mathbf{z}_k \mid A_{ki}=1, \mathbf{w}_k, \mathbf{l}_k)$<br>$\times P(\mathbf{m}_k \mid A_{ki}=0, \mathbf{w}_k, \mathbf{l}_k)$ |
| PI attributes | $E[\mathbf{Y}(\mathbf{Z}^*_{pi})\|\mathbf{A}=1] - E[\mathbf{Y}\|\mathbf{A}=1]$<br>$E[\mathbf{Y}\|\mathbf{A}=0] - E[\mathbf{Y}(\mathbf{Z}^*_{pi})\|\mathbf{A}=1]$ | $Y_{ki}(\mathbf{z}_{pi,k}) \perp \mathbf{Z}_{pi,k} \mid \mathbf{W}_k, \mathbf{A}_k, \mathbf{Z}_{prop,k}, \mathbf{L}_k$<br>$P(\mathbf{A}_k \mid \mathbf{W}_k, \mathbf{L}_k) > 0$ a.e.<br>$P(\mathbf{Z}_{pi,k} \mid \mathbf{A}_k, \mathbf{W}_k, \mathbf{L}_k, \mathbf{Z}_{prop,k}) > 0$ a.e. | $\frac{1}{M}\sum_{k=1}^{N}\sum_{i=1}^{n_k}\sum_{\mathbf{z}_{pi,k},\mathbf{z}_{prop,k},\mathbf{w}_k,\mathbf{l}_k}$<br>$E[Y_{ki} \mid A_{ki}=1, \mathbf{z}_{pi,k}, \mathbf{z}_{prop,k}, \mathbf{w}_k, \mathbf{l}_k]$<br>$\times P(\mathbf{z}_{prop,k} \mid A_{ki}=1, \mathbf{w}_k, \mathbf{l}_k)$<br>$\times P(\mathbf{z}_{pi,k} \mid A_{ki}=0, \mathbf{w}_k, \mathbf{l}_k) P(\mathbf{w}_k, \mathbf{l}_k \mid A_{ki}=1)$ |
| Proposal attributes | $E[\mathbf{Y}(\mathbf{Z}^*_{prop})\|\mathbf{A}=1] - E[\mathbf{Y}\|\mathbf{A}=1]$<br>$E[\mathbf{Y}\|\mathbf{A}=0] - E[\mathbf{Y}(\mathbf{Z}^*_{prop})\|\mathbf{A}=1]$ | $Y_{ki}(\mathbf{z}_{prop,k}) \perp \mathbf{Z}_{prop,k} \mid \mathbf{W}_k, \mathbf{A}_k, \mathbf{Z}_{pi,k}, \mathbf{L}_k$<br>$P(\mathbf{A}_k \mid \mathbf{W}_k, \mathbf{L}_k) > 0$ a.e.<br>$P(\mathbf{Z}_{prop,k} \mid \mathbf{A}_k, \mathbf{W}_k, \mathbf{L}_k, \mathbf{Z}_{pi,k}) > 0$ a.e. | $\frac{1}{M}\sum_{k=1}^{N}\sum_{i=1}^{n_k}\sum_{\mathbf{z}_{prop,k},\mathbf{z}_{pi,k},\mathbf{w}_k,\mathbf{l}_k}$<br>$E[Y_{ki} \mid A_{ki}=1, \mathbf{z}_{prop,k}, \mathbf{z}_{pi,k}, \mathbf{w}_k, \mathbf{l}_k]$<br>$\times P(\mathbf{z}_{pi,k} \mid A_{ki}=1, \mathbf{w}_k, \mathbf{l}_k)$<br>$\times P(\mathbf{z}_{prop,k} \mid A_{ki}=0, \mathbf{w}_k, \mathbf{l}_k)$<br>$\times P(\mathbf{w}_k, \mathbf{l}_k \mid A_{ki}=1)$ |
| PI & Proposal attributes | $E[\mathbf{Y}(\mathbf{Z}^*)\|\mathbf{A}=1] - E[\mathbf{Y}\|\mathbf{A}=1]$<br>$E[\mathbf{Y}\|\mathbf{A}=0] - E[\mathbf{Y}(\mathbf{Z}^*)\|\mathbf{A}=1]$ | $Y_{ki}(\mathbf{z}_k) \perp \mathbf{Z}_k \mid \mathbf{W}_k, \mathbf{A}_k, \mathbf{L}_k$<br>$P(\mathbf{A}_k \mid \mathbf{W}_k, \mathbf{L}_k) > 0$ a.e.<br>$P(\mathbf{Z}_k \mid \mathbf{A}_k, \mathbf{W}_k, \mathbf{L}_k) > 0$ a.e. | $\frac{1}{M}\sum_{k=1}^{N}\sum_{i=1}^{n_k}\sum_{\mathbf{z}_k,\mathbf{w}_k,\mathbf{l}_k} P(\mathbf{w}_k,\mathbf{l}_k \mid A_{ki}=1)$<br>$\times E[Y_{ki} \mid A_{ki}=1, \mathbf{z}_k, \mathbf{w}_k, \mathbf{l}_k]$<br>$\times P(\mathbf{z}_k \mid A_{ki}=0, \mathbf{w}_k, \mathbf{l}_k)$ |
| Structural attributes | $E[\mathbf{Y}(\mathbf{L}^*)\|\mathbf{A}=1] - E[\mathbf{Y}\|\mathbf{A}=1]$<br>$E[\mathbf{Y}\|\mathbf{A}=0] - E[\mathbf{Y}(\mathbf{L}^*)\|\mathbf{A}=1]$ | $Y_{ki}(\mathbf{l}_k) \perp \mathbf{L}_k \mid \mathbf{W}_k, \mathbf{A}_k, \mathbf{Z}_k$<br>$P(\mathbf{A}_k \mid \mathbf{W}_k) > 0$ a.e.<br>$P(\mathbf{L}_k \mid \mathbf{A}_k, \mathbf{W}_k, \mathbf{Z}_k) > 0$ a.e. | $\frac{1}{M}\sum_{k=1}^{N}\sum_{i=1}^{n_k}\sum_{\mathbf{z}_k,\mathbf{w}_k,\mathbf{l}_k} P(\mathbf{w}_k,\mathbf{z}_k \mid A_{ki}=1)$<br>$\times E[Y_{ki} \mid A_{ki}=1, \mathbf{z}_k, \mathbf{w}_k, \mathbf{l}_k]$<br>$\times P(\mathbf{l}_k \mid A_{ki}=0, \mathbf{w}_k)$ |

## *A.5 Identification Proofs*

We provide identification proofs for the Disparity Reduction, Disparity Remaining, and Thresholded Disparity estimands under no selection bias.

### *A.5.1 Identification Proofs for Disparity Reduction and Disparity Remaining*

We identify Disparity Reduction and Disparity Remaining for the case of equalizing Preliminary Overall Impact Scores. We estimate these quantities under several interventions, equalizing the Black and white PI distributions of: 1) mean Preliminary Overall Impact Score, 2) application type and amended status jointly, 3) degree category, funding bin, and

Table A.7: Summary of the estimands of interest, identification assumptions, and identification forms for interventions targeting the applicant-related criterion scores under the Simultaneous and Normatively Backwards Evaluation Models. We study both equalization-based and thresholding-based interventions on disparities in selection into discussion. In addition to the notation introduced in Section 3.4, we denote applicant-related criterion scores by $M_e$ and proposal-related criterion scores by $M_p$.

| Model | Estimands of Interest | Identification Assumptions | Identification Form of Counterfactual Quantity (e.g. $E[\mathbf{Y}(\mathbf{M}_e^t)\|\mathbf{A}=1]$) |
|---|---|---|---|
| Simultaneous | | Equalization | |
| | $E[\mathbf{Y} \mid \mathbf{A}=1] - E[\mathbf{Y}(\mathbf{M}_e^*) \mid \mathbf{A}=1]$<br>$E[\mathbf{Y}(\mathbf{M}_e^*) \mid \mathbf{A}=1] - E[\mathbf{Y} \mid \mathbf{A}=0]$ | $Y_{ki}(\mathbf{m}_{e,k}) \perp \mathbf{M}_{e,k} \mid \mathbf{W}_k, \mathbf{A}_k, \mathbf{Z}_k, \mathbf{L}_k, \mathbf{M}_{p,k}$<br>$P(\mathbf{A}_k \mid \mathbf{W}_k, \mathbf{L}_k) > 0$ a.e.<br>$P(\mathbf{M}_{e,k} \mid \mathbf{A}_k, \mathbf{W}_k, \mathbf{L}_k, \mathbf{Z}_k, \mathbf{M}_{p,k}) > 0$ a.e. | $E[\mathbf{Y}(\mathbf{M}_e^*) \mid \mathbf{A}=1] =$<br>$\frac{1}{M}\sum_{k=1}^{N}\sum_{i=1}^{n_k}\sum_{\mathbf{m}_{e,k},\mathbf{m}_{p,k},\mathbf{z}_k,\mathbf{w}_k,\mathbf{l}_k}$<br>$P(\mathbf{w}_k, \mathbf{l}_k \mid A_{ki}=1)$<br>$\times E[Y_{ki} \mid A_{ki}=1, \mathbf{m}_{e,k}, \mathbf{m}_{p,k}, \mathbf{z}_k, \mathbf{w}_k, \mathbf{l}_k]$<br>$\times P(\mathbf{z}_k, \mathbf{m}_{p,k} \mid A_{ki}=1, \mathbf{w}_k, \mathbf{l}_k)$<br>$\times P(\mathbf{m}_{e,k} \mid A_{ki}=0, \mathbf{w}_k, \mathbf{l}_k)$ |
| | Thresholding | | |
| | $E[\mathbf{Y}(\mathbf{M}_e^t) \mid \mathbf{A}=0] - E[\mathbf{Y}(\mathbf{M}_e^t) \mid \mathbf{A}=1]$ | $Y_{ki}(\mathbf{m}_{e,k}) \perp \mathbf{M}_{e,k} \mid \mathbf{W}_k, \mathbf{A}_k, \mathbf{Z}_k, \mathbf{M}_{p,k}, \mathbf{L}_k$<br>$P(\mathbf{A}_k \mid \mathbf{W}_k, \mathbf{L}_k, \mathbf{Z}_k, \mathbf{M}_{p,k}) > 0$ a.e.<br>$P(\mathbf{M}_{e,k}^t \mid \mathbf{A}_k, \mathbf{W}_k, \mathbf{L}_k, \mathbf{Z}_k, \mathbf{M}_{p,k}) > 0$ a.e. | $E[\mathbf{Y}(\mathbf{M}_e^t) \mid \mathbf{A}=1] =$<br>$\frac{1}{M}\sum_{k=1}^{N}\sum_{i=1}^{n_k}\sum_{\mathbf{l}_k,\mathbf{w}_k,\mathbf{z}_k,\mathbf{m}_{p,k},\mathbf{m}_{e,k}\in\{m^t\}}$<br>$P(\mathbf{w}_k, \mathbf{l}_k, \mathbf{z}_k, \mathbf{m}_{p,k} \mid A_{ki}=1)$<br>$\times E[Y_{ki} \mid A_{ki}=1, \mathbf{m}_{e,k}, \mathbf{z}_k, \mathbf{w}_k, \mathbf{l}_k, \mathbf{m}_{p,k}]$<br>$\times P(\mathbf{M}_e^t = \mathbf{m}_{e,k} \mid A_{ki}=1, \mathbf{w}_k, \mathbf{l}_k, \mathbf{z}_k, \mathbf{m}_{p,k})$ |
| Normatively Backwards | | Equalization | |
| | $E[\mathbf{Y} \mid \mathbf{A}=1] - E[\mathbf{Y}(\mathbf{M}_e^*) \mid \mathbf{A}=1]$<br>$E[\mathbf{Y}(\mathbf{M}_e^*) \mid \mathbf{A}=1] - E[\mathbf{Y} \mid \mathbf{A}=0]$ | $Y_{ki}(\mathbf{m}_{e,k}) \perp \mathbf{M}_{e,k} \mid \mathbf{W}_k, \mathbf{A}_k, \mathbf{Z}_k, \mathbf{L}_k$<br>$P(\mathbf{A}_k \mid \mathbf{W}_k, \mathbf{L}_k) > 0$ a.e.<br>$P(\mathbf{M}_{e,k} \mid \mathbf{A}_k, \mathbf{W}_k, \mathbf{L}_k, \mathbf{Z}_k) > 0$ a.e. | $E[\mathbf{Y}(\mathbf{M}_e^*) \mid \mathbf{A}=1] =$<br>$\frac{1}{M}\sum_{k=1}^{N}\sum_{i=1}^{n_k}\sum_{\mathbf{m}_{e,k},\mathbf{z}_k,\mathbf{w}_k,\mathbf{l}_k}$<br>$P(\mathbf{w}_k, \mathbf{l}_k \mid A_{ki}=1)$<br>$\times E[Y_{ki} \mid A_{ki}=1, \mathbf{m}_{e,k}, \mathbf{z}_k, \mathbf{w}_k, \mathbf{l}_k]$<br>$\times P(\mathbf{z}_k \mid A_{ki}=1, \mathbf{w}_k, \mathbf{l}_k)$<br>$\times P(\mathbf{m}_{e,k} \mid A_{ki}=0, \mathbf{w}_k, \mathbf{l}_k)$ |
| | Thresholding | | |
| | $E[\mathbf{Y}(\mathbf{M}_e^t) \mid \mathbf{A}=0] - E[\mathbf{Y}(\mathbf{M}_e^t) \mid \mathbf{A}=1]$ | $Y_{ki}(\mathbf{m}_{e,k}) \perp \mathbf{M}_{e,k} \mid \mathbf{W}_k, \mathbf{A}_k, \mathbf{Z}_k, \mathbf{L}_k$<br>$P(\mathbf{A}_k \mid \mathbf{W}_k, \mathbf{L}_k, \mathbf{Z}_k) > 0$ a.e.<br>$P(\mathbf{M}_{e,k}^t \mid \mathbf{A}_k, \mathbf{W}_k, \mathbf{L}_k, \mathbf{Z}_k) > 0$ a.e. | $E[\mathbf{Y}(\mathbf{M}_e^t) \mid \mathbf{A}=1] =$<br>$\frac{1}{M}\sum_{k=1}^{N}\sum_{i=1}^{n_k}\sum_{\mathbf{l}_k,\mathbf{w}_k,\mathbf{z}_k,\mathbf{m}_{e,k}\in\{m^t\}}$<br>$P(\mathbf{w}_k, \mathbf{l}_k, \mathbf{z}_k \mid A_{ki}=1)$<br>$\times E[Y_{ki} \mid A_{ki}=1, \mathbf{m}_{e,k}, \mathbf{z}_k, \mathbf{w}_k, \mathbf{l}_k]$<br>$\times P(\mathbf{M}_e^t = \mathbf{m}_{e,k} \mid A_{ki}=1, \mathbf{w}_k, \mathbf{l}_k, \mathbf{z}_k)$ |

career stage jointly, and 4) administering organization and IRG jointly. Identification of these estimands under such interventions requires: 1) partial interference, 2) consistency at the SRG level, 3) no unmeasured confounding of the intervention-outcome relationship at the SRG level, and 4) positivity at the SRG level. The same proof strategy applies to the other interventions we study. Table A.6 summarizes the interventions, estimands of interest, corresponding identification assumptions, and identification forms for the equalization analysis in Section 3.6.1.

*Proof.*

$$
\begin{aligned}
P(Y_{ki}(\mathbf{M}_k = \mathbf{M}_k^\star)|A_{ki}=1,\mathbf{w}_k,\mathbf{l}_k) &= \sum_{\mathbf{m}_k} P(Y_{ki}(\mathbf{M}_k=\mathbf{m}_k)|A_{ki}=1,\mathbf{w}_k,\mathbf{l}_k,\mathbf{M}_k^\star=\mathbf{m}_k) \\
&\quad \times P(\mathbf{M}_k^\star=\mathbf{m}_k|A_{ki}=1,\mathbf{w}_k,\mathbf{l}_k) \quad \text{(LOTP \& consistency)} \\
&= \sum_{\mathbf{m}_k} P(Y_{ki}(\mathbf{M}_k=\mathbf{m}_k)|A_{ki}=1,\mathbf{w}_k,\mathbf{l}_k) \\
&\quad \times P(\mathbf{M}_k^\star=\mathbf{m}_k|A_{ki}=1,\mathbf{w}_k,\mathbf{l}_k) \\
&\text{since } Y_{ki}(\mathbf{M}_k=\mathbf{m}_k) \perp \mathbf{M}_k^\star=\mathbf{m}_k|\{\mathbf{A}_k,\mathbf{W}_k,\mathbf{L}_k\} \text{ by definition of} \\
&\mathbf{M}_k^\star \text{ as a random draw given } \mathbf{W}=w_k,\mathbf{L}_k=l_k \\
&= \sum_{\mathbf{m}_k} P(Y_{ki}(\mathbf{M}_k=\mathbf{m}_k)|A_{ki}=1,\mathbf{w}_k,\mathbf{l}_k) \\
&\quad \times P(\mathbf{M}_k^\star=\mathbf{m}_k|A_{ki}=0,\mathbf{w}_k,\mathbf{l}_k) \\
&\text{We have } \mathbf{M}_k^\star=\mathbf{m}_k^\star \perp \mathbf{A}_k|\mathbf{W}_k,\mathbf{L}_k \text{ by definition of } \mathbf{M}_k^\star \text{ as a} \\
&\text{random draw given } \mathbf{W}_k,\mathbf{L}_k \\
&= \sum_{\mathbf{m}_k} P(Y_{ki}(\mathbf{M}_k=\mathbf{m}_k)|A_{ki}=1,\mathbf{w}_k,\mathbf{l}_k) \\
&\quad \times P(\mathbf{M}_k=\mathbf{m}_k|A_{ki}=0,\mathbf{w}_k,\mathbf{l}_k) \\
&\text{definition of } \mathbf{M}_k^\star \sim P(\mathbf{M}_k|A_{ki}=0,\mathbf{W}_k,\mathbf{L}_k) \\
&= \sum_{\mathbf{m}_k,\mathbf{z}_k} P(Y_{ki}(\mathbf{M}_k=\mathbf{m}_k)|A_{ki}=1,\mathbf{w}_k,\mathbf{l}_k,\mathbf{z}_k) \\
&\quad \times P(\mathbf{z}_k|A_{ki}=1,\mathbf{w}_k,\mathbf{l}_k) \\
&\quad \times P(\mathbf{M}_k=\mathbf{m}_k|A_{ki}=0,\mathbf{w}_k,\mathbf{l}_k) \quad \text{(LOTP)} \\
&= \sum_{\mathbf{m}_k,\mathbf{z}_k} P(Y_{ki}(\mathbf{M}_k=\mathbf{m}_k)|A_{ki}=1,\mathbf{w}_k,\mathbf{l}_k,\mathbf{m}_k,\mathbf{z}_k) \\
&\quad \times P(\mathbf{z}_k|A_{ki}=1,\mathbf{w}_k,\mathbf{l}_k) \\
&\quad \times P(\mathbf{M}_k=\mathbf{m}_k|A_{ki}=0,\mathbf{w}_k,\mathbf{l}_k), \text{ since } (\mathbf{Y}(\mathbf{m}_k) \perp \mathbf{M}_k|\mathbf{A}_k,\mathbf{W}_k,\mathbf{L}_k,\mathbf{Z}_k) \\
&= \sum_{\mathbf{m}_k,\mathbf{z}_k} P(Y_{ki}|A_{ki}=1,\mathbf{w}_k,\mathbf{l}_k,\mathbf{m}_k,\mathbf{z}_k)P(\mathbf{z}_k|A_{ki}=1,\mathbf{w}_k,\mathbf{l}_k) \\
&\quad \times P(\mathbf{M}_k=\mathbf{m}_k|A_{ki}=0,\mathbf{w}_k,\mathbf{l}_k) \quad \text{(consistency)}
\end{aligned}
$$

Let $M = \sum_{k=1}^{N} n_k$.

$$
\begin{aligned}
E[Y(\mathbf{M}^*)|A=1] &= \frac{1}{M}\sum_{k=1}^{N}\sum_{i=1}^{n_k}\sum_{\mathbf{w}_k,\mathbf{l}_k,\mathbf{m}_k,\mathbf{z}_k} P(Y_{ki}(\mathbf{M}^*)|A_{ki}=1,\mathbf{w}_k,\mathbf{l}_k)P(\mathbf{w}_k,\mathbf{l}_k|A_{ki}=1) \\
&= \frac{1}{M}\sum_{k=1}^{N}\sum_{i=1}^{n_k}\sum_{\mathbf{w}_k,\mathbf{l}_k,\mathbf{m}_k,\mathbf{z}_k} P(Y_{ki}|A_{ki}=1,\mathbf{w}_k,\mathbf{l}_k,\mathbf{m}_k,\mathbf{z}_k) \\
&\quad \times P(\mathbf{z}_k,\mathbf{w}_k,\mathbf{l}_k|A_{ki}=1)P(\mathbf{M}_k=\mathbf{m}_k|A_{ki}=0,\mathbf{w}_k,\mathbf{l}_k)
\end{aligned}
$$

We can apply this derivation to multiple mediators. When there are multiple mediators we can model the mediators as follows:

$$
\begin{aligned}
P(\mathbf{m}=(m_1,m_2,\cdots,m_k)|w,l,A=0) &= P(m_1|w,l,A=0) \\
&\quad \times P(m_2|m_1,w,l,A=0) \\
&\quad \times \cdots P(m_k|m_1,\cdots,m_{k-1},w,l,A=0)
\end{aligned}
$$

□

### *A.5.2 Identification of Disparity after Thresholding Scores*

We also use the causal decomposition framework to analyze the impact of thresholding the Environment and Investigator criterion scores using the Thresholded Disparity. Specifically, we study this estimand under both the Simultaneous and Normatively Backwards Evaluation Models. The identification argument closely parallels that for Disparity Reduction and Disparity Remaining in Section A.5.1. We give the proof under the Normatively Backwards Evaluation Model; analogous steps yield the corresponding identification form under the Simultaneous Evaluation Model. In our analysis, we connect the binarization analysis to the equalization analysis by noting that extreme thresholds effectively equalize the distributions between Black and white PIs. Table A.7 summarizes the estimands of interest, identification

assumptions, and identification forms for interventions on the applicant-related criterion scores under both the Simultaneous and Normatively Backwards Evaluation Models.

*Proof.*

$$
\begin{aligned}
P(Y_{ki}(\mathbf{M}_{e,k} = \mathbf{M}^t_{e,k})|\mathbf{a}_k, \mathbf{w}_k, \mathbf{l}_k, \mathbf{z}_k) &= \sum_{\mathbf{m}_{e,k} \in \{m^t\}} P(Y_{ki}(\mathbf{M}_{e,k} = \mathbf{m}_{e,k})|\mathbf{a}_k, \mathbf{w}_k, \mathbf{l}_k, \mathbf{z}_k, \mathbf{M}^t_{e,k} = \mathbf{m}_{e,k}) \\
&\quad \times P(\mathbf{M}^t_{e,k} = \mathbf{m}_{e,k}|\mathbf{a}_k, \mathbf{w}_k, \mathbf{l}_k, \mathbf{z}_k) \quad \text{(LOTP \& consistency)} \\
&= \sum_{\mathbf{m}_{e,k} \in \{m^t\}} P(Y_{ki}(\mathbf{M}_{e,k} = \mathbf{m}_{e,k})|\mathbf{a}_k, \mathbf{w}_k, \mathbf{l}_k, \mathbf{z}_k) \\
&\quad \times P(\mathbf{M}^t_{e,k} = \mathbf{m}_{e,k}|\mathbf{a}_k, \mathbf{w}_k, \mathbf{l}_k, \mathbf{z}_k) \\
&\quad Y_{ki}(\mathbf{M}_{e,k} = \mathbf{m}_{e,k}) \perp \mathbf{M}^t_{e,k} = \mathbf{m}_{e,k}|\{\mathbf{A}_k, \mathbf{W}_k, \mathbf{L}_k, \mathbf{Z}_k\} \text{ since} \\
&\quad Y_{ki}(\mathbf{M}_{e,k} = \mathbf{m}_{e,k}) \perp \mathbf{M}^e_k = \mathbf{m}_{e,k} \mid \{\mathbf{A}_k, \mathbf{W}_k, \mathbf{L}_k, \mathbf{Z}_k\}) \\
&\quad \text{and we fix } \mathbf{M}^t_{e,k} = \mathbf{M}_{e,k} = \mathbf{m}_{e,k} \\
&= \sum_{\mathbf{m}_{e,k} \in \{m^t\}} P(Y_{ki}(\mathbf{M}_{e,k} = \mathbf{m}_{e,k})|\mathbf{a}_k, \mathbf{w}_k, \mathbf{l}_k, \mathbf{z}_k, \mathbf{M}_{e,k} = \mathbf{m}_{e,k}) \\
&\quad \times P(\mathbf{M}^t_{e,k} = \mathbf{m}_{e,k}|\mathbf{a}_k, \mathbf{w}_k, \mathbf{l}_k, \mathbf{z}_k) \\
&\quad \text{(by independence: } Y_{ki}(\mathbf{m}_{e,k}) \perp \mathbf{M}_{e,k} \mid \{\mathbf{A}_k, \mathbf{W}_k, \mathbf{L}_k, \mathbf{Z}_k\}) \\
&= \sum_{\mathbf{m}_{e,k} \in \{m^t\}} P(Y_{ki}|\mathbf{a}_k, \mathbf{w}_k, \mathbf{l}_k, \mathbf{z}_k, \mathbf{M}_{e,k} = \mathbf{m}_{e,k}) \\
&\quad \times P(\mathbf{M}^t_{e,k} = \mathbf{m}_{e,k}|\mathbf{a}_k, \mathbf{w}_k, \mathbf{l}_k, \mathbf{z}_k) \quad \text{(consistency)}
\end{aligned}
$$

Let $M = \sum_{k=1}^{N} n_k$. Then we have that under the Normatively Backwards Evaluation Model:

$$
\begin{aligned}
E_{CD}[Y(\mathbf{M}^{e,t})|A = 1] &= \frac{1}{M} \sum_{k=1}^{N} \sum_{i=1}^{n_k} \sum_{\mathbf{w}_k, \mathbf{l}_k, \mathbf{z}_k} E[Y_{ki}(\mathbf{M}^e_k = \mathbf{M}^{e,t}_k)|\mathbf{w}_k, \mathbf{l}_k, \mathbf{z}_k, \mathbf{A}_k = 1] P(\mathbf{w}_k, \mathbf{l}_k, \mathbf{z}_k|\mathbf{A}_k = 1) \\
&= \frac{1}{M} \sum_{k=1}^{N} \sum_{i=1}^{n_k} \sum_{\mathbf{w}_k, \mathbf{l}_k, \mathbf{z}_k, \mathbf{m}^e_k \in \{m^t\}} E[Y_{ki}|\mathbf{w}_k, \mathbf{l}_k, \mathbf{z}_k, \mathbf{A}_k = 1, \mathbf{M}^e_k = \mathbf{m}^e_k] \\
&\quad \times P(\mathbf{M}^{e,t}_k = \mathbf{m}^e_k|\mathbf{w}_k, \mathbf{l}_k, \mathbf{z}_k, \mathbf{A}_k = 1) P(\mathbf{w}_k, \mathbf{l}_k, \mathbf{z}_k|\mathbf{A}_k = 1) \qquad \text{(S8)}
\end{aligned}
$$

□

Similarly, the identification form for the Thresholded Disparity under the Simultaneous

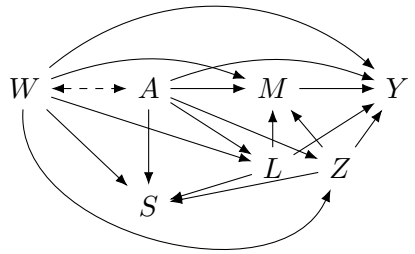



Figure A.3: Causal Diagram of NIH Peer Review Process after accounting for Selection Bias. This diagram illustrates the causal relationships within the NIH peer review process. In this representation, $A$ corresponds to the PI race, with $A = 0$ representing white PIs and $A = 1$ representing Black PIs. The outcome, $Y$, signifies the selection of proposals for discussion. The selection is denoted as $S$. $W$ denotes the PI's ethnicity and gender, regarded as protected attributes not currently under examination. We allow for an association between ethnicity, gender ($W$), and PI race ($A$), as indicated by the dotted bi-directional arrow. Intermediate variables linking PI race to selection into discussion include $M$ (Preliminary Overall Impact Score); $Z$ (Application Type, Amended Status, Career Stage, Degree Category, and Funding Bin, jointly); and $L$ (Integrated Review Group [IRG], Scientific Review Group [SRG], and Administering Organization).

Evaluation Model is

$$
\begin{aligned}
E_{Doc}[Y(\mathbf{M}^{e,t})|A=1] &= \frac{1}{M}\sum_{k=1}^{N}\sum_{i=1}^{n_k}\sum_{\mathbf{w}_k,\mathbf{l}_k,\mathbf{z}_k,\mathbf{m}_{p,k}} E[Y_{ki}(\mathbf{M}_k^e=\mathbf{M}_k^{e,t})|\mathbf{w}_k,\mathbf{l}_k,\mathbf{m}_{p,k},\mathbf{z}_k,\mathbf{A}_k=1] \\
&\quad \times P(\mathbf{w}_k,\mathbf{l}_k,\mathbf{z}_k,\mathbf{m}_{p,k}|\mathbf{A}_k=1) \\
&= \frac{1}{M}\sum_{k=1}^{N}\sum_{i=1}^{n_k}\sum_{\mathbf{w}_k,\mathbf{l}_k,\mathbf{z}_k,\mathbf{m}_{p,k},\mathbf{m}_k^e\epsilon\{m^t\}} E[Y_{ki}|\mathbf{w}_k,\mathbf{l}_k,\mathbf{z}_k,\mathbf{m}_{p,k},\mathbf{A}_k=1,\mathbf{M}_k^e=\mathbf{m}_k^e] \\
&\quad \times P(\mathbf{M}_k^{e,t}=\mathbf{m}_k^e|\mathbf{w}_k,\mathbf{l}_k,\mathbf{z}_k,\mathbf{m}_{p,k},\mathbf{A}_k=1)P(\mathbf{w}_k,\mathbf{l}_k,\mathbf{z}_k,\mathbf{m}_{p,k}|\mathbf{A}_k=1) \qquad \text{(S9)}
\end{aligned}
$$

### *A.6 Selection*

We define self-selection as the process of applicants' choice to apply for the NIH grant in the first place (denoted as $S = 1$). Figure A.3, shows our DAG after accounting for this selection. In this setting, we are interested in the following estimands:

- Observed Disparity: $E[Y|A=1,S=1] - E[Y|A=0,S=1]$

- Thresholded Disparity: $E[Y(M^t)|A=0,S=1] - E[Y(M^t)|A=1,S=1] \quad s.t \quad P(Y_k(M^t)=1|S=1) = d_k$
- Disparity Reduction: $E[Y|A=1,S=1] - E[Y(M^*)|A=1,S=1] \quad s.t \quad P(Y_k(M^*)=1|S=1) = d_k$
- Disparity Remaining: $E[Y(M^*)|A=1,S=1] - E[Y|A=0,S=1] \quad s.t \quad P(Y_k(M^*)=1|S=1) = d_k$

Where $M^* \sim P(M|A=0,S=1,W,L)$ and $M^t$ has the same definition as in Section 3.4. We are interested in these quantities for the selected population or the population that applied for NIH grants. The identification assumptions and proof are identical to what was shown previously in Section 3.4, A.5.1 and A.5.2 with the extra conditioning on $S=1$.

So the final identification forms are the following:

$$
\begin{aligned}
E[Y(\mathbf{M}^*)|A=1,S=1] &= \frac{1}{M}\sum_{k=1}^{N}\sum_{i=1}^{n_k}\sum_{\mathbf{w}_k,\mathbf{l}_k,\mathbf{m}_k,\mathbf{z}_k,S_{ki}=1} P(Y_{ki}|A_{ki}=1,\mathbf{w}_k,\mathbf{l}_k,\mathbf{m}_k,\mathbf{z}_k,S_{ki}=1) \\
&\quad \times P(\mathbf{z}_k,\mathbf{w}_k,\mathbf{l}_k|A_{ki}=1,S_{ki}=1)P(\mathbf{M}_k=\mathbf{m}_k|A_{ki}=0,\mathbf{w}_k,\mathbf{l}_k,S_{ki}=1) \\
E_{CD}[Y(\mathbf{M}^{e,t})|A=1,S=1] &= \frac{1}{M}\sum_{k=1}^{N}\sum_{i=1}^{n_k}\sum_{\mathbf{w}_k,\mathbf{l}_k,\mathbf{z}_k,\mathbf{m}_k^e\in\{m^t\}} E[Y_{ki}|\mathbf{w}_k,\mathbf{l}_k,\mathbf{z}_k,A_{ki}=1,\mathbf{M}_k^e=\mathbf{m}_k^e,S_{ki}=1] \\
&\quad \times P(\mathbf{M}_k^{e,t}=\mathbf{m}_k^e|\mathbf{w}_k,\mathbf{l}_k,\mathbf{z}_k,A_{ki}=1,S_{ki}=1)P(\mathbf{w}_k,\mathbf{l}_k,\mathbf{z}_k|A_{ki}=1,S_{ki}=1) \\
E_{Doc}[Y(\mathbf{M}^{e,t})|A=1,S=1] &= \frac{1}{M}\sum_{k=1}^{N}\sum_{i=1}^{n_k}\sum_{\mathbf{w}_k,\mathbf{l}_k,\mathbf{z}_k,\mathbf{m}_{p,k},\mathbf{m}_k^e\in\{m^t\}} \\
&\quad E[Y_{ki}|\mathbf{w}_k,\mathbf{l}_k,\mathbf{z}_k,\mathbf{m}_{p,k},A_{ki}=1,\mathbf{M}_k^e=\mathbf{m}_k^e,S_{ki}=1] \\
&\quad \times P(\mathbf{M}_k^{e,t}=\mathbf{m}_k^e|\mathbf{w}_k,\mathbf{l}_k,\mathbf{z}_k,\mathbf{m}_{p,k},A_{ki}=1,S_{ki}=1) \\
&\quad \times P(\mathbf{w}_k,\mathbf{l}_k,\mathbf{z}_k,\mathbf{m}_{p,k}|A_{ki}=1,S_{ki}=1)
\end{aligned}
$$

## *A.7 Full Thresholded Disparity Results*

Figure A.4 shows the results for all binarization thresholds. In the Normatively Backwards Evaluation Model, thresholds of 4 or higher substantially reduce disparities. However, as

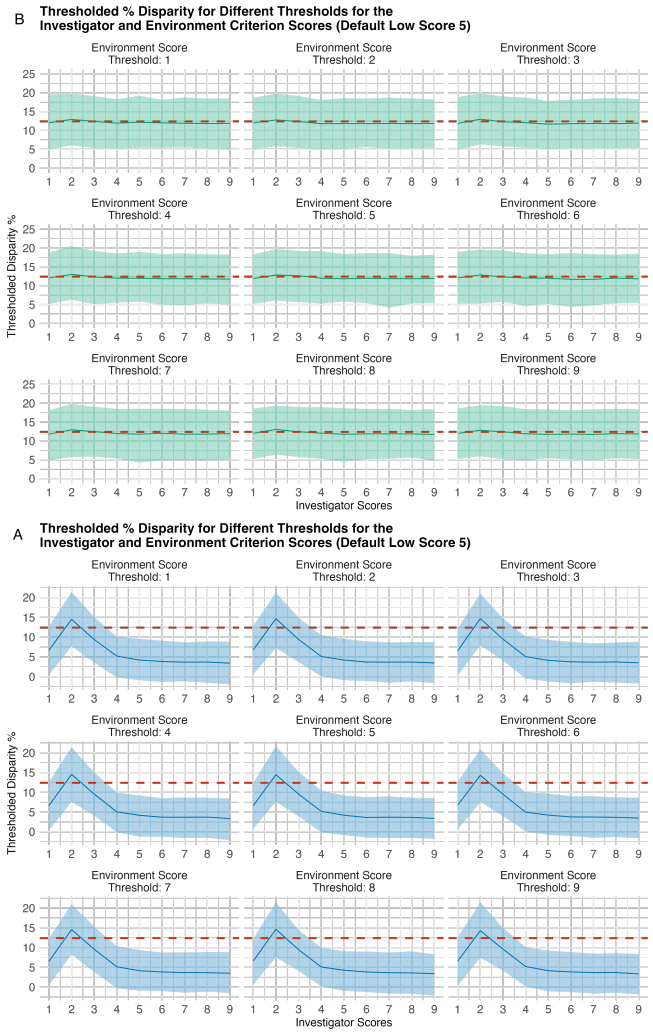



Figure A.4: **Thresholded disparity results:** Each plot represents different thresholds for the Environment criterion score for thresholds from 1 to 9. Each plot shows the % Thresholded Disparity for that Environment criterion score threshold and varying Investigator criterion score thresholds for thresholds from 1 to 9 also based on historical NIH discussion rates. The top plot (all in green) corresponds to the Thresholded Disparity under Model 6a. The green shaded region represents the point-wise Credible Intervals corresponding to that model. The bottom plot (all in blue) corresponds to the Thresholded Disparity under Model 6b. The blue shaded region represents the point-wise 95% Credible Intervals corresponding to that model. For both models, scores that fell below the Environment and Investigator criterion score thresholds were treated as if they received an Environment and Investigator criterion score of 5. The red dotted line is the overall observed disparity, which is 12.3% (11.3%, 16.3%).

noted earlier, such high thresholds may be unlikely given historical NIH discussion rates. In the Simultaneous Evaluation Model, binarizing the Investigator and Environment criterion scores has no impact on disparities, regardless of the threshold.

# Appendix B

# APPENDIX TO CHAPTER 4

The appendix for this chapter is organized as follows. In Section B.1, we describe the unidentifiable cases in PNL models. Section B.2 provides additional intuition for how $\xi$ behaves under different properties. Finally, Section B.3 presents simulation results under model misspecification for PNLs.

## *B.1 Unidentifiable Cases in Post-Nonlinear Models*

Following Zhang and Hyvarinen (2012), the causal direction in the Post-Nonlinear (PNL) model is unidentifiable in the following five cases:

1. Both $f_1$ and $f_2$ are linear and the noise $\eta$ is Gaussian.

2. The model reduces to an invertible linear transformation with symmetric distributions for $X$ and $\eta$.

3. The noise $\eta$ and the cause $X$ have certain exponential family relationships that render $f_1$ and $f_2$ indistinguishable.

4. The functional relationships satisfy specific differential equation constraints that make both directions equally valid.

5. Degenerate cases where $f_2$ is constant or non-invertible, leading to lack of model identifiability.

## B.2 Example Behaviors of $\xi$

To understand the behavior of $\xi$ under the ANM model class ($f_2, g_2$ are identity functions), we examine several illustrative cases.

1. Suppose $f_1$ is additive, i.e. $f_1(a+b) = f_1(a) + f_1(b)$, and $g_1 = f_1^{-1}$:

$$\xi = f_1^{-1}(Y) - f_1^{-1}(\eta) - f_1^{-1}(Y) = f_1^{-1}(\eta),$$

   here $\xi$ is only a function of $\eta$.

2. Suppose $\eta = 0$, and $g_1 = f_1^{-1}$. Then

$$\xi = f_1^{-1}(Y) - f_1^{-1}(Y) = 0,$$

   so the reverse-direction residual vanishes.

3. Suppose $f_1$ is nonlinear, and $g_1$ is restricted to be linear, $g_1(Y) = \gamma Y$ for some $\gamma \in \mathbb{R}$. Then

$$\xi = f_1^{-1}(Y - \eta) - \gamma Y.$$

   In this case, $\xi$ depends on both $Y$ and $\eta$, and hence $Y \not\perp \xi$.

Taken together, these cases illustrate that under the ANM model class, when $f_1$ is nonlinear, there is no choice of $g_1$ within a restricted model class that yields $\xi \perp Y$. Consequently, except in the standard unidentifiable cases (e.g., the linear–Gaussian setting), we have $Y \not\perp \xi$ when fitting the model in the incorrect causal direction.

## B.3 PNL Model Misspecification Simulation Results

Table B.1 presents the simulation results for the test-based approach paired with the PNL model class under varying levels of model misspecification.

| Degree of misspecification | Test-based approach | |
|---|---|---|
| | Decision rates | Average causal outcome rates |
| No misspecification | $X \to Y$**: 100%**<br>$Y \to X$: 0%<br>Inconclusive: 0% | Reject both: 1.9%<br>Fail to reject both: 0%<br>Reject only $X \to Y$: 98.1%<br>Reject only $Y \to X$: 0% |
| Partial misspecification | $X \to Y$: 0%<br>$Y \to X$: 0%<br>**Inconclusive: 100%** | Reject both: 100%<br>Fail to reject both: 0%<br>Reject only $X \to Y$: 0%<br>Reject only $Y \to X$: 0% |
| Full misspecification | $X \to Y$: 0%<br>$Y \to X$: 0%<br>**Inconclusive: 100%** | Reject both: 100%<br>Fail to reject both: 0%<br>Reject only $X \to Y$: 0%<br>Reject only $Y \to X$: 0% |

Table B.1: Test-based approach under varying degrees of model misspecification. Rows correspond to settings with no misspecification (both $f_1$ and $f_2$ match their true forms), partial misspecification (only $f_1$ is misspecified as linear), and full misspecification (both $f_1$ and $f_2$ are misspecified as linear). We report the resulting decision rates and average causal outcome rates across $M$ simulated datasets.

# Appendix C

# APPENDIX TO CHAPTER 5

The appendix for this chapter is organized as follows. In Section C.1, we prove the asymptotic joint distribution of $(T_{1b,X\to Y}, T_{1c,Y\to X})$. Section C.2 provides details for our simulation setup. Finally, Section C.3 presents simulation results under severe model misspecification.

## *C.1 Asymptotic Joint Distribution of* $(T_{1b,X\to Y}, T_{1c,Y\to X})$

In this appendix we show that the joint distribution of the test statistics $(T_{1b,X\to Y}, T_{1c,Y\to X})$ is asymptotically bivariate normal as stated in Section 5.2.1, adopting the independence–and–goodness-of-fit framework of Sen and Sen (2014). We work in the scalar case $X \in \mathbb{R}$, $Y \in \mathbb{R}$, which is the setting used throughout our paper.

*Setup and notation*

The true data-generating mechanism is

$$Y = m(X) + \eta, \qquad E[\eta \mid X] = 0. \tag{C.1}$$

Assume $m$ is bijective on its image. For any measurable function $l(Y)$, define

$$\xi = m^{-1}(Y - \eta) - l(Y).$$

For simplicity, let $l(Y) = E[X \mid Y]$, so that $E[\xi \mid Y] = 0$.

Let $(X_1, Y_1), \ldots, (X_Y, n_n)$ be i.i.d. observations generated from (C.1). For $i = 1, \ldots, n$, define

1. $e_i = Y_i - X_i\hat{\beta}_n$ (residual in the true direction),

2. $\varepsilon_i = m(X_i) - X_i\tilde{\beta}_0 + \eta_i$ (misspecification error in the true direction),

3. $d_i = X_i - Y_i\hat{\gamma}_n$ (residual in the incorrect direction),

4. $\delta_i = l(Y_i) - Y_i\tilde{\gamma}_0 + \xi_i$ (misspecification error in the incorrect direction).

Let $\theta(U, V)$ denote the Hilbert–Schmidt independence criterion (Gretton et al., 2008):

$$\theta(U,V) = E[k(U,U')h(V,V')] + E[k(U,U')]E[h(V,V')] - 2E[k(U,U')h(V,V'')], \quad \text{(C.2)}$$

where $(U', V')$ and $(U'', V'')$ are independent copies of $(U, V)$.

We consider the test statistics

$$T_{1b,X\to Y} = \frac{1}{n^2}\sum_{i,j=1}^{n} k_{ij}h_{ij} + \frac{1}{n^4}\sum_{i,j,q,r=1}^{n} k_{ij}h_{qr} - \frac{2}{n^3}\sum_{i,j,q=1}^{n} k_{ij}h_{iq}, \quad \text{(C.3)}$$

$$T_{1c,Y\to X} = \frac{1}{n^2}\sum_{i,j=1}^{n} k'_{ij}h'_{ij} + \frac{1}{n^4}\sum_{i,j,q,r=1}^{n} k'_{ij}h'_{qr} - \frac{2}{n^3}\sum_{i,j,q=1}^{n} k'_{ij}h'_{iq}, \quad \text{(C.4)}$$

where

$$k_{ij} = k(X_i, X_j), \qquad h_{ij} = h(e_i, e_j), \qquad k'_{ij} = k'(Y_i, Y_j), \qquad h'_{ij} = h'(d_i, d_j),$$

and $k, k', h, h'$ are characteristic kernels defined on $\mathbb{R} \times \mathbb{R}$.

Under $(H^{1b}_{X\to Y}, H^{1c}_{Y\to X})$, the corresponding population targets are

$$\theta(X, \varepsilon) \qquad \text{and} \qquad \theta(Y, \delta).$$

Let

$$T_n = \begin{pmatrix} T_{1b,X\to Y} \\ T_{1c,Y\to X} \end{pmatrix}.$$

*Conditions:*

*Condition 1.*

$$A = E[X^2] > 0, \qquad B = E[Y^2] > 0.$$

*Condition 2.* The kernels $k, k', h, h'$ are characteristic, $k, k'$ are continuous, and $h, h'$ are twice continuously differentiable. Their first- and second-order partial derivatives are Lipschitz continuous.

*Condition 3.*

$$E[X^2] < \infty, \quad E[Y^2] < \infty, \quad E[\eta^2] < \infty, \quad E[\xi^2] < \infty, \quad E[X^2\varepsilon^2] < \infty, \quad \text{and} \quad E[Y^2\delta^2] < \infty.$$

*Condition 4.* The kernel–moment conditions required by Lemma 1 in the Supplement of Sen and Sen (2014) hold for both $(X, \varepsilon)$ and $(Y, \delta)$, so that the Taylor expansion remainders below satisfy

$$R_n = o_p(n^{-1/2}), \qquad R_n' = o_p(n^{-1/2}).$$

*Condition 5.* Both

$$m(X) - X\tilde{\beta}_0 \qquad \text{and} \qquad l(Y) - Y\tilde{\gamma}_0$$

are nonconstant with positive probability.

**Theorem C.1.1.** *Suppose that Conditions 1–5 hold. Then under* $(H^{1b}_{X\to Y}, H^{1c}_{Y\to X})$,

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} N_2(0, \Sigma),$$

*where*

$$\theta = \begin{pmatrix} \theta(X, \varepsilon) \\ \theta(Y, \delta) \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \sigma^2_{1b,X\to Y} & \sigma_{1b,1c;X\to Y,Y\to X} \\ \sigma_{1b,1c;X\to Y,Y\to X} & \sigma^2_{1c,Y\to X} \end{pmatrix},$$

*and*

$$\theta(X, \varepsilon) > 0, \qquad \theta(Y, \delta) > 0.$$

*Moreover,*

$$\sigma^2_{1b,X\to Y} = \mathrm{Var}\Big(h^{(0)}_{Y,1}(W_1) + \nu A^{-1} X_1 \varepsilon_1\Big), \tag{C.5}$$

$$\sigma_{1b,1c;X\to Y,Y\to X} = \mathrm{Cov}\Big(h^{(0)}_{Y,1}(W_1) + \nu A^{-1} X_1 \varepsilon_1,\, h^{(0)}_{X,1}(W_1') + \phi B^{-1} Y_1 \delta_1\Big), \tag{C.6}$$

*and*

$$\sigma^2_{1c,Y\to X} = \mathrm{Var}\Big(h^{(0)}_{X,1}(W_1') + \phi B^{-1} Y_1 \delta_1\Big), \tag{C.7}$$

*where* $W_i = (X_i, \varepsilon_i)$, $W_i' = (Y_i, \delta_i)$, *and* $h^{(0)}_{Y,1}, h^{(0)}_{X,1}, \nu, \phi$ *are defined in the proof.*

*Proof.* The least squares estimate

$$\hat\beta_n = A_n^{-1}\left(\frac{1}{n}\sum_{i=1}^n X_i Y_i\right), \qquad A_n = \frac{1}{n}\sum_{i=1}^n X_i^2,$$

admits the following expansion around $\tilde\beta_0$:

$$\begin{aligned}
\sqrt{n}(\hat\beta_n - \tilde\beta_0) &= \sqrt{n}\left(A_n^{-1}\frac{1}{n}\sum_{i=1}^n X_i Y_i - \tilde\beta_0\right) \\
&= \sqrt{n}\left(A_n^{-1}\frac{1}{n}\sum_{i=1}^n X_i\{m(X_i) - X_i\tilde\beta_0 + \eta_i\}\right) \\
&= \{1 + o_p(1)\}\frac{1}{\sqrt{n}}\sum_{i=1}^n A^{-1} X_i \varepsilon_i.
\end{aligned} \tag{C.8}$$

Here we use that $A_n \to A$ almost surely and $A > 0$ by Condition 1. Similarly,

$$\sqrt{n}(\hat\gamma_n - \tilde\gamma_0) = \{1 + o_p(1)\}\frac{1}{\sqrt{n}}\sum_{i=1}^n B^{-1} Y_i \delta_i. \tag{C.9}$$

The population normal equation gives

$$E[X\{m(X) - X\tilde\beta_0\}] = 0,$$

and since $E[\eta \mid X] = 0$,

$$E[X\eta] = E[XE(\eta \mid X)] = 0.$$

Hence $E[X\varepsilon] = 0$. Similarly, because $E[\xi \mid Y] = 0$,

$$E[Y\delta] = E[Y\{l(Y) - Y\tilde{\gamma}_0\}] + E[Y\xi] = 0.$$

By Condition 3, the variances

$$A^{-1}E[X^2\varepsilon^2]A^{-1} \qquad \text{and} \qquad B^{-1}E[Y^2\delta^2]B^{-1}$$

are finite, so by the central limit theorem, $\sqrt{n}(\hat{\beta}_n - \tilde{\beta}_0)$ and $\sqrt{n}(\hat{\gamma}_n - \tilde{\gamma}_0)$ are asymptotically Gaussian.

We next expand $h_{ij} = h(e_i, e_j)$ around $h(\varepsilon_i, \varepsilon_j)$:

$$h_{ij} = h(\varepsilon_i, \varepsilon_j) + (e_i - \varepsilon_i)h_x(\lambda_{ijn}, \tau_{ijn}) + (e_j - \varepsilon_j)h_y(\lambda_{ijn}, \tau_{ijn}),$$

where $(\lambda_{ijn}, \tau_{ijn})$ lies on the line segment joining $(e_i, e_j)$ and $(\varepsilon_i, \varepsilon_j)$. Similarly,

$$h'_{ij} = h'(\delta_i, \delta_j) + (d_i - \delta_i)h'_x(\lambda'_{ijn}, \tau'_{ijn}) + (d_j - \delta_j)h'_y(\lambda'_{ijn}, \tau'_{ijn}),$$

where $(\lambda'_{ijn}, \tau'_{ijn})$ lies on the line segment joining $(d_i, d_j)$ and $(\delta_i, \delta_j)$.

Using

$$e_i - \varepsilon_i = -X_i(\hat{\beta}_n - \tilde{\beta}_0), \qquad d_i - \delta_i = -Y_i(\hat{\gamma}_n - \tilde{\gamma}_0),$$

we obtain the decompositions

$$\begin{pmatrix} T_{1b,X\to Y} \\ T_{1c,Y\to X} \end{pmatrix} = \begin{pmatrix} T^{(0)}_{1b,X\to Y} + (\hat{\beta}_n - \tilde{\beta}_0)T^{(1)}_{1b,X\to Y} + R_n \\ T^{(0)}_{1c,Y\to X} + (\hat{\gamma}_n - \tilde{\gamma}_0)T^{(1)}_{1c,Y\to X} + R'_n \end{pmatrix}, \tag{C.10}$$

where

$$T^{(0)}_{1b,X\to Y} = \frac{1}{n^2}\sum_{i,j=1}^{n} k(X_i,X_j)h(\varepsilon_i,\varepsilon_j) + \frac{1}{n^4}\sum_{i,j,q,r=1}^{n} k(X_i,X_j)h(\varepsilon_q,\varepsilon_r) - \frac{2}{n^3}\sum_{i,j,q=1}^{n} k(X_i,X_j)h(\varepsilon_i,\varepsilon_q),$$
$$T^{(1)}_{1b,X\to Y} = \frac{1}{n^2}\sum_{i,j=1}^{n} k(X_i,X_j)h^{(1)}_{ij} + \frac{1}{n^4}\sum_{i,j,q,r=1}^{n} k(X_i,X_j)h^{(1)}_{qr} - \frac{2}{n^3}\sum_{i,j,q=1}^{n} k(X_i,X_j)h^{(1)}_{iq},$$
$$T^{(0)}_{1c,Y\to X} = \frac{1}{n^2}\sum_{i,j=1}^{n} k'(Y_i,Y_j)h'(\delta_i,\delta_j) + \frac{1}{n^4}\sum_{i,j,q,r=1}^{n} k'(Y_i,Y_j)h'(\delta_q,\delta_r) - \frac{2}{n^3}\sum_{i,j,q=1}^{n} k'(Y_i,Y_j)h'(\delta_i,\delta_q),$$
$$T^{(1)}_{1c,Y\to X} = \frac{1}{n^2}\sum_{i,j=1}^{n} k'(Y_i,Y_j)h'^{(1)}_{ij} + \frac{1}{n^4}\sum_{i,j,q,r=1}^{n} k'(Y_i,Y_j)h'^{(1)}_{qr} - \frac{2}{n^3}\sum_{i,j,q=1}^{n} k'(Y_i,Y_j)h'^{(1)}_{iq},$$

with

$$h^{(1)}_{ij} = -\big\{h_x(\varepsilon_i,\varepsilon_j)X_i + h_y(\varepsilon_i,\varepsilon_j)X_j\big\} \quad \text{and} \quad h'^{(1)}_{ij} = -\big\{h'_x(\delta_i,\delta_j)Y_i + h'_y(\delta_i,\delta_j)Y_j\big\}.$$

By Lemma 1 in the Supplement of Sen and Sen (2014) and Condition 4,

$$R_n = o_p(n^{-1/2}), \qquad R'_n = o_p(n^{-1/2}).$$

We now show that $\theta(X,\varepsilon) > 0$ and $\theta(Y,\delta) > 0$. Since

$$E[\varepsilon \mid X] = m(X) - X\tilde{\beta}_0 + E[\eta \mid X] = m(X) - X\tilde{\beta}_0,$$

and

$$E[\delta \mid Y] = l(Y) - Y\tilde{\gamma}_0 + E[\xi \mid Y] = l(Y) - Y\tilde{\gamma}_0,$$

Condition 5 implies that both conditional means are nonconstant with positive probability. Hence $X \not\perp \varepsilon$ and $Y \not\perp \delta$. Since the kernels are characteristic, HSIC detects dependence, so $\theta(X,\varepsilon) > 0$ and $\theta(Y,\delta) > 0$.

Next, let $W_i = (X_i,\varepsilon_i)$ and $W'_i = (Y_i,\delta_i)$. Then $T^{(0)}_{1b,X\to Y}$ and $T^{(0)}_{1c,Y\to X}$ are V-statistics with finite second moments and means $\theta(X,\varepsilon)$ and $\theta(Y,\delta)$, respectively. Therefore, by the standard asymptotic theory of nondegenerate V-statistics, there exist mean-zero, square-

integrable functions $h^{(0)}_{Y,1}$ and $h^{(0)}_{X,1}$ such that

$$\sqrt{n}\{T^{(0)}_{1b,X\to Y} - \theta(X,\varepsilon)\} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} h^{(0)}_{Y,1}(W_i) + o_p(1), \tag{C.11}$$

$$\sqrt{n}\{T^{(0)}_{1c,Y\to X} - \theta(Y,\delta)\} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} h^{(0)}_{X,1}(W'_i) + o_p(1). \tag{C.12}$$

By the weak law of large numbers for V-statistics,

$$T^{(1)}_{1b,X\to Y} \xrightarrow{p} \nu \quad \text{and} \quad T^{(1)}_{1c,Y\to X} \xrightarrow{p} \phi, \tag{C.13}$$

for finite constants $\nu$ and $\phi$.

Combining (C.10), (C.8), (C.9), (C.11), (C.12), and (C.13), we obtain

$$\begin{aligned}\sqrt{n}\begin{pmatrix} T_{1b,X\to Y} - \theta(X,\varepsilon) \\ T_{1c,Y\to X} - \theta(Y,\delta)\end{pmatrix} &= \begin{pmatrix} \sqrt{n}\{T^{(0)}_{1b,X\to Y} - \theta(X,\varepsilon)\} + \sqrt{n}(\hat{\beta}_n - \tilde{\beta}_0)T^{(1)}_{1b,X\to Y} + \sqrt{n}R_n \\ \sqrt{n}\{T^{(0)}_{1c,Y\to X} - \theta(Y,\delta)\} + \sqrt{n}(\hat{\gamma}_n - \tilde{\gamma}_0)T^{(1)}_{1c,Y\to X} + \sqrt{n}R'_n\end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{h^{(0)}_{Y,1}(W_i) + \nu A^{-1}X_i\varepsilon_i\right\} \\ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{h^{(0)}_{X,1}(W'_i) + \phi B^{-1}Y_i\delta_i\right\}\end{pmatrix} + o_p(1).\end{aligned}$$

Define

$$\psi_i = \begin{pmatrix} h^{(0)}_{Y,1}(W_i) + \nu A^{-1}X_i\varepsilon_i \\ h^{(0)}_{X,1}(W'_i) + \phi B^{-1}Y_i\delta_i\end{pmatrix}.$$

Then $\psi_1, \psi_2, \ldots$ are i.i.d., $E[\psi_i] = 0$, and by Condition 3 they have finite second moments. By the multivariate central limit theorem,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i \xrightarrow{d} N_2(0,\Sigma),$$

where $\Sigma = \mathrm{Var}(\psi_1)$, with entries given in (C.5)–(C.7). This completes the proof. □

### *C.2 Simulation Setup Details*

**Purpose.** To study how CDSP responds under varying degrees of misspecification of the linear model, controlled by degrees $d \in \{1, 1.2, 1.25, 1.3, 1.4, 1.5, 3\}$. Note that all simulation experiments were conducted using a SLURM-based cluster computing environment to efficiently handle the computational demands.

**Inputs.** Fix significance level $\alpha$, number of replications $M$, dataset size $N$, and a large Monte Carlo sample size $N_{\mathrm{MC}}$ used to approximate population quantities.

**Global oracle objects (computed once).** Using a very large Monte Carlo sample from the *true DGP*, $Y = f(X)+\eta$, $f(x) = \beta\,\mathrm{sign}(x-a)\,|x-a|^d$, with $X \sim P_X$ (truncated exponential) and $\eta \sim P_\eta$ (mixture), estimates of the population HSIC quantities $\theta(Y, \delta)$ and $\theta(X, \varepsilon)$ by computing the corresponding test statistics using this large Monte Carlo dataset. These quantities are treated as oracle (population) values.

**Per-replication experiments.** For each misspecification setting $d$ and for $m = 1, \ldots, M$:

1. **Simulate data from the true DGP.** Generate a dataset $\mathcal{D}^{(m)} = \{(X_i^{(m)}, Y_i^{(m)})\}_{i=1}^N$ using $X \sim P_X$, $\eta \sim P_\eta$, and $Y = f(X) + \eta$.

2. **LiNGAM decision.** Fit DirectLiNGAM and record the chosen direction.

3. **Estimate directional detectability indices using bootstrap.** Using bootstrap resampling on $\mathcal{D}^{(m)}$, obtain bootstrap estimates of the directional detectability indices $\widehat{I}^{(m)}_{X\to Y,n}$ and $\widehat{I}^{(m)}_{Y\to X,n}$. Use this to estimate the causal direction via effect-size asymmetry.

4. **Store per-replication test statistics.** Record $T^{(m)}_{1b,X\to Y}$ and $T^{(m)}_{1c,Y\to X}$ for use in estimating the oracle asymptotic standard deviations.

**Population Effect-Size Asymmetry** For each misspecification setting, use the following oracle estimates to evaluate the population effect-size asymmetry assumption via $I_{Y\to X}$ and $I_{X\to Y}$:

- **Oracle standard deviations.** Estimate $\sigma_{1b,X\to Y} \approx \sqrt{N}\,\mathrm{sd}_M\big(T^{(m)}_{1b,X\to Y}\big)$, $\sigma_{1c,Y\to X} \approx \sqrt{N}\,\mathrm{sd}_M\big(T^{(m)}_{1c,Y\to X}\big)$.

- **Population HSIC.** Use the Monte Carlo estimates of $\theta(Y,\delta)$ and $\theta(X,\varepsilon)$ computed from the global oracle objects.

*CDSP* **and LiNGAM Accuracy.** Across $m = 1,\ldots,M$ for each misspecification setting:

- Estimate *CDSP* accuracy as the proportion of replications in which the direction favored by the bootstrap-estimated directional detectability indices $I_{Y\to X,n}$ and $I_{X\to Y,n}$ matches the direction favored by the population quantities $I_{Y\to X}$ and $I_{X\to Y}$.

- Estimate LiNGAM accuracy as the proportion of replications in which LiNGAM identifies the true causal direction.

### *C.3 Severe Model Misspecification Simulation Results*

We present CDSP under severe linear model misspecification with exponent $d = 3$. Under this setting, the effect-size asymmetry assumption no longer holds. The results for this setting are shown in Figure C.1.

In contrast to the moderate misspecification settings, we again observe no overlap between the distributions of $I_{X\to Y,n}$ and $I_{Y\to X,n}$, but now the ordering is reversed. Consequently, direction detection via effect-size asymmetry favors the incorrect direction $Y \to X$ across all Monte Carlo replications as the effect-size asymmetry assumption does not hold.

This severe misspecification setting highlights an important aspect of CDSP. While the method is robust to model misspecification, there exists a threshold at which the effect-size asymmetry assumption breaks down. As the degree of misspecification increases, the overlap

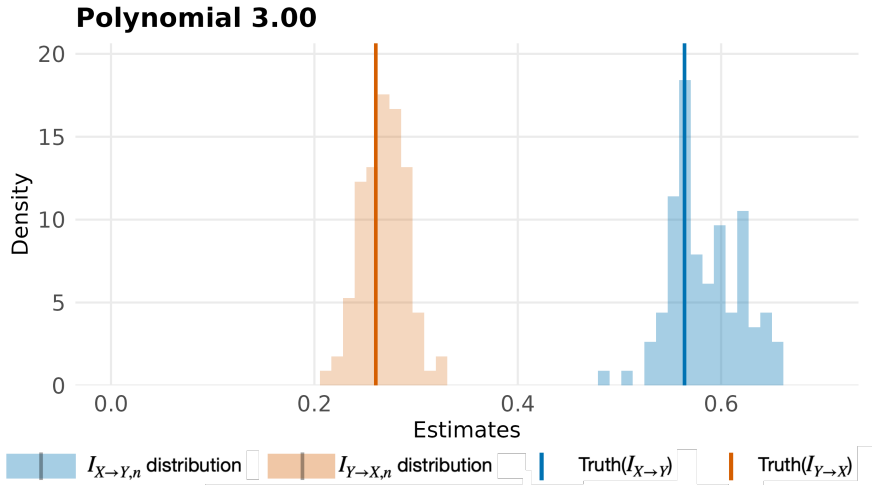



Figure C.1: Distribution of the estimated directional detectability indices $I_{X\to Y,n}$ (blue) and $I_{Y\to X,n}$ (orange) under $d = 3$. Solid vertical lines indicate the corresponding population values ($I_{X\to Y}$ in blue and $I_{Y\to X}$ in orange).

between the $I_{X\to Y,n}$ and $I_{Y\to X,n}$ distributions again diminishes, but in favor of the opposite direction. However, as discussed in Remark 3, this regime is not the primary setting for which CDSP is intended, since practitioners would typically avoid linear-model-based approaches under such severe misspecification.