BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

ESSAYS ON CAUSAL INFERENCE, STRUCTURAL ESTIMATION, AND THEIR APPLICATIONS

by

LIANG ZHONG

B.S., Zhejiang University, 2017
M.A., Boston University, 2019

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2025

© Copyright by
LIANG ZHONG
2025

Approved by

First Reader ______________________________

Hiroaki Kaido, Ph.D
Associate Professor of Economics

Second Reader ______________________________

Jean-Jacques Forneron, Ph.D
Assistant Professor of Economics

Third Reader ______________________________

Daniele Paserman, Ph.D
Professor of Economics

Fourth Reader ______________________________

Kevin Lang, Ph.D
Professor of Economics

ACKNOWLEDGMENTS

I would like to express my heartfelt gratitude to Hiroaki Kaido, who introduced me to econometrics even before I entered the Ph.D. program. He encouraged me to explore topics that intrigued me and provided sharp insights from different perspectives. His influence shaped my approach to research, instilling the importance of rigor in writing and careful evaluation of costs and benefits in decision-making. I am deeply indebted to him for his patience, support, and rigorous training throughout my journey.

I am profoundly grateful to Jean-Jacques Forneron for his invaluable guidance throughout my Ph.D. He introduced me to computational econometrics and consistently helped me structure papers and presentations, even during his busiest times. His mentorship extended beyond research, offering opportunities in co-authorship, teaching, and research assistance. His innumerable insights will continue to shape my work in econometrics.

I sincerely thank Daniele Paserman for his unwavering support since my time as a master's student. His encouragement led me to pursue research I was passionate about, and his mentorship culminated in our co-authorship. He also introduced me to Angela Crema, whose advice during the job market was invaluable. Without Danieles guidance over the past eight years, I would not have been as committed to an academic career.

I am also indebted to Kevin Lang, who generously agreed to write letters for me and joined my committee just before I entered the job market. His experience was invaluable, helping me prepare for flyouts and navigate job negotiations. I also appreciate our conversations during my time as a teaching and research assistant, which expanded my knowledge and led to an ongoing project I am thrilled to be

part of.

I also owe special thanks to the support from Zhongjun Qu and Hsueh-Ling Huynh, who played a pivotal role in shaping my career by assigning me to important research assistant and teaching assistant positions. These opportunities significantly fostered my growth during the most transformative phase of my Ph.D.

I am also grateful to Ivan Fernandez-Val, Pierre Perron, Marc Rysman, Xunkang Tian, Stella Hong, Benjamin Marx, Haoran Pan, Lei Ma, Justin Hong, Qingyuan Chai, Jizhou Liu, Julius Owusu, Panos Toulis, Yan Liu, Matthew Hong, Qifan Han, Juan Ortner, Albert Ma, and Yuzhao Yang for their critical feedback on various projects throughout my Ph.D.

The completion of this dissertation would not have been possible without my familys love and support. I am especially grateful to my wife, Yingqi Wang, and our beloved pets, Cheese and Guoba, whose companionship sustained me through this journey. In moments of doubt and exhaustion, their presence provided the strength and motivation to keep pushing forward. I am also deeply thankful to our parents for their unconditional support and patience, which made this journey possible.

Finally, I want to express my profound appreciation to my closest friends. Zhirui Wang's pursuit of a music career alongside academia has been truly inspiring, and I hope he finds a way to balance both passions successfully. I am also grateful for the opportunity to coauthor with Zixian Liu, a collaboration we had envisioned since the beginning of our Ph.D. I also thank Jiayuan Tian, Ciyue Shen, Yinan Zhang, and Zhenyi Long, who have been an essential part of my social life, making this journey far more enjoyable.

ESSAYS ON CAUSAL INFERENCE, STRUCTURAL ESTIMATION, AND THEIR APPLICATIONS

LIANG ZHONG

Boston University, Graduate School of Arts and Sciences, 2025

Major Professor: Hiroaki Kaido, Associate Professor of Economics

ABSTRACT

This dissertation comprises three chapters that explore two interconnected areas: the development of innovative econometric tools to reduce computational complexities and the analysis of strategic behaviors for actionable policy insights. The first two chapters introduce new statistical approaches that link advanced econometric methods with empirical research, while the third chapter connects economic theory to practical applications by leveraging big data techniques.

When conducting causal inference or designing policy, researchers are often concerned with the existence and extent of interference between units. However, complex correlations across units pose significant challenges for inference. Chapter 1 introduces the pairwise imputation-based randomization test (PIRT), a novel framework for testing interference in experimental settings. PIRT employs a design-based approach, combining unconditional randomization testing with pairwise comparisons to facilitate straightforward implementation and ensure finite-sample validity under minimal assumptions about network structure. To illustrate the method's broad applicability, I apply it to a large-scale experiment by Blattman et al. (2021) in Bogotá, Colombia, which evaluates the impact of hotspot policing on crime using street segments as units of analysis. The re-

sults indicate that increasing police patrolling time in hotspots has a significant displacement effect on violent crime but not on property crime.

Chapter 2, coauthored with Jean-Jacques Forneron at BU, studies the Generalized Method of Moments and the Simulated Method of Moments for estimating structural economic models. These methods are often reported to pose optimization challenges, largely because the corresponding objective functions are non-convex. For smooth problems, Chapter 2 shows that convexity is not required: under a global rank condition involving the Jacobian of the sample moments, certain algorithms are globally convergent. These include a gradient-descent and a Gauss-Newton algorithm with appropriate choice of tuning parameters. The results are robust to 1) non-convexity, 2) one-to-one non-linear reparameterizations, and 3) moderate misspecification. In contrast, Newton-Raphson and quasi-Newton methods can fail to converge because of non-convexity. The condition precludes non-global optima. Numerical and empirical examples illustrate the condition, non-convexity, and convergence properties of different optimizers.

Chapter 3, coauthored with Daniele Paserman at BU and Angela Crema at the University of Rochester, develops a model of discrimination that helps interpret observed outcome differences across groups, conditional on passing a screening test, as taste-based (employer), statistical, or customer discrimination. The framework is applied to examine non-white underrepresentation in the US motion picture industry. Leveraging a novel dataset that provides racial identifiers for the casts of 7,000 motion pictures, we show that, once a movie is produced, non-white movies tend to have higher average revenues and a smaller variance. These findings are consistent with the model if non-white movies face higher production standards.

CONTENTS

Acknowledgements **iv**

Abstract **vi**

List of Tables **xiii**

List of Figures **xv**

List of Symbols and Abbreviations **xvii**

1 Unconditional Randomization Tests for Interference **1**

1.1 Introduction 1

1.2 Setup and Null Hypothesis of Interest 7

1.2.1 Partially Sharp Null Hypothesis 10

1.2.2 Two Technical Challenges for Randomization Tests 15

1.3 Pairwise Imputation-based Randomization Test (PIRT) 19

1.3.1 Pairwise Imputable Statistics 21

1.3.2 Unconditional Randomization Test 26

1.3.3 The Pairwise Comparison-Based p-values 30

1.3.4 Comparison to Previous Literature 34

1.4 Empirical Application 35

1.4.1 Power Comparison of Spatial Interference: A Simulation Study 37

1.4.2 Implementation of PIRTs for Testing the Existence of a Displacement Effect 41

1.4.3 PIRT on Actual Data 42

1.5 Conclusion 48

2 Convexity Not Required: Estimation of Smooth Moment Condition Models 50

2.1 Introduction 50

2.2 GMM Estimation without Convexity 53

2.2.1 Correctly-specified models 58

2.2.2 Misspecified models 62

2.3 Assumption 2 and its relation to the Literature 65

2.4 Applications 75

2.4.1 A pen and pencil example: the MA(1) model 75

2.4.2 Estimation of a Random Coefficient Demand Model Revisited 78

2.4.3 Innovation, Productivity, and Monetary Policy 81

2.5 Conclusion 84

3 Racial Screening on the Big Screen? Evidence from the Motion Picture Industry 86

3.1 Introduction 86

3.2 Institutional background of film production 93

3.3 A model of the screening process 94

3.3.1 Predictions for empirical work 98

3.4 Data 102

3.4.1 Facial classification 102

3.4.2 Additional variables 103

3.4.3 Summary statistics 104

3.5 Results 105

3.5.1 Non-parametric analysis 105

3.5.2 OLS regressions 105

3.5.3 Quantile regressions 107

3.5.4 Robustness 108

3.5.5 The white-nonwhite gap in residual variance 111

3.5.6 Heterogeneity Analysis 112

3.5.7 Producer analysis 113

3.6 Alternative explanation: is the industry surprised? 115

3.7 Conclusion 118

A Supplementary Materials for Chapter One 132

A.1 Proof of the Theorems 132

A.2 Framework for Intersection of Null Hypotheses 134

A.2.1 Testing the Intersection of partially sharp null Hypotheses . . 136

A.3 The Minimization-Based PIRT 137

A.4 Discussion on Some Extreme Cases 140

A.5 Framework to Determine the Boundary of Interference 143

A.5.1 A Valid Procedure to Determine the Neighborhood of Interference 145

A.5.2 Rationale for Using the FWER 147

A.6 Incorporating Covariate Adjustment 148

A.6.1 Investigation on the Power of Incorporating Covariates 149

A.6.2 Robustness of Results to Adjustment Methods 151

A.7 Algorithm for Simulation Exercise 153

B Supplementary Materials for Chapter Two 155

B.1 Proofs for the Main Results 155
  B.1.1 Implications of Assumption 1 155
  B.1.2 Proofs for Section 2.2.1 159
  B.1.3 Proofs for Section 2.2.2 169
B.2 Proofs and additional results for Section 2.3 174
  B.2.1 Additional Results for Over-Identified Models 174
  B.2.2 Proofs 175
B.3 Common methods and their properties 181
  B.3.1 A survey of empirical practice 181
  B.3.2 A brief summary of the Algorithms' properties 183
  B.3.3 Revisting some empirical results 184
B.4 R Code for the MA(1) Example 184
B.5 Additional Simulation, Empirical Results 188
  B.5.1 Estimating an MA(1) model 188
  B.5.2 Demand for Cereal 191
  B.5.3 Impulse Response Matching 191
  B.5.4 Comparison of Rank Condition and Convexity 195
B.6 Additional Material for Algorithms 197
  B.6.1 General overview of Algorithms properties 197
  B.6.2 Implementation of the algorithms 199

C Supplementary Materials for Chapter Three 206

C.1 Proofs 206
C.2 Alternative distributional assumptions 213
  C.2.1 Beta-Binomial distribution 213
  C.2.2 Pareto-Normal distribution 219

C.3 Machine learning algorithm for facial classification 226
C.4 Other Figures 228

Bibliography **231**

Curriculum Vitae **245**

LIST OF TABLES

1.1 Potential Outcome Schedule in the Example 11

1.2 Potential Outcome Schedule Under Partially Sharp Null H_0^0 16

1.3 Traditional Test Statistics Under Partially Sharp Null H_0^0 18

1.4 Constructing a Pairwise Imputable Statistic 23

1.5 Simple Randomization Test in the Example 28

1.6 PIRT in the Example 32

1.7 Pairwise Comparison for PIRT 33

1.8 Descriptive Statistics During the Intervention 46

1.9 p-Values: Testing the Displacement Effect at Different Distances . . . 47

1.10 p-Values: Heterogeneous Patterns Within Violent Crimes 47

2.1 Optimizers considered in (2.1) 55

2.2 MA(1): search paths for NR, GN, BFGS, and L-BFGS-B 78

2.3 Demand for Cereal: performance comparison 80

2.4 Impulse Response Matching: performance comparison 82

3.1 Model Predictions 121

3.2 Summary Statistics 122

3.3 The non-white revenue premium 123

3.4 Robustness: different definitions of “non-white” movies 124

3.5 Robustness to different dependent variables 125

3.6 Conditional residual variance regressions: robustness with respect to different definitions of non-white movies 126

3.7 Heterogeneity Analysis 127
3.8 Producer Analysis 128
3.9 Regressions of first-weekend theaters on number of theaters 129

A.6.1 p-Values: PIRT with Different Specifications 152
A.7.1 Potential Outcome Schedule in the Simulation 154

B.3.1 American Economic Review 2016-2018: GMM and related empirical estimations 182
B.5.1 Demand for Cereal: GN with different learning rates 192
B.5.2 Without reparameterization : GN with different learning rates 193
B.5.3 With reparameterization : GN with different learning rates 194

B.1 Simulation results: Beta-binomial distribution 218
B.2 Simulation results: Pareto distribution 225

LIST OF FIGURES

1.1 Example Network Structure and Distance Matrix 10

1.2 Example Network Structure with Treated, Neighbor, and Control Units 17

1.3 Example Network Structure with Imputable Units 20

1.4 Imputable Neighbor and Control Units for $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D)$ 24

1.5 Imputable Neighbor and Control Units for $T(Y_{\mathbb{I}(D)}(D^{obs}), D^{obs})$ 32

1.6 Map of Experimental Sample and Treatment Conditions 36

1.7 Power Comparison of Testing Methods for Different Hypotheses . . 39

2.1 Relationship between conditions for correctly specified models . . . 70

2.2 MA(1): illustration of non-convexity and the rank condition 77

2.3 Demand for Cereal: distribution of minimized objective values . . . 80

2.4 Demand for Cereal: Gauss-Newton iterations for 5 starting values . . 81

2.5 Impulse Response Matching: distribution of minimized objective values 83

2.6 Impulse Response Matching: Gauss-Newton iterations for 5 starting values 84

3.1 Revenue distribution by number of non-white members 130

3.2 Coefficients are decreasing over quantiles 131

A.6.1 Power Comparison of Testing Methods for Different Covariate Adjustments 150

A.7.1 Unit Distribution 154

B.5.1 GN iterations: different learning rates 188
B.5.2 Non-convexity and the rank condition ($p = 12$, equal weighting $W_n = I_d$) 189
B.5.3 Non-convexity and the rank condition ($p = 12$, optimal weighting $W_n = V_n^{-1}$) 190
B.5.4 GN iterations: equal and optimal weighting, different learning rates . 191
B.5.5 Non-convexity and the rank condition 196

C.1 Output of facial classification 228
D.1 Trend in consumer spending on digital home entertainment, by category 229
D.2 Distribution of log cast star power across racial movie types 230

LIST OF SYMBOLS AND ABBREVIATIONS

$\lambda_{\min}$	The smallest eigenvalue of a square positive semidefinite matrix.
$\lambda_{\max}$	The largest eigenvalue of a square positive semidefinite matrix.
$\sigma_{\min}$	The smallest singular value of a matrix A, defined as $\sigma_{\min}(A) = \sqrt{\lambda_{\min}(A'A)}$.
$\sigma_{\max}$	The largest singular value of a matrix A, defined as $\sigma_{\max}(A) = \sqrt{\lambda_{\max}(A'A)}$.

CHAPTER 1

Unconditional Randomization Tests for Interference

1.1 INTRODUCTION

In social networks or spatial experiments, the outcome for one unit often depends on the treatment assigned to another. This phenomenon is known as interference.[1] Researchers often test for interference based on distance, proximity, and connection strength between units for two reasons: First, identifying the extent of interference helps refine causal inference and model specification;[2] Second, understanding interference facilitates efficient resource allocation, especially when treatments are costly (Brollo et al., 2020). For example, Bond et al. (2012) examine whether spillovers extend beyond immediate friends. Blattman et al. (2021) test spillover distance before assessing hotspot policing impacts in Colombia. Rajkumar et al. (2022) analyze job mobility in relation to link intensity, distinguishing strong from weak ties.

However, testing for interference poses econometric challenges, as large-sample approximations can become intractable due to complex clustering patterns (Kelly, 2021; Blattman et al., 2021). Even in randomized experiments, a valid inference may require assumptions beyond treatment assignment randomness (Aronow, 2012; Pollmann, 2023). Consequently, recent studies (e.g., Bond et al., 2012; Blattman et al., 2021) highlight randomization tests for detecting interference. These studies typically use Fisher randomization tests (FRTs), which

---

[1] Blattman et al. (2021), p. 2027: "Many urban programs are both place-based and vulnerable to spillovers. This includes efforts to improve traffic flow, beautify blighted streets and properties, foster community mobilization, and rezone land use. The same challenges could arise with experiments in social and family networks."

[2] See, for example, Angrist (2014), Sacerdote (2001), Cai et al. (2015), Paluck et al. (2016), Miguel & Kremer (2004), Wang et al. (2024), and Jayachandran et al. (2017).

are not always valid for testing interference (Athey et al., 2018). The core issue lies in the null hypothesis: FRTs test the *sharp null hypothesis* of no effect and rely on imputability, meaning all potential outcomes are assumed known under the null across all assignments (Rosenbaum, 2007; Hudgens & Halloran, 2008). In network settings, this implies potential outcomes remain constant irrespective of which units are treated, excluding both direct effects and interference under the sharp null. Thus, rejecting the sharp null could reflect either nonzero treatment effects or interference, without distinguishing between them.

In this chapter, I introduce the pairwise imputation-based randomization test (PIRT), an unconditional framework for detecting and analyzing interference in experimental settings. PIRT adopts a design-based approach, treating potential outcomes as fixed and using random treatment assignment as the sole source of uncertainty (Abadie et al., 2020, 2022).[3] In the main analysis, I focus on a hypothesis comparing the potential outcomes of units under different treatment assignments when they lie beyond a distance threshold ϵ_s from treated units. The method reassigns treatments while keeping outcomes fixed, calculates carefully designed test statistics, and constructs a novel p-value. A sufficiently small p-value signals evidence against the null hypothesis. PIRT requires minimal assumptions about network structure–making it suitable for dense networks–and relies solely on random assignment, ensuring validity without further assumptions.

More generally, I define *partially sharp null hypotheses* in which only a subset of potential outcomes is assumed known across treatment assignments (Zhang & Zhao, 2023). Testing for interference arises as a special case, requiring isolation of direct treatment effects while assessing whether a unit's outcome depends on

---

[3] As Blattman et al. (2021) note, a design-based approach and randomization inference may be particularly suitable in network contexts where spillover effects are unknown.

others' treatment statuses. Overall, PIRT is a non-parametric, finite-sample valid, and easily implementable method for any partially sharp null hypothesis.[4]

Testing partially sharp null hypotheses with PIRT involves addressing two technical challenges. First, only a subset of potential outcomes is "imputable," meaning their values can be inferred from observed data under the partially sharp null hypothesis. For example, under a partially sharp null hypothesis of no peer effects on non-treated units in a social network, outcomes can only be imputed for non-treated units; outcomes for treated units remain unknown. Second, the set of units with imputable outcomes changes with each treatment assignment, as the non-treated units vary across assignments. Together, these challenges complicate the direct application of traditional randomization inference methods, highlighting the need for specialized approaches.

To address the first challenge, I propose a class of statistics termed *pairwise imputable statistics*, each defined with two treatment assignment arguments. The first assignment identifies the imputable units, while the second determines how these units are grouped or compared. These statistics resemble conventional test statistics defined by Imbens & Rubin (2015) but are restricted to imputable units as specified by the partially sharp null under both assignments. Despite this restriction, pairwise imputable statistics accommodate various commonly used test statistics. For example, a difference-in-means estimator might compare imputable individuals who have treated friends to those who do not. Here, the first assignment determines the imputable individuals included in the calculation, while the second assignment defines the groupings, similar to conventional test statistics.

To tackle the second challenge, I draw on recent advances in selective infer-

---

[4] It is finite-sample exact, meaning the probability of a false rejection in finite samples does not exceed the user-prescribed nominal rate (Pouliot, 2024).

ence (Wen et al., 2023; Guan, 2023) and construct PIRT p-values via pairwise comparisons of two pairwise imputable statistics. The first statistic uses the randomized assignment to select imputable units, while the observed assignment defines groupings and comparisons. Conversely, the second statistic selects imputable units based on the observed assignment, with groupings and comparisons determined by the randomized assignment. The validity of this procedure relies on the symmetry of these pairwise comparisons, analogous to the conformal lemma of Guan (2023).

To illustrate PIRT's applicability, I apply it to a large-scale experiment by Blattman et al. (2021) evaluating a policing strategy that concentrated resources on high-crime hotspots in Bogotá, Colombia, using street segments as units. I assess the policy's overall effectiveness and examine criminal behavior and incentives by testing for interference–such as crime displacement or deterrence–in nearby neighborhoods.[5] The authors report significant displacement effects of increased police patrol on property crime but not on violent crime. However, using PIRT to specifically test against no displacement, I find, contrary to Blattman et al. (2021), a marginally significant displacement effect on violent crime at the 10% level and an insignificant effect on property crime.[6] This result could reshape our understanding of criminal behavior and inform welfare analysis, especially if more severe violent crime warrants stricter interventions.

A simulation study calibrated to this dataset further demonstrates the strong empirical properties of PIRT compared to existing methods. In particular, I test for

---

[5] This assumes that interactions pass through neighboring units, resulting in spillover effects.

[6] I also propose a multiple hypothesis testing adjustment that ensures control of the family-wise error rate (FWER) when defining the "neighborhood" of interference. This adjustment is particularly useful when spillover effects are positive, as it helps policymakers design cost-effective interventions. Conversely, if spillover effects are negative, identifying their range aids in evaluating the overall effectiveness of the policy.

displacement effects, where interference causes outcomes to “spill over” to neighboring units. PIRT at the α rejection level successfully controls type I error rates, demonstrating robustness under worst-case scenarios. In contrast, classical FRT may over-reject under partially sharp null hypotheses. Regarding power, PIRT at the α rejection level outperforms competing alternative methods, which is especially valuable in network analysis, where data collection is costly and interference effects are subtle (Taylor & Eckles, 2018; Breza et al., 2020). However, there exists a trade-off between ease of implementation and conservatism under the null, as PIRT may exhibit conservatism in some cases.

Literature Review This chapter contributes to three strands of literature. First, it advances network analysis. Since the seminal work of Manski (1993), several studies have adopted model-based approaches relying on parametric assumptions (Sacerdote, 2001; Bowers et al., 2013; Toulis & Kao, 2013; Graham, 2017; de Paula et al., 2018). These papers typically impose specific structures on networks and must carefully handle the high dimensionality of network interactions. In contrast, my method is non-parametric and leverages the null hypothesis to reduce dimensionality.

Second, this chapter contributes to design-based causal inference methods under interference. Two main frameworks for causal inference under interference are the Fisherian and Neymanian perspectives (Li et al., 2018). The Neymanian approach emphasizes randomization-based unbiased estimation and variance calculations (Hudgens & Halloran, 2008; Aronow & Samii, 2017; Pollmann, 2023), typically using asymptotic normal approximations and often requiring sparse networks or local interference assumptions.[7]

---

[7] Also see Basse & Airoldi (2018), Viviano (2022), Wang et al. (2023), Vazquez-Bare (2023), Leung

In contrast, this chapter adopts the Fisherian perspective, focusing on detecting causal effects using finite-sample valid, randomization-based tests (Dufour & Khalaf, 2003; Lehmann & Romano, 2005; Rosenbaum, 2020). Acknowledging the limitations of FRTs for testing interference, prior literature has introduced conditional randomization tests (CRTs), which restrict testing to a conditioning event–a subset of units and assignments for which the null hypothesis is sharp.[8] However, many CRT methods are tailored to specific scenarios, such as clustered interference (Basse et al., 2019, 2024), limiting their generalizability. Additionally, designing conditioning events that can detect interference is challenging and often leads to power loss (Puelz et al., 2021). Furthermore, implementing CRTs under general interference can be computationally demanding, requiring significant resources. This chapter builds on this foundation by introducing an alternative approach that applies broadly, is straightforward to implement, and remains valid even when designing conditioning events is difficult. Confidence intervals for specific causal parameters can then be constructed by inverting these tests.[9]

Finally, this chapter contributes to the literature by extending randomization testing beyond sharp null hypotheses. While the primary focus is on partially sharp null hypotheses defined by distance measures, the principles of PIRT seem generalizable beyond network contexts. Since Neyman et al. (2018) acknowledged that FRTs are limited to testing sharp null hypotheses, researchers have developed various strategies to address weak nulls (Ritzwoller et al., 2025). For example,

---

(2020), Leung (2022), and Shirani & Bayati (2024).

[8] See, for example, Aronow (2012), Athey et al. (2018), Basse et al. (2019), Puelz et al. (2021), Zhang & Zhao (2021), Basse et al. (2024), and Hoshino & Yanagi (2023).

[9] Furthermore, randomization-based methods can be integrated with model-based frameworks, such as the linear-in-means model (Manski, 1993), to increase power or broaden applicability beyond randomized experiments while preserving test validity (Wu & Ding, 2021; Basse et al., 2024; Borusyak & Hull, 2023).

Ding et al. (2016), Li et al. (2016), and Zhao & Ding (2020) examine randomization tests for the null hypothesis of no average treatment effect, while Caughey et al. (2023) validate these tests under bounded null hypotheses. Zhang & Zhao (2021) construct CRTs for partial sharp nulls, applying an approach similar to Athey et al. (2018) and Puelz et al. (2021) in time-staggered adoption designs. To my knowledge, PIRT is the first method addressing partially sharp null hypotheses through unconditional randomization testing.

The rest of the chapter is structured as follows. Section 1.2 introduces the general setup and establishes all necessary notation. Section 1.3 presents the PIRT procedure, which includes the pairwise imputable statistics and the p-value based on pairwise comparisons. Section 1.4 applies the method to a large-scale policing experiment in Bogotá, Colombia, with Section 1.4.1 reporting the results of a Monte Carlo experiment calibrated to this setting. Finally, Section 1.5 concludes. The appendix provides additional empirical and theoretical results as well as proofs.

1.2 SETUP AND NULL HYPOTHESIS OF INTEREST

Consider N units indexed by $i \in \{1, 2, \ldots, N\}$, connected through an undirected network observed by the researcher. The researcher is interested in understanding the extent of interference based on factors such as distance, neighboring units, and connection strength, which are captured by an $N \times N$ proximity matrix G. The (i, j)-th component $G_{i,j} \geq 0$ represents a “distance measure" between units i and j, which can be either a continuous or discrete variable. I normalize $G_{i,i} = 0$ for all $i = 1, 2, \ldots, N$, and assume $G_{i,j} > 0$ for all $i \neq j$. This distance measure is context-specific:[10]

---

[10] Researchers can also define distance in product space, particularly in the context of firms selling differentiated products. Here, units represent products, and $G_{i,j}$ can be the Euclidean dis-

Example 1 (Spatial distance)**.** *In settings where units interact locally through shared space, such as street segments in a city (Blattman et al., 2021),* $G_{i,j}$ *represents the spatial distance between units* i *and* j*.*

Example 2 (Network distance)**.** *In social network settings, such as friendships on Facebook (Bond et al., 2012),* $G_{i,j}$ *measures the distance between units* i *and* j*, where* $G_{i,j} = 1$ *for friends,* $G_{i,j} = 2$ *for friends of friends, and* $G_{i,j} = \infty$ *if* i *and* j *are not connected. This framework accommodates disconnected networks and captures partial interference, such as cluster-level interference (Sobel, 2006; Basse et al., 2019).*

Example 3 (Link intensity)**.** *Researchers may observe not only whether two units are linked but also the intensity of the link* $int_{i,j}$*, such as frequency of interaction or volume of email correspondence (Goldenberg et al., 2009; Bond et al., 2012; Rajkumar et al., 2022). Building on the classic study by Granovetter (1973), one might examine how interference differs across weak and strong ties, defined by this intensity measure. Let* $\bar{int} = \max_{i,j \in \{1,\ldots,N\}} int_{i,j}$*, and define* $G_{i,j} = \bar{int} - int_{i,j}$*. In this way, an increase in* $G_{i,j}$ *implies a weaker connection, analogous to Examples 1 and 2.*

In this chapter, I focus on experimental settings where treatment assignment is random and follows a known probability distribution P, where $P(d) = \Pr(D = d)$ is the probability that the treatment assignment D equals d. Let X represent the collected pre-treatment characteristics, such as age and gender, which can be used to control for unit heterogeneity. However, I do not attempt to evaluate their direct effects on the outcome. The probability distribution may or may not depend on covariates X. In cases of complete or cluster randomization, it does not depend on X, while in stratified or matched-pair designs, it does.

---

tance between them in a multi-dimensional space of product characteristics, as in Pollmann (2023). This measure is useful for defining market boundaries, such as when a merger authority assesses whether two products belong to the same relevant market.

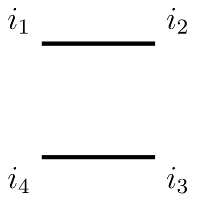
I adopt the potential outcomes framework with a binary treatment assignment vector $D = (D_1, \ldots, D_N) \sim P$, where $D \in \{0,1\}^N$ and $D_i \in \{0,1\}$ denotes unit i's treatment. Let $Y(d) = (Y_1(d), \ldots, Y_N(d)) \in \mathbb{R}^N$ be the potential outcomes under treatment assignment d, where the potential outcome of unit i is $Y_i(d) = Y_i(d_1, \ldots, d_N)$. This allows unit i's potential outcome to depend on the treatment assignment of unit j, violating the classic *Stable Unit Treatment Value Assumption (SUTVA)* proposed by Cox (1958), and accommodating cases where spatial or network interference exists. However, the distance measure between treatment and individuals is unaffected by the treatment.

Throughout the chapter, I assume that the following are observed: 1) the realized vector of treatments for all units, denoted by D^{obs}; 2) the realized outcomes for all units, denoted by $Y^{obs} \equiv Y(D^{obs}) = (Y_1(D^{obs}), \ldots, Y_N(D^{obs}))$; 3) the proximity matrix G; 4) the covariates X; and 5) the probability distribution of the treatment assignment P. I adopt a design-based inference approach, where D is treated as random, while G, X, P, and the unknown potential outcome schedule $Y(\cdot)$ are considered fixed. For simplicity in notation, these elements will not be treated as arguments of functions in the rest of the chapter. To illustrate these notations, consider the following running example.

Running Example. Consider four street segments, where two segments are adjacent if they are connected, as shown in Figure 1.1. Units i_1 and i_2 are connected, forming one area, while units i_3 and i_4 are connected, forming another. For simplicity, the distance between units in the same area is set to 1. In practice, the distance between units in different areas could be up to infinity but for the sake of this example, assume it is 2.

Suppose the outcome of interest, Y, is the total number of crimes over a year,

Figure 1.1: Example Network Structure and Distance Matrix



(a) Network Structure

$$G = \begin{pmatrix} 0 & 1 & 2 & 2 \\ 1 & 0 & 2 & 2 \\ 2 & 2 & 0 & 1 \\ 2 & 2 & 1 & 0 \end{pmatrix}$$

(b) Distance Matrix

Notes: Panel (a) presents the network structure of four units, while panel (b) displays the corresponding distance matrix.

and a random treatment D is applied to increase policing in one unit. Assume the treatment is randomly assigned with $P(d) = 1/4$ for each possible assignment. Let the observed treatment be $D^{obs} = (1, 0, 0, 0)$ and the observed outcomes $Y^{obs} = (2, 4, 3, 1)$.

Table 1.1 illustrates the potential outcome schedule under the design-based framework for all assignments that have positive probability. The first row corresponds to the observed dataset. Although all potential outcomes are fixed values, only the outcomes under the observed treatment are known. In general, since potential outcomes can depend on assignments across all units, there could theoretically be up to 2^N potential outcomes.

1.2.1 Partially Sharp Null Hypothesis

The term "partially sharp null" was first introduced by Zhang & Zhao (2023), and I begin by providing a formal definition of the partially sharp null hypothesis.

Definition 1 (Partially sharp null hypothesis)**.** *A partially sharp null hypothesis holds*

Table 1.1: Potential Outcome Schedule in the Example

Assignment D	Potential Outcome Y_i			
	i_1	i_2	i_3	i_4
$(1,0,0,0)$	2	4	3	2
$(0,1,0,0)$	?	?	?	?
$(0,0,1,0)$	?	?	?	?
$(0,0,0,1)$	?	?	?	?

Notes: The table shows the potential outcome schedule under the design-based view. The first row represents the observed assignment D^{obs}, while potential outcomes denoted by ? are unobserved values.

if there exists a collection of subsets $\{\mathcal{D}_i\}_{i=1}^N$*, where each* $\mathcal{D}_i \subset \{0,1\}^N$*, such that*

$$H_0 : Y_i(d) = Y_i(d') \text{ for all } i \in \{1, \ldots, N\}, \text{ and any } d, d' \in \mathcal{D}_i.$$

The partially sharp null hypothesis reduces dimensionality by restricting potential outcomes only across certain subsets of assignments. The set $\mathcal{D}_i$ can vary across units and is always a strict subset of $\{0,1\}^N$, and thus offers greater flexibility than the sharp null hypothesis, which corresponds to the case where $\mathcal{D}_i = \{0,1\}^N$. For instance, researchers can specify $\mathcal{D}_i$ based on an exposure mapping–a function linking treatment assignments to exposure levels–to test outcome constancy within each exposure level, especially when concerned about potential misspecification (Hoshino & Yanagi, 2023).

More generally, researchers can define alternative forms of $\mathcal{D}_i$ that reflect specific hypotheses and research contexts, including cases where the null hypothesis is expressed as the intersection of multiple $\mathcal{D}_i$ sets (Owusu, 2023; Puelz et al., 2021). Appendix A.2 discusses extensions of the current framework to these more general and complex hypotheses. Although the method introduced in this chapter is applicable to any partially sharp null hypothesis, I specifically focus on cases where

$\mathcal{D}_i$ is defined based on a distance measure.

Definition 2 (Distance interval assignment set)**.** *For a unit $i \in \{1, \ldots, N\}$ and a given distance ϵ_s, the distance interval assignment set is defined as*

$$\mathcal{D}_i(\epsilon_s) \equiv \left\{ d \in \{0,1\}^N : \sum_{j=1}^{N} 1\{G_{i,j} \leq \epsilon_s\} d_j = 0 \right\}.$$

When $d \in \mathcal{D}_i(\epsilon_s)$, unit i is said to be in the distance interval (ϵ_s, ∞).

This definition involves two key concepts: $\mathcal{D}_i(\epsilon_s)$ and the interval (ϵ_s, ∞), both of which are specific to unit i. The distance interval assignment set $\mathcal{D}_i(\epsilon_s)$ maps a distance ϵ_s to a set of treatment assignments where unit i is at least a distance ϵ_s away from any treated units. For any $\epsilon_s \geq 0$, since $G_{i,i} = 0$, it follows that $1\{G_{i,i} \leq \epsilon_s\} = 1$, implying that unit i is untreated ($d_i = 0$) for any assignment $d \in \mathcal{D}_i(\epsilon_s)$. Specifically, when $\epsilon_s = 0$, all $G_{i,j}$ for $i \neq j$ are positive, which ensures that $1\{G_{i,j} \leq \epsilon_s\} = 0$. As a result, there is no restriction on the treatment status of other units d_j, and $\mathcal{D}_i(0)$ includes all treatment assignments d where $d_i = 0$, while allowing others to be treated.[11]

The distance interval assignment set $\mathcal{D}_i(a)/\mathcal{D}_i(b)$ corresponds to treatment assignments where unit i is within the distance interval $(a, b]$. For any treatment assignment d, the set $\{i : d \in \mathcal{D}_i(a)/\mathcal{D}_i(b)\}$ contains all units that fall within the distance interval $(a, b]$ relative to treated units.

Using the concept of distance interval assignment sets, I now define the partially sharp null hypothesis of interference based on distance.

Definition 3 (Partially sharp null hypothesis of interference on distance $\epsilon_s \geq 0$)**.**

---

[11] For any $\epsilon_s < 0$, since $G_{i,j} \geq 0$ for all i, j, we have $1\{G_{i,j} \leq \epsilon_s\} = 0$, meaning that $\mathcal{D}_i(\epsilon_s) = \{0,1\}^N$, where all treatment assignments are included.

The partially sharp null hypothesis of interference on distance $\epsilon_s \geq 0$ *is defined as*

$$H_0^{\epsilon_s} : Y_i(d) = Y_i(d') \text{ for all } i \in \{1, \ldots, N\}, \text{ and any } d, d' \in \mathcal{D}_i(\epsilon_s).$$

Under $\mathcal{D}_i(\epsilon_s)$, all units within ϵ_s distance of unit i, as well as unit i itself, are not treated. Hence, this hypothesis asserts that no interference occurs beyond distance ϵ_s, meaning the potential outcomes for unit i remain unchanged for any treatment assignment where unit i is at least a distance ϵ_s away from all treated units. Under this null hypothesis, the potential outcomes for unit i can be imputed for treatment assignment vectors that satisfy this distance condition, allowing for a partial imputation of outcomes. The interpretation of distance here is context-specific and depends on the nature of the interference in the particular application.

Example 1 (Spatial distance continued)**.** *In a setting where units represent street segments, for a given spatial distance* ϵ_s *(e.g., 500 meters),* $\mathcal{D}_i(\epsilon_s)$ *consists of all treatment assignments where unit* i *is at least 500 meters away from any treated street segments. The partially sharp null hypothesis* $H_0^{\epsilon_s}$ *tests whether spillover effects occur on an untreated unit located 500 meters away from any treated units.*

Example 2 (Network distance continued)**.** *Consider two schools, each with 100 students, where the goal is to test for cluster interference within schools. We assume that students within the same school are 100 units apart from each other and are infinitely distant from students in the other school. Setting* $\epsilon_s = 0$*, we test for interference within schools. Cluster interference is present if students' outcomes are affected by treatment assignments in their own school but not in the other school.*[12]

Example 3 (Link intensity continued)**.** *Consider a scenario where units represent indi-*

---

[12] Setting $\epsilon_s = 101$ would test for interference across schools but such a test may lack power in practice, as noted by Puelz et al. (2021).

viduals with cell phones, and the intensity of their connection is measured by the number of text messages exchanged, with a maximum of 50 messages per week. We define the "distance" between two individuals as 50 *minus the number of messages exchanged. For* $\epsilon_s = 40$*,* $\mathcal{D}_i(\epsilon_s)$ *represents all treatment assignments where unit* i *has exchanged fewer than 10 messages with any treated units. The partially sharp null hypothesis* $H_0^{\epsilon_s}$ *tests whether interference occurs for an untreated unit that has exchanged fewer than 10 messages with treated units.*

The null hypothesis defined in Definition 3 is useful for assessing the existence or extent of interference within a network, as researchers are often interested in whether interference occurs beyond a certain distance ϵ_s. If $\epsilon_s > 0$, researchers can use this approach to identify the neighborhood of interference or to find a suitable comparison group for subsequent estimation.

Comparison to the Traditional t-Test. The traditional t-test compares units at varying distances from treated units but faces two key challenges: First, units' distances to treated units are not random even under random assignment, potentially causing bias without additional assumptions (Aronow, 2012; Pollmann, 2023); Second, large-sample approximations become difficult due to complex clustering patterns (Kelly, 2021; Blattman et al., 2021). In contrast, the partially sharp null hypothesis from Definition 3 directly evaluates the same unit's potential outcomes at distances greater than ϵ_s from treated units. Its advantage is that it requires only random treatment assignment and avoids biases from comparing outcomes across potentially non-comparable units.

If $\epsilon_s = 0$, we test the partially sharp null hypothesis of no interference as $\mathcal{D}_i(0)$ consists of all treatment assignments where $d_i = 0$, meaning the unit is untreated. Treatment assignments where $d_i = 1$ are excluded, ensuring that the hypothesis solely focuses on spillover effects. In contrast, the traditional sharp null hypothesis includes potential outcomes where $d_i = 1$, which also involves direct treatment effects. We can simplify H_0^0 further, as illustrated in the following running example.

Running Example Continued. Suppose researchers want to test for the existence of spillover effects using the partially sharp null hypothesis in Definition 3 with $\epsilon_s = 0$:

$$H_0^0 : Y_i(d) = Y_i(d') \text{ for all } i \in \{1, \ldots, N\},$$

$$\text{and any } d, d' \in \{0,1\}^N \text{ such that } d_i = d'_i = 0.$$

Throughout the chapter, I use the above H_0^0 for illustration in the running example. This hypothesis implies that the potential outcome for any untreated unit i remains unchanged regardless of the treatment assignments of other units. The potential outcome schedule under H_0^0 is shown in Table 1.2.

As shown in Table 1.2, the null hypothesis H_0^0 allows us to impute many of the previously missing potential outcomes. For example, since we observe the outcome when unit i_2 is not treated, we can impute other outcomes as long as unit i_2 remains untreated. Consequently, the outcome for i_2 when either unit i_3 or i_4 is treated is also 4.

1.2.2 Two Technical Challenges for Randomization Tests

As discussed in Zhang & Zhao (2023), the partially sharp null hypothesis implies that only a subset of potential outcomes remain unknown. Although this reduces the number of missing outcomes, technical challenges persist. As illustrated by

Table 1.2: Potential Outcome Schedule Under Partially Sharp Null H_0^0

Assignment D	Potential Outcome Y_i			
	i_1	i_2	i_3	i_4
$(1,0,0,0)$	2	4	3	2
$(0,1,0,0)$	?	?	3	2
$(0,0,1,0)$	?	4	?	2
$(0,0,0,1)$	?	4	3	?

Notes: The table shows the potential outcome schedule with the partially sharp null hypothesis under Definition 3 for the toy example. Assignment D includes all potential assignments, with the first row representing the observed assignment D^{obs}. Potential outcomes marked in ? are non-imputable values under the partially sharp null.

Table 1.2, the potential outcome schedule under H_0^0 still contains missing values, complicating the use of traditional methods.

Traditional Test Statistics. In practice, researchers often specify a distance ϵ_c such that units farther than ϵ_c from treated units are assumed to experience no interference. For instance, in a spatial setting, we might assume that no interference occurs for units more than $\epsilon_c = 1,000$ meters away. For cluster interference, we might assume that no spillover occurs once ϵ_c exceeds the maximum distance within a cluster, indicating no interference across clusters.

A natural test statistic compares units within the distance interval $(\epsilon_s, \epsilon_c]$ to the treated group, while using units in the distance interval (ϵ_c, ∞) as a pure control group. The idea behind ϵ_c is to identify a threshold beyond which the influence of the treatment is negligible, allowing researchers to separate units likely to be impacted by interference from those that serve as clean controls. If the researcher has no prior value for ϵ_c, Section A.5.2 proposes a sequential testing procedure to help select an appropriate ϵ_c. Even if ϵ_c is misspecified and does not provide a clean control group, the proposed testing procedure remains valid, though it may reduce test power (Basse et al., 2024).

For example, consider the difference in means with control distance ϵ_c:

$$T(Y(D^{obs}), D) = \underbrace{\bar{Y}(D^{obs})_{\{i:D\in\mathcal{D}_i(\epsilon_s)/\mathcal{D}_i(\epsilon_c)\}}}_{\text{Mean of neighbor}} - \underbrace{\bar{Y}(D^{obs})_{\{i:D\in\mathcal{D}_i(\epsilon_c)\}}}_{\text{Mean of control}},$$

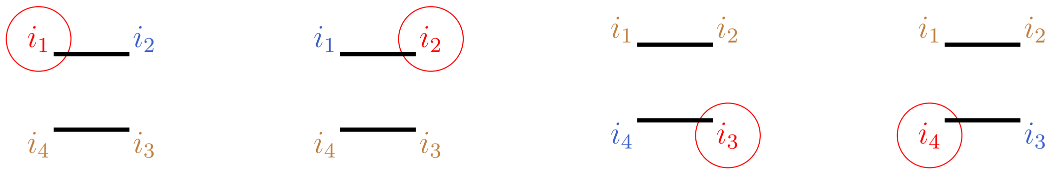
where for sets $A_i \subset \{0,1\}^N$, we have

$$\bar{Y}(D^{obs})_{\{i:D\in A_i\}} = \sum_{i=1}^{N} 1\{D \in A_i\} Y_i(D^{obs}) / \sum_{i=1}^{N} 1\{D \in A_i\},$$

In particular, $A_i = \mathcal{D}_i(\epsilon_s)/\mathcal{D}_i(\epsilon_c)$ corresponds to the distance interval $(\epsilon_s,\, \epsilon_c]$, while $A_i = \mathcal{D}_i(\epsilon_c)$ corresponds to $(\epsilon_c,\, \infty)$. The difference-in-means estimator is widely used in the literature(see, e.g., Basse et al., 2019; Puelz et al., 2021).

Running Example Continued. For the rest of the discussion in the running example, I would use $\epsilon_c = 1$. Therefore, there are two relevant distance intervals for the difference-in-means estimator: $(0, 1]$ and $(1, \infty)$. Figure 1.2 shows how these intervals change with different treatment assignments.

Figure 1.2: Example Network Structure with Treated, Neighbor, and Control Units



(a) Treated Unit: i_1 **(b)** Treated Unit: i_2 **(c)** Treated Unit: i_3 **(d)** Treated Unit: i_4

Notes: Units with red circles are treated, units in blue are neighbors in the interval $(0, 1]$, and units in brown are control units in the interval $(1, \infty)$.

Applying traditional test statistics, such as the difference-in-means estimator, can be problematic when some potential outcomes remain unknown under H_0^0. Although the first row can be computed as $4 - (3 + 2)/2 = 1.5$, Table 1.3 shows that test statistics under non-observed treatment assignments still involve missing values. This occurs because randomization requires knowledge of all $Y_i(d)$ values for the relevant assignment. This renders FRT inapplicable under the partially sharp null hypothesis and highlights two specific challenges that persist in more general settings:

Table 1.3: Traditional Test Statistics Under Partially Sharp Null H_0^0

Assignment D	Potential Outcome Y_i				$T(Y(D^{obs}), D)$
	i_1	i_2	i_3	i_4	
$(1, 0, 0, 0)$	2	4	3	2	1.5
$(0, 1, 0, 0)$	?	?	3	2	?
$(0, 0, 1, 0)$	?	4	?	2	?
$(0, 0, 0, 1)$	?	4	3	?	?

Notes: The table shows the potential outcome schedule under the partially sharp null hypothesis for the example. Assignment D includes all potential assignments, with the first row representing the observed assignment D^{obs}. Potential outcomes marked in red question marks are non-imputable under the partially sharp null.

First, only a subset of potential outcomes can be observed or imputed. For example, under H_0^0, if unit i_2 is treated, the hypothesis provides no information about the potential outcomes of unit i_1, leaving the potential outcomes for both i_1 and i_2 missing.

Second, the set of units with imputable outcomes depends on the treatment assignment. For instance, if unit i_3 is treated instead, the missing values now belong to i_1 and i_3, differing from other assignments.

The remainder of the chapter focuses on addressing these two technical challenges.

1.3 PAIRWISE IMPUTATION-BASED RANDOMIZATION TEST (PIRT)

For simplicity, I initially fix ϵ_s and ϵ_c, deferring discussion of their selection until the end of this section. For each treatment assignment d, I focus on the units imputable under $H_0^{\epsilon_s}$ given the observed information.

Definition 4 (Imputable units)**.** *Given a treatment assignment $d \in \{0,1\}^N$ and a partially sharp null hypothesis $H_0^{\epsilon_s}$,*

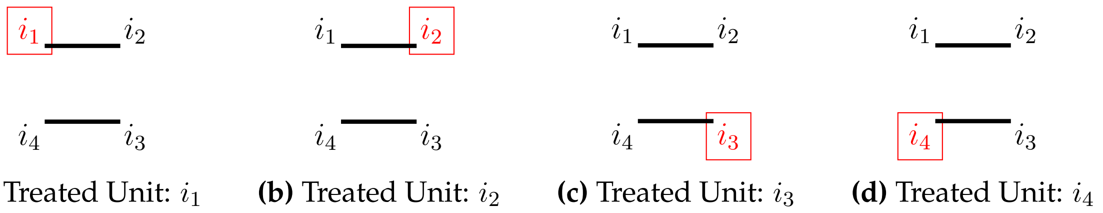
$$\mathbb{I}(d) \equiv \{i \in \{1,\ldots,N\} : d \in \mathcal{D}_i(\epsilon_s)\} \subseteq \{1,\ldots,N\}$$

is called the set of imputable units under treatment assignment d.

The set of imputable units is the subset of units for which imputation is possible, corresponding to those in the distance interval (ϵ_s, ∞) under the partially sharp null hypothesis $H_0^{\epsilon_s}$. It shares a similar spirit with the "super focal units" in Owusu (2023). Specifically, given the observed treatment D^{obs}, the set $\mathbb{I}(D^{obs})$ includes all units with an imputable observed outcome. Units outside this set provide no additional information because their observed outcomes cannot be imputed to other treatment assignments under the partially sharp null. For example, if $\epsilon_s = 0$, then $\mathcal{D}_i(\epsilon_s)$ includes all assignments d where $d_i = 0$, meaning $\mathbb{I}(D^{obs})$ consists of all units not treated under D^{obs}.

Running Example Continued. Under H_0^0, the imputable units for a treatment assignment d can be expressed as $\mathbb{I}(d) \equiv \{i \in \{1,\ldots,N\} : d_i = 0\}$. That is, under the null hypothesis of no interference, all non-treated units are imputable. For example, as shown in Figure 1.3, when unit i_1 is treated, units i_2 to i_4 belong to the imputable set, and when unit i_2 is treated, units i_1, i_3, and i_4 are imputable. It is worth noting that this is a special case where all untreated units are imputable. In more general settings, this depends on the ϵ_s in the null hypothesis.

Figure 1.3: Example Network Structure with Imputable Units



(a) Treated Unit: i_1 **(b)** Treated Unit: i_2 **(c)** Treated Unit: i_3 **(d)** Treated Unit: i_4

Notes: Treated units are marked with red rectangles, while imputable units are shown in black.

As shown in Figure 1.3, generally, $\mathbb{I}(d) \neq \mathbb{I}(d')$ for different assignments d and d'. For example, when testing for spillover effects among friends, the set of friends affected will change with different treatment assignments due to varying social connections.

In practice, $\mathbb{I}(D^{obs})$ could sometimes be empty, depending on the network structure and the specific partially sharp null hypothesis. If no units meet the required criteria (i.e., $\mathbb{I}(D^{obs})$ is empty), one approach is to reject the null hypothesis α percent of the time, in line with the desired significance level. This ensures control of the test's size, even in cases where the imputable set is empty. However, to achieve power in such cases, additional data or a different study design may be necessary. See Appendix A.4 for further discussion.

The set of imputable units can also be defined under the sharp null hypothesis, though in this case, $\mathbb{I}(d) = \{1, \ldots, N\}$ for any assignment d, meaning all units are imputable under the sharp null. Therefore, there has been less focus on the imputable units set in the randomization tests literature. To help define the test statistics later, I further define the following.

Definition 5 (Imputable outcome vector). *For any treatment assignment $d \in \{0,1\}^N$ and a partially sharp null hypothesis $H_0^{\epsilon_s}$,*

$$Y_{\mathbb{I}(d)} \equiv \{Y_i\}_{i \in \mathbb{I}(d)}$$

is called the imputable outcome vector for the treatment assignment d, with each component representing the potential outcome for the units in $\mathbb{I}(d)$. When the value of Y is determined by an alternative treatment assignment d', we denote

$$Y_{\mathbb{I}(d)}(d') \equiv \{Y_i(d')\}_{i \in \mathbb{I}(d)}$$

as the imputable outcome vector for d, with each component representing the potential outcome under the alternative treatment assignment d' for the units in $\mathbb{I}(d)$.

Two factors influence the imputable outcome vector. First, the value of the potential outcome depends on d'. For example, when $d' = D^{obs}$, $Y(d') = Y^{obs}$ is the observed outcome in the dataset. Second, the set of units included in the vector is determined by assignment d. For the potential outcome vector $Y(d')$ under treatment assignment d', $Y_{\mathbb{I}(d)}(d')$ is a subvector of it. For different assignments d, it leads to different sets of units in the imputable outcome vector when testing the partially sharp null hypothesis. In contrast, under the sharp null hypothesis, $\mathbb{I}(d) = \{1, \ldots, N\}$, so $Y_{\mathbb{I}(d)}(d') = Y(d')$.

1.3.1 Pairwise Imputable Statistics

To address the first technical challenge, I construct a valid test statistic that accounts for missing potential outcomes. The definitions above provide the foundation for formally defining the core idea behind the test statistics.

Definition 6 (Pairwise Imputable Statistic)**.** *Let* $T : \mathbb{R}^N \times \{0,1\}^N \times \{0,1\}^N \longrightarrow \mathbb{R} \cup \{\infty\}$ *be a measurable function. We say that* T *is a* pairwise imputable statistic *if, for any* $d, d' \in \{0,1\}^N$*, the following holds:*

$$\textit{whenever } Y_i = Y_i' \textit{ for all } i \in \mathbb{I}(d) \cap \mathbb{I}(d'), \quad \textit{then} \quad T\big(Y_{\mathbb{I}(d)}, d'\big) \; = \; T\big(Y'_{\mathbb{I}(d)}, d'\big).$$

In words, T *depends only on the portion of* Y *(or* Y'*) in the intersection* $\mathbb{I}(d) \cap \mathbb{I}(d')$*.*

The set $\mathbb{I}(d) \cap \mathbb{I}(d')$ in Definition 6 is similar to the set H in Definition 1 of Zhang & Zhao (2023). Intuitively, it excludes units that are not imputable under the partially sharp null hypothesis in the test statistics. Under the sharp null hypothesis, where all units are imputable regardless of the treatment assignment d, $\mathbb{I}(d) \cap \mathbb{I}(d') = \{1, \ldots, N\}$ for any d and d', and all formulas reduce to the classical form as defined in Imbens & Rubin (2015).

At first glance, the pairwise imputable statistic may seem to restrict the form of the test statistics but it is actually general enough to accommodate commonly used test statistics. For instance, the classic difference in means can be written as

$$T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) = \underbrace{\bar{Y}_{\mathbb{I}(D^{obs})}(D^{obs})_{\{i : D \in \mathcal{D}_i(\epsilon_s)/\mathcal{D}_i(\epsilon_c)\}}}_{\text{Mean of } \textit{imputable} \text{ neighbor}} - \underbrace{\bar{Y}_{\mathbb{I}(D^{obs})}(D^{obs})_{\{i : D \in \mathcal{D}_i(\epsilon_c)\}}}_{\text{Mean of } \textit{imputable} \text{ control}}.$$

where for sets $A_i \subset \{0,1\}^N$, we have

$$\bar{Y}_{\mathbb{I}(D^{obs})}(D^{obs})_{\{i : D \in A_i\}} = \sum_{i \in \mathbb{I}(D^{obs})} 1\{D \in A_i\} Y_i(D^{obs}) / \sum_{i \in \mathbb{I}(D^{obs})} 1\{D \in A_i\},$$

In particular, $A_i = \mathcal{D}_i(\epsilon_s)/\mathcal{D}_i(\epsilon_c)$ corresponds to units in the distance interval $(\epsilon_s, \epsilon_c]$, while $A_i = \mathcal{D}_i(\epsilon_c)$ corresponds to units in the distance interval (ϵ_c, ∞).

This formula coincides with the classic difference in means when $\mathbb{I}(D^{obs}) =$

$\{1, \ldots, N\}$, and whether unit i lies in $(\epsilon_s, \epsilon_c]$ or (ϵ_c, ∞) depends on D. In practice, one of these mean values may be undefined if no unit in $\mathbb{I}(D^{obs})$ falls into one of those intervals. In that situation, I set $T = \max(Y^{obs}) - \min(Y^{obs})$ to ensure the test remains valid but conservative.[13] However, this issue did not arise in my empirical application and is generally less concerning for bipartite experiments when ϵ_s is moderate. For further discussion, see Appendix A.4.[14]

Running Example Continued. Consider the test statistic

$$T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) = \bar{Y}_{\mathbb{I}(D^{obs})}(D^{obs})_{\{i:D\in\mathcal{D}_i(0)/\mathcal{D}_i(1)\}} - \bar{Y}_{\mathbb{I}(D^{obs})}(D^{obs})_{\{i:D\in\mathcal{D}_i(1)\}}.$$

Table 1.4 presents the corresponding values for the first and second terms of the test statistic, while Figure 1.4 provides a visual representation of how we determine the imputable neighbor units and imputable control units.

Table 1.4: Constructing a Pairwise Imputable Statistic

Assignment D	Potential Outcome Y_i				$\bar{Y}_{\mathbb{I}(D^{obs})}(D^{obs})$		$T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D)$
	i_1	i_2	i_3	i_4	$\{i : D \in \mathcal{D}_i(0)/\mathcal{D}_i(1)\}$	$\{i : D \in \mathcal{D}_i(1)\}$	
$(1, 0, 0, 0)$	2	4	3	2	4	2.5	1.5
$(0, 1, 0, 0)$	🟥	?	3	2	?	2.5	2
$(0, 0, 1, 0)$	🟥	4	?	2	2	4	-2
$(0, 0, 0, 1)$	🟥	4	3	?	3	4	-1

Notes: Assignment D includes all potential assignments, with the first row corresponding to the observed assignment D^{obs}. Potential Outcome Y_i is the potential outcome of each unit under the null H_0^0, with red question marks representing missing values. Unit i_1 does not belong to set $\mathbb{I}(D^{obs})$, so the entire column is marked in red. $\bar{Y}_{\mathbb{I}(D^{obs})}(D^{obs})$ with $\{i : D \in \mathcal{D}_i(0)/\mathcal{D}_i(1)\}$ is the mean potential outcome for units in the distance interval $(0, 1]$, marked in blue. $\bar{Y}_{\mathbb{I}(D^{obs})}(D^{obs})$ with $\{i : D \in \mathcal{D}_i(1)\}$ is the mean potential outcome for units in the distance interval $(1, \infty)$. $T = \max(Y^{obs}) - \min(Y^{obs})$, marked in red when one of the mean values is undefined.

---

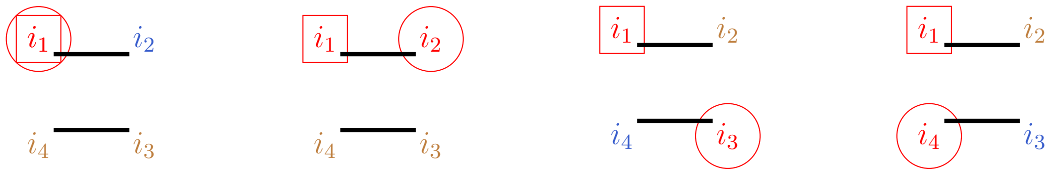
[13] Any number greater than the observed statistic would also suffice.

[14] To increase test power, one could combine this approach with conditional randomization testing, which trims treatment assignments to avoid undefined cases (Zhang & Zhao, 2023).

As illustrated in Figure 1.4, D^{obs} refers to the scenario where unit i_1 is treated, so the set of imputable units remains the same across different potential assignments D. However, the potential assignment D can change, altering which units belong to the neighborhood set and the control set. When $D = D^{obs}$ and unit i_1 is treated, the first term $\bar{Y}_{\mathbb{I}(D^{obs})}(D^{obs})_{\{i:D\in\mathcal{D}_i(0)/\mathcal{D}_i(1)\}}$ corresponds to the outcome of i_2, while the second term $\bar{Y}_{\mathbb{I}(D^{obs})}(D^{obs})_{\{i:D\in\mathcal{D}_i(1)\}}$ is the mean outcome of i_3 and i_4. When unit i_2 is treated, there are no imputable units in the neighborhood set, so I define $T = \max(Y^{obs}) - \min(Y^{obs}) = 2$ to ensure the test's validity.

The proposed test statistic satisfies Definition 6 because only units in the intersection $\mathbb{I}(D^{obs}) \cap \mathbb{I}(D)$ are used to construct it. For example, if D^{obs} has i_1 treated and D has i_3 treated, then $\mathbb{I}(D^{obs}) = \{i_2, i_3, i_4\}$, and $\mathbb{I}(D) = \{i_1, i_2, i_4\}$. As a result, their intersection is $\{i_2, i_4\}$. As shown in the third row of Table 1.4, the test statistic only depends on the outcomes of units i_2 and i_4.

Figure 1.4: Imputable Neighbor and Control Units for $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D)$



(a) D: i_1; D^{obs}: i_1 **(b)** D: i_2; D^{obs}: i_1 **(c)** D: i_3; D^{obs}: i_1 **(d)** D: i_4; D^{obs}: i_1

Notes: Red circles indicate treated units in D, which determine neighbor units in the interval $(0, 1]$ and control units in $(1, \infty)$. Red rectangles indicate treated units in D^{obs}, which determine imputable units.

Additionally, I can incorporate rank statistics by excluding non-imputable units and reranking the remaining units. Following Imbens & Rubin (2015), I define the rank as

$$R_i \equiv R_i\big(Y_{\mathbb{I}(D^{obs})\cap\mathbb{I}(D)}(D^{obs})\big)$$

$$
\begin{aligned}
&= \sum_{j \in \mathbb{I}(D^{obs}) \cap \mathbb{I}(D)} 1\{Y_j(D^{obs}) < Y_i(D^{obs})\} \\
&\quad + 0.5\Big(1 + \sum_{j \in \mathbb{I}(D^{obs}) \cap \mathbb{I}(D)} 1\{Y_j(D^{obs}) = Y_i(D^{obs})\}\Big) \\
&\quad - \frac{1 + \|\mathbb{I}(D^{obs}) \cap \mathbb{I}(D)\|}{2}.
\end{aligned}
$$

Thus, the test statistic becomes

$$T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) = \bar{R}_{\{i: D \in \mathcal{D}_i(\epsilon_s)/\mathcal{D}_i(\epsilon_c)\}} - \bar{R}_{\{i: D \in \mathcal{D}_i(\epsilon_c)\}}.$$

When $Y_i(D^{obs}) = Y_i(D)$ for all $i \in \mathbb{I}(D^{obs}) \cap \mathbb{I}(D)$, $R_i(Y_{\mathbb{I}(D^{obs}) \cap \mathbb{I}(D)}(D^{obs})) = R_i(Y_{\mathbb{I}(D^{obs}) \cap \mathbb{I}(D)}(D))$, meaning the ranks remain unchanged. Therefore, $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) = T(Y_{\mathbb{I}(D^{obs})}(D), D)$, satisfying Definition 6.

For further discussion on selecting test statistics in randomization tests and network settings, see Section 5 of Imbens & Rubin (2015), Athey et al. (2018), and Hoshino & Yanagi (2023). All test statistics can also be adapted to their absolute-value forms for two-sided testing.

While the method remains valid without covariate adjustments, incorporating them may improve the test's power in practice (Wu & Ding, 2021). See Appendix A.6 for a discussion on incorporating covariates. Moreover, since the proposed method is finite-sample valid, researchers can conduct subgroup analysis when expecting different patterns of interference across covariates.

Following Definition 6 of pairwise imputable statistics, I can derive a property to calculate test statistics using only the observed information:

Proposition 1. *Suppose the partially sharp null hypothesis* $H_0^{\epsilon_s}$ *is true. Suppose*

$T(Y_{\mathbb{I}(d)}(d), d')$ is a pairwise imputable statistic. Then,

$$T(Y_{\mathbb{I}(d)}(d), d') = T(Y_{\mathbb{I}(d)}(d'), d')$$

for any $d, d' \in \{0,1\}^N$.

The proof is provided in Appendix A.1. Let $d = D^{obs}$ and $d' = D$. By Proposition 1, $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) = T(Y_{\mathbb{I}(D^{obs})}(D), D)$ under the null $H_0^{\epsilon_s}$, ensuring I observe a counterfactual test statistic for comparison.

1.3.2 Unconditional Randomization Test

In this chapter, I am interested in the unconditional randomization test framework that satisfies the following definition:

Definition 7 (Unconditional randomization test). *An unconditional randomization test $\phi : \{0,1\}^N \to [0,1]$ is defined such that for any $D^{obs} \in \{0,1\}^N$,*

$$\phi(D^{obs}) = Q(\tilde{p}(D^{obs}), \alpha),$$

where $Q : [0,1] \times [0,1] \to [0,1]$ is a measurable function, α is the nominal level, and $\tilde{p}(D^{obs})$ can be written as

$$\tilde{p}(D^{obs}) = \sum_{d \in \{0,1\}^N} g(D^{obs}, d) P(D = d),$$

with P being the pre-specified probability distribution on the treatment assignment, and $g : \{0,1\}^N \times \{0,1\}^N \to \{0,1\}$ a measurable function.

The key feature of the unconditional randomization test is that the probability of rejection, $\phi(D^{obs})$, is computed by randomizing the treatment assignment according to the same probability distribution P that governs the original treatment assignment. This contrasts with methods in the existing literature, such as Athey et al. (2018), where the rejection function is based on randomizing the treatment assignment within a conditional probability space, conditioned on certain events. One example is the simple randomization test, which uses pairwise imputable statistics, with *p*-values constructed similarly to the classic FRT.

One example is the simple randomization test, which uses pairwise imputable statistics, with *p*-values constructed similarly to the classic FRT.

Definition 8 (Simple randomization test)**.** *A simple randomization test is an unconditional randomization test defined by* $\phi(D^{obs}) = 1\{pval(D^{obs}) \leq \alpha\}$*, where* $pval(D^{obs}) : \{0,1\}^N \to [0,1]$ *is the p-value function given by*

$$pval(D^{obs}) = P(T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) \geq T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D^{obs})) \textit{ for } D \sim P,$$

and $T(Y_{\mathbb{I}(d)}(d), d')$ *denotes a pairwise imputable statistic.*

Running Example Continued. Using pairwise imputable statistics $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D)$ and following Table 1.4, we can construct Table 1.5 with the test statistics for each assignment.

Following Definition 8, the *p*-value is given by

$$pval(D^{obs}) = P(T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) \geq T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D^{obs}))$$

with respect to $D \sim P$, where D is drawn independently from D^{obs}. Based on Table 1.5, this results in a *p*-value of 2/4. However, one might question whether this procedure guarantees finite-sample validity—specifically, whether it satisfies the condition $E_P(\phi(D^{obs})) \leq \alpha$ under the null hypothesis.

Table 1.5: Simple Randomization Test in the Example

Assignment D	Potential Outcome Y_i				$T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D)$
	i_1	i_2	i_3	i_4	
$(1,0,0,0)$	2	4	3	2	1.5
$(0,1,0,0)$	🟥	?	3	2	2
$(0,0,1,0)$	🟥	4	?	2	-2
$(0,0,0,1)$	🟥	4	3	?	-1

Notes: Assignment D includes all potential assignments, with the first row representing the observed assignment D^{obs}. Potential Outcome Y_i is the potential outcome of each unit under the null H_0^0, while red question marks denote missing values. Unit i_1 does not belong to the set $\mathbb{I}(D^{obs})$, so the column is marked in red. Blue cells represent the units used to calculate the mean value in the first term of the test statistics. $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D)$ are test statistics for different D, fixing D^{obs} where unit i_1 is treated.

Investigating Finite-Sample Validity. Although pairwise imputable statistics are used, naively constructing the *p*-value as defined in the classic FRT does not guarantee the test's validity. For the test to be valid, the following condition must hold under the partially sharp null hypothesis:

$$T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) \overset{d}{=} T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D^{obs}),$$

where $\overset{d}{=}$ indicates equality in distribution. The distribution on the left-hand side (LHS) is with respect to D, while the distribution on the right-hand side (RHS) is with respect to D^{obs}.

By Proposition 1, under the null hypothesis, we also have:

$$T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) \overset{H_0}{=} T(Y_{\mathbb{I}(D^{obs})}(D), D).$$

Here, $\overset{H_0}{=}$ denotes equality under the null hypothesis. However, the term $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D^{obs})$, being induced by the randomness of D^{obs}, satisfies:

$$T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D^{obs}) \overset{d}{=} T(Y_{\mathbb{I}(D)}(D), D).$$

Thus, for the test to maintain validity, we require:

$$T(Y_{\mathbb{I}(D^{obs})}(D), D) \overset{d}{=} T(Y_{\mathbb{I}(D)}(D), D).$$

This condition is not guaranteed under the partially sharp null hypothesis because $\mathbb{I}(D^{obs}) \neq \mathbb{I}(D)$ in general. Different treatment assignments D result in different sets of imputable units, leading to variability in $\mathbb{I}(D)$. This is a key technical challenge. In the special case of testing the sharp null hypothesis, where $\mathbb{I}(D^{obs}) = \{1, \ldots, N\} = \mathbb{I}(D)$, the validity trivially holds.

To address the challenges posed by varying imputable unit sets, previous literature suggests a remedy through the design of a conditioning event that consists of a fixed subset of imputable units, known as *focal units*, and a fixed subset of assignments, known as *focal assignments*. CRTs are then performed by conducting FRTs within this conditioning event. However, using conditioning events in practice introduces two key drawbacks.

First, as Zhang & Zhao (2023) pointed out, there is a trade-off between the sizes of focal units and focal assignments: a larger subset of treatment assignments typically corresponds to a smaller subset of experimental units. This inevitably results in a loss of information, with fewer units and assignments within the conditioning events, potentially affecting the test's power. Second, constructing the conditioning event adds a layer of computational complexity. This raises the question: can unconditional randomization testing still be valid in finite samples?

While previous approaches rely on carefully designing a fixed subset of units to maintain the validity of randomization testing, my method avoids fixing the subset of units during implementation. Instead, it achieves valid testing through a carefully designed *p*-value calculation, ensuring finite-sample validity without the need for conditioning events.

1.3.3 The Pairwise Comparison-Based *p*-values

Building on the selective inference literature (Wen et al., 2023; Guan, 2023), the key idea is to compute *p*-values by summing pairwise inequality comparisons between $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), d^r)$ and $T(Y_{\mathbb{I}(d^r)}(D^{obs}), D^{obs})$. When the null hypothesis is false, $T(Y_{\mathbb{I}(d^r)}(D^{obs}), D^{obs})$ remains relatively large across different d^r since the distance interval for each unit is fixed by D^{obs}. The change in d^r only alters the set of units used in the test statistics, and rejection of the null is still possible when the units in the neighborhood set tend to have high outcome values. As a result, we would expect a small *p*-value, as the probability that $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), d^r)$ exceeds $T(Y_{\mathbb{I}(d^r)}(D^{obs}), D^{obs})$ is low.

Formally, I refer to any randomization test with *p*-values constructed through this pairwise comparison method as a "PIRT."

Definition 9 (PIRT)**.** *The PIRT is an unconditional randomization test defined by* $\phi^{pair}(D^{obs}) = 1\{pval^{pair}(D^{obs}) \leq \alpha/2\}$*, where* $pval^{pair}(D^{obs}) : \{0,1\}^N \to [0,1]$ *is the p-value function given by*

$$pval^{pair}(D^{obs}) = P(T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) \geq T(Y_{\mathbb{I}(D)}(D^{obs}), D^{obs})) \text{ for } D \sim P,$$

and $T(Y_{\mathbb{I}(d)}(d), d')$ *denotes a pairwise imputable statistic.*

Theorem 1. *Suppose the partially sharp null hypothesis* $H_0^{\epsilon_s}$ *holds. Then, the PIRT, as* *defined in Definition 9, satisfies* $\mathbb{E}_P[\phi^{pair}(D^{obs})] < \alpha$ *for any* $\alpha \in (0,1)$*, where the expectation is taken with respect to* $D^{obs} \sim P$*.*

See the proof in Appendix A.1. Theorem 1 provides a worst-case guarantee, similar to cross-conformal prediction and jackknife+, due to certain pathological cases (Vovk et al., 2018; Barber et al., 2021; Guan, 2023). As in that literature, it empirically achieves size control at $\alpha/2$, as demonstrated in Section 1.4.1.[15]

Even with a large sample, an unbiased estimator of the *p*-value can be computed using Algorithm 1, which calculates the *p*-value as the average of $1 + R$ draws, with $r = 0$ corresponding to $d = D^{\text{obs}}$. See Appendix A.1 for a detailed discussion.

```
Algorithm 1 PIRT p-value
Inputs : Test statistic $T = T(Y(d), d)$, observed assignment $D^{obs}$, observed out-
         come $Y^{obs}$, treatment assignment mechanism $P$, and size $\alpha$.
for $r = 1$ to $R$ do
    Randomly sample $d^r \sim P$, and store $T_r \equiv T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), d^r)$.
    Store $T_r^{obs} \equiv T(Y_{\mathbb{I}(d^r)}(D^{obs}), D^{obs})$.
end
Output : p-value: $\hat{pval}^{pair} = (1 + \sum_{r=1}^{R} 1\{T_r \geq T_r^{obs}\})/(1 + R)$.
```

Running Example Continued. Using the difference-in-mean estimator as before,

$$T(Y_{\mathbb{I}(D)}(D^{obs}), D^{obs}) = \bar{Y}_{\mathbb{I}(D)}(D^{obs})_{\{i: D^{obs} \in \mathcal{D}_i(0)/\mathcal{D}_i(1)\}} - \bar{Y}_{\mathbb{I}(D)}(D^{obs})_{\{i: D^{obs} \in \mathcal{D}_i(1)\}}.$$

As shown in Figure 1.5, for each treatment assignment D, the test statistic is calculated as the mean value of i_2 (excluding missing values) minus the mean value of i_3 and i_4 (excluding missing values).

---

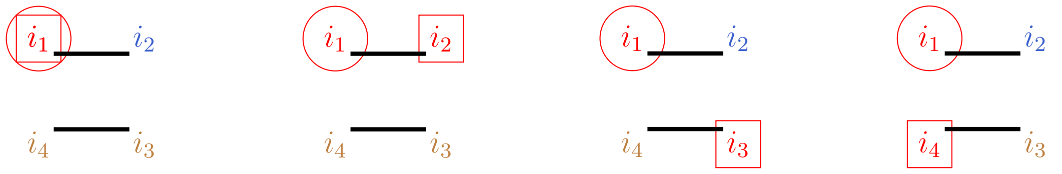
[15] See Appendix A.3 for a more conservative minimization-based PIRT that achieves theoretical size control with a rejection threshold of α.

Table 1.6: PIRT in the Example

Assignment D	Potential Outcome Y_i				$\bar{Y}_{\mathbb{I}(D)}(D^{obs})$		$T(Y_{\mathbb{I}(D)}(D^{obs}), D^{obs})$
	i_1	i_2	i_3	i_4	$\{i : D^{obs} \in \mathcal{D}_i(0)/\mathcal{D}_i(1)\}$	$\{i : D^{obs} \in \mathcal{D}_i(1)\}$	
$(1,0,0,0)$	2	4	3	2	4	2.5	1.5
$(0,1,0,0)$	🟥	?	3	2	?	2.5	2
$(0,0,1,0)$	🟥	4	?	2	4	2	2
$(0,0,0,1)$	🟥	4	3	?	4	3	1

Notes: Assignment D includes all potential assignments, with the first row representing the observed assignment D^{obs}. Potential Outcome Y_i is the potential outcome of each unit under the null H_0^0, with red question marks indicating missing values. Unit i_1 does not belong to either the neighborhood set or the control set under D^{obs}, so the column is marked red. Unit i_2 is in the distance interval $(0, 1]$ under D^{obs}, so the column is marked blue. Units i_3 and i_4 are in the distance interval $(1, \infty)$ under D^{obs}, so those columns are marked brown. $T(Y_{\mathbb{I}(D)}(D^{obs}), D^{obs})$ is calculated as the mean of non-missing potential outcomes in the blue columns minus the mean of non-missing potential outcomes in the brown columns.

Figure 1.5: Imputable Neighbor and Control Units for $T(Y_{\mathbb{I}(D)}(D^{obs}), D^{obs})$



(a) D^{obs}: i_1, D: i_1 **(b)** D^{obs}: i_1, D: i_2 **(c)** D^{obs}: i_1, D: i_3 **(d)** D^{obs}: i_1, D: i_4

Notes: Treated units in D^{obs} are marked with red circles and determine the neighbor units in the interval $(0, 1]$ and control units in $(1, \infty)$. Treated units in D are marked with red rectangles and determine the imputable units.

Based on Tables 1.6 and 1.4, I can construct Table 1.7, where each row represents the values used to compare and construct the p-value for each (D^{obs}, D) pair.

Only when D involves treating units i_1 or i_2 does $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) \geq T(Y_{\mathbb{I}(D)}(D^{obs}), D^{obs})$. Hence, $pval^{pair} = 2/4$. In practice, similar to Guan (2023), using $1/2$ to discount the number of equalities can reduce the p-value without compromising test validity. Additionally, in simulation experiments, using a uniform random number multiplied by the number of equalities also maintains test validity.

The validity of Algorithm 1 follows from the symmetry between $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), d^r)$ and $T(Y_{\mathbb{I}(d^r)}(D^{obs}), D^{obs})$ under the null hypothesis $H_0^{\epsilon_s}$. Intuitively, for each pair of assignments D^{obs} and d^r, both terms are restricted to units $i \in \mathbb{I}(D^{obs}) \cap \mathbb{I}(d^r)$ by Definition 6. Moreover, by Proposition 1, under the null, with $d = D$ and $d' = D^{obs}$, we have $T(Y_{\mathbb{I}(D)}(D^{obs}), D^{obs}) = T(Y_{\mathbb{I}(D)}(D), D^{obs})$, which is the counterfactual value of $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D)$ when flipping the observed assignment and randomized assignment between D and D^{obs}. Thus, the pairwise comparison is symmetric, and its validity follows from the conformal lemma in the conformal prediction literature (Guan, 2023).

Table 1.7: Pairwise Comparison for PIRT

Assignment D	Potential Outcome Y_i				$T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D)$	$T(Y_{\mathbb{I}(D)}(D^{obs}), D^{obs})$
	i_1	i_2	i_3	i_4		
$(1,0,0,0)$	2	4	3	2	1.5	1.5
$(0,1,0,0)$	?	?	3	2	2	2
$(0,0,1,0)$	?	4	?	2	-2	2
$(0,0,0,1)$	?	4	3	?	-1	1

Notes: Assignment D includes all potential assignments, with the first row representing the observed assignment D^{obs}. Potential Outcome Y_i is the potential outcome of each unit under the null H_0^0, with red question marks indicating missing values. $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D)$ are test statistics under different D while fixing D^{obs} for imputable units, with the same values as in Table 1.5. $T(Y_{\mathbb{I}(D)}(D^{obs}), D^{obs})$ are test statistics under different D for imputable units, with the same values as in Table 1.6.

Similar to Guan (2023), for non-directional tests, the absolute value of the test statistic can be used. For directional tests, the statistic can be applied to test for positive effects, while the negation of the statistic can be used to test for negative effects.

1.3.4 Comparison to Previous Literature

As previously discussed, a key distinction in PIRT is that the set of units included in $\mathbb{I}(d)$ varies across different assignments d, allowing all imputable units to be used for testing. However, under the sharp null hypothesis, $\mathbb{I}(d) = \{1, \ldots, N\}$ for all $d \in \{0, 1\}^N$, leading to

$$T(Y_{\mathbb{I}(d)}(D^{obs}), D^{obs}) = T(Y(D^{obs}), D^{obs}) \quad \text{for any } d, D^{obs} \in \{0, 1\}^N.$$

Thus, PIRT reduces to the classical FRT. The proposed method generalizes the FRT framework, ensuring validity under the partially sharp null hypothesis by allowing the set of units in the test statistic to vary across different assignments.

Comparison to the CRTs When testing under a partially sharp null hypothesis, the p-values constructed in Definitions 9 aligns closely with those from CRTs if we interpret $\mathbb{I}(D^{obs})$ as a focal unit set and $\{0, 1\}^N$ as a focal assignment set. The pair $(\mathbb{I}(D^{obs}), \{0, 1\}^N)$ represents a broader conditioning event than the traditional events in CRTs, which may or may not yield higher statistical power depending on whether the additional potential units and assignments contribute meaningful information.

In scenarios where a conditioning event can be specified over all imputable units in $\mathbb{I}(D^{obs})$, as demonstrated in Basse et al. (2024), CRTs with a well-defined focal assignment set may facilitate more targeted comparisons, potentially leading to a higher power. However, in cases where designing a suitable conditioning event is either infeasible or would produce only a limited number of focal units and assignments, PIRT can serve as a more practical alternative. Broadly, when including all assignments from $\{0, 1\}^N$ is suboptimal, combining PIRTs with CRTs could improve power by focusing on more pertinent test statistics and selected assignments (Lehmann & Romano, 2005; Hennessy et al., 2015). This approach highlights a promising avenue for future research on optimizing power through the flexibility of PIRTs and CRTs.

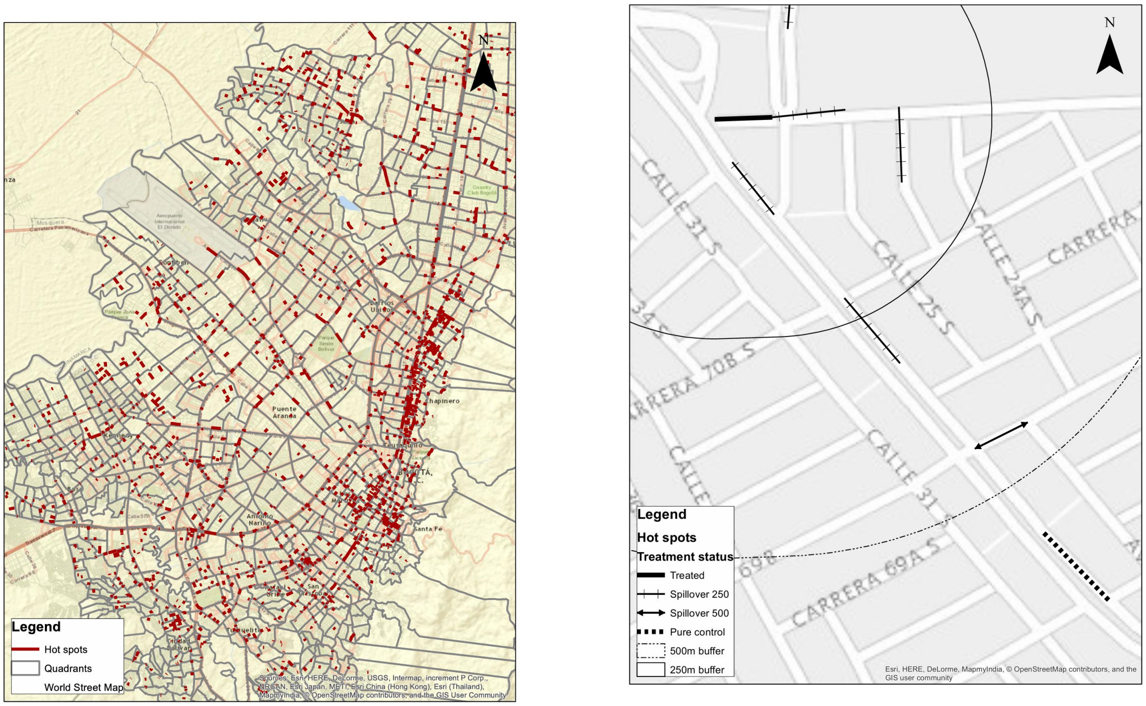
In practice, a researcher may want to test multiple distance levels ϵ_s rather than just a single one to estimate the boundary of interference. A sequential testing approach can achieve this by starting from the smallest ϵ_s and increasing until the test fails to reject. This method provides an estimate of the interference boundary while automatically controlling the family-wise error rate without requiring additional size adjustments. See Appendix A.5 for a detailed discussion on implementing this framework and adjusting for sequential testing.

1.4 EMPIRICAL APPLICATION

In 2016, a large-scale experiment was conducted in Bogotá, Colombia, as described by Blattman et al. (2021). The study involved 136,984 street segments, with 1,919 identified as crime hotspots. Among these, 756 were randomly assigned to a treatment involving increased daily police patrolling from 92 to 169 minutes over eight months. It also included a secondary intervention aimed at enhancing municipal services, though this is peripheral to the primary focus of my empirical application. The key outcome of interest was the number of crimes per street segment, encompassing both property crimes and violent crimes (e.g., assault, rape, and murder).

Figure 1.6a shows the distribution of hotspots, with many located close to each other. While only 756 street segments received the treatment, every segment potentially experienced spillover effects, creating a dense network. This complicates the application of cluster-robust standard errors for addressing unit correlation. The study estimated a negative treatment effect and used FRTs with a sharp null hypothesis of no effect for inference.

Figure 1.6: Map of Experimental Sample and Treatment Conditions


N
Legend
Hot spots
Quadrants
World Street Map
Puente Aranda
Chapinero
Santa Fe
Sources: Esri, HERE, DeLorme, USGS, Intermap, increment P Corp., NRCAN, Esri Japan, METI, Esri China (Hong Kong), Esri (Thailand), MapmyIndia, © OpenStreetMap contributors, and the GIS User Community
N
CALLE 31 S
CALLE 25 S
CALLE 24A S
CARRERA
CARRERA 70B S
CALLE 31
CARRERA 69A S
Legend
Hot spots
Treatment status
Treated
Spillover 250
Spillover 500
Pure control
500m buffer
250m buffer
Esri, HERE, DeLorme, MapmyIndia, © OpenStreetMap contributors, and the GIS user community

(a) Experimental Sample Map

(b) Assignment to Experimental Conditions

Notes: Panel (a) displays a map of the experimental sample, with hotspot street segments marked in red. Panel (b) shows an example of the assignment to the four experimental conditions. Source: Blattman et al. (2021).

In evaluating the total welfare impact of the policy, it is essential to assess whether interference occurred following treatment assignment, such as crime displacement or deterrence in neighboring areas. I aim to address three key questions: 1) Does interference exist? 2) If so, is it displacement or deterrence? 3) At what distance is this interference effective? Given the complexity of modeling correlation across units in such a dense network, testing a partially sharp null hypothesis, as proposed by Blattman et al. (2021) and Puelz et al. (2021), becomes relevant. I specify the distance threshold sequence $(\epsilon_0, \epsilon_1, \epsilon_2, \epsilon_3) = (0, 125, 250, 500)$ for $K = 3$, where the interval $(500, \infty)$ represents a pure control group with no treated units within 500 meters. Figure 1.6b provides an example of the distance intervals identified in Blattman et al. (2021).

Table 1.8 presents descriptive statistics for the number of crimes during the intervention period. The t-statistics for t-tests between each of the two columns reveal two key findings: treated hotspots experience significantly fewer crimes, and non-hotspot areas report even fewer crimes when located farther from any treated units. However, the extremely high values of the t-statistics raise concerns about interpreting these t-test results as evidence of a displacement effect. As noted earlier, standard errors may be under-estimated, and units at varying distances from treated areas may not be directly comparable. Both factors could contribute to the high t-statistics observed in the table.

In Blattman et al. (2021), the authors reported no significant displacement effect for violent crimes and a marginally significant displacement effect for property crimes.[16] However, as previously illustrated, p-values for the t-test may not adequately capture the extent of interference. Moreover, using FRTs to test partially sharp null hypotheses may not be valid. Thus, how might these conclusions change if a valid testing approach is implemented?

1.4.1 Power Comparison of Spatial Interference: A Simulation Study

To evaluate the methodology in a large-scale setting, I conduct a simulation study to preselect the preferred method. For simplicity, the sample size is set to 1,000

---

[16] The treatment effect for violent crime was significant but property crime effects were insignificant.

units, including 20 hotspots and 7 randomly treated units, mirroring proportions in the original Bogotá study. I focus on two distance thresholds, using $(\epsilon_0, \epsilon_1, \epsilon_2) = (0, 0.1, 0.2)$. See Appendix A.7 for a detailed discussion.

To approximate the Bogotá setting, I calibrate the potential outcome schedule using gamma distributions that match the observed mean and variance of total crimes (Table A.7.1). A negative treatment effect of 1 is imposed while maintaining non-negative crime counts for treated units. Additionally, a decreasing displacement effect is introduced via a positive τ, which is the primary focus of this analysis.

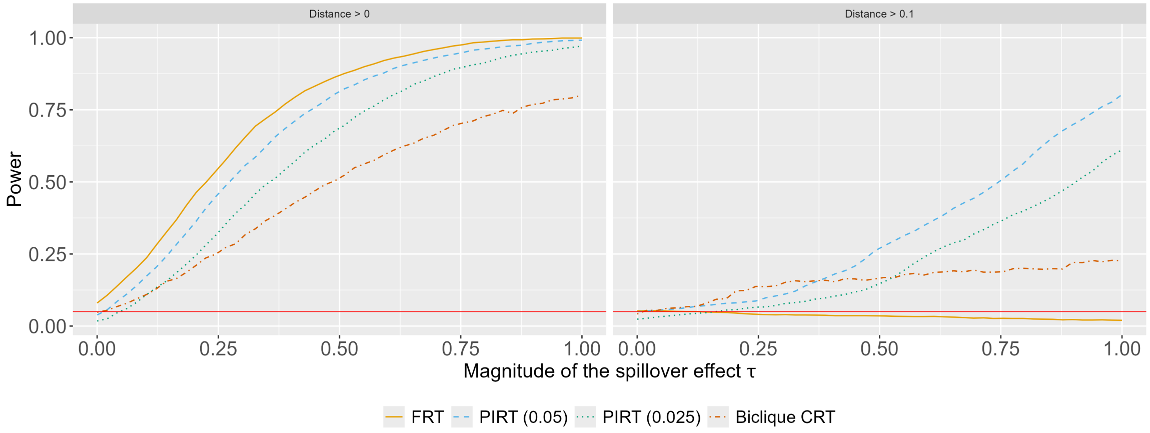
The partially sharp null hypothesis for $k = 0$ and 1 is given by

$$H_0^{\epsilon_k} : Y_i(d) = Y_i(d') \textit{ for all } i \in \{1, \ldots, N\}, \textit{ and any } d, d' \in \mathcal{D}_i(\epsilon_k).$$

In the analysis, I compare five methods: 1) the classic FRT, using the sharp null hypothesis of no effect, as used in Blattman et al. (2021) for spillover effect inference; 2) the biclique CRT proposed by Puelz et al. (2021), a benchmark for CRTs due to their strong power properties in simulations involving general interference; 3) the PIRT with rejection based on $\alpha/2$, ensuring validity in the worst-case scenario; and 4) the PIRT with rejection based on α.

Two main criteria guide the testing procedure selection. First, in the absence of a spillover effect ($\tau = 0$), the partially sharp null hypothesis should be rejected no more than 5% of the time, maintaining type I error control. Second, when a spillover effect exists ($\tau > 0$), the partially sharp null hypothesis should be rejected as frequently as possible to maximize power. To assess power, I consider 50 equally spaced τ values between 0 and 1, conducting 2,000 simulations for each τ to compute the average rejection rate for each method. The algorithm is detailed in Appendix A.7, with a focus on displacement effects and one-sided testing using the non-absolute difference in mean.

Figure 1.7: Power Comparison of Testing Methods for Different Hypotheses



Notes: The left panel shows the power comparison for testing H_0^0, while the right panel illustrates the power comparison for $H_0^{0.1}$. The red line indicates the size level $\alpha = 0.05$. PIRT (0.025) indicates the PIRT with a rejection threshold of $\alpha/2$, and PIRT (0.05) denotes the PIRT with rejection based on α.

Figure 1.7 (left panel) shows that the FRT over-rejects the true partially sharp null hypothesis when $\tau = 0$, consistent with Athey et al. (2018)'s observation that testing the sharp null of no effect is invalid for partially sharp null hypotheses. In my simulation, with only seven treated units (0.7% of the total), the rejection rate is around 10%. Surprisingly, the PIRT without adjustment at the α level maintains good size control, indicating that the $\alpha/2$ rejection level primarily provides a worst-case guarantee and can be conservative, with rejection rates below 5%. The biclique CRT remains valid, with a rejection rate near 5%.

Regarding power, the FRT is excluded from the comparison due to its invalidity. The unadjusted PIRT demonstrates the best performance, outperforming other methods across all effect magnitudes τ. Among methods with theoretical size control, the PIRT with $\alpha/2$ rejection is optimal, though it slightly lags behind the biclique CRT for small τ values. Despite its validity, the power of the biclique CRT increases slowly as the spillover effect grows, with a rejection rate remaining below 90% even when $\tau = 1$.

The right panel of Figure 1.7 contrasts with the results in the left panel. First, all methods, including the FRT, maintain validity under the null hypothesis. This may be due to the fact that hotspots rarely belong to the exposure levels $(0.1, 0.2]$ or $(0.2, \infty)$. As a result, despite the negative treatment effect, its impact on the test statistics used for this analysis remains minimal. Similar to the case of H_0^0, both the PIRT and biclique CRT methods exhibit rejection rates close to 5%, while the PIRT with a rejection level of $\alpha/2$ remains conservative.

Second, all methods demonstrate significantly lower power compared to the results under H_0^0. This is largely because only 60% of the units are relevant to the partially sharp null hypothesis in this scenario, and the spillover effect magnitude is halved (0.5τ). Nevertheless, the PIRT method still exhibits sufficient power when the spillover magnitude τ is large enough, outperforming other methods, particularly when using the unadjusted rejection level α. Surprisingly, the FRT exhibits under-rejection, with almost no power across all values of τ. This can be explained by the nature of the FRT: the *p*-value remains large unless the observed test statistic exceeds most test statistics generated from randomized treatment assignments. However, since units within the group $(0, 0.1]$ under the observed assignment contribute to test statistics for other randomized assignments d–and these units are subject to a spillover effect of τ–the observed test statistics constructed from units in $(0.1, 0.2]$ and $(0.2, \infty)$ fail to exhibit high values, even when τ is large. This results in large *p*-values for the FRT, explaining its under-rejection.

Taken together with the earlier discussion for H_0^0, these findings illustrate how the FRT, when used to test partially sharp null hypotheses, may lead to either over-rejection or under-rejection depending on the scenario. Finally, although the biclique CRT method demonstrates power, its increase is slower than that of the PIRT methods, which may be attributed to the complexity of finding an optimal conditioning event in the presence of spatial interference. One point worth noting is that the performance of the biclique method depends on the parameter values used for solving the bicliques; thus, with more advanced computing resources, its power performance could improve. Overall, the key advantage of my method lies in its computational simplicity, and future research could further explore power comparisons between different methods.

Overall, the results favor PIRT methods, especially the unadjusted PIRT. Thus, I apply PIRTs to replicate the results of Blattman et al. (2021), using the non-absolute difference-in-mean estimator.

1.4.2 Implementation of PIRTs for Testing the Existence of a Displacement Effect

Consider the experimental setting described in Blattman et al. (2021), where the observed treatment assignment is denoted by D^{obs}. Similar to Blattman et al. (2021), we can regress the number of crimes, Y, on a spillover proximity indicator $S(D^{obs})$, which indicates whether units are within 125 meters of any treated unit. This proximity indicator is directly determined by the treatment assignment D^{obs} and would change with a different treatment assignment D. While additional covariates can be included in the regression, the key test statistic remains the coefficient from regressing Y on the proximity indicator $S(D)$.

To test the partially sharp null hypothesis using the PIRT, follow these steps:

1. Randomly reassign treatments D to the units.
2. For each reassignment D, identify the subsample of units that are untreated under both D^{obs} and D. These units form the set of imputable units, as their potential outcomes remain unaffected by treatment under both assignments.
3. Collect the coefficient β from the regression of Y on $S(D^{obs})$ within the subsample of imputable units.
4. Collect the coefficient β' from the regression of Y on $S(D)$ within the same subsample of imputable units.

Both steps 3 and 4 construct the pairwise imputable statistics, which use only untreated units under either the observed treatment assignment D^{obs} or the randomized assignment D.

For the rejection decision when testing for displacement effects, the PIRT computes the *p*-value as the fraction of reassignments D such that $\beta' \geq \beta$. The null hypothesis of no displacement effect is rejected if the *p*-value is less than or equal to $\alpha/2$. Simulation results suggest that using α as the rejection threshold is also empirically valid. The method accommodates two-sided tests by comparing $\|\beta'\| \geq \|\beta\|$ or by testing for deterrence effects using the negative of the coefficient, comparing $-\beta' \geq -\beta$.

1.4.3 PIRT on Actual Data

I conduct the analysis using the publicly available dataset from Blattman et al. (2021), which includes street-level treatment assignments and distance intervals with thresholds at 125, 250, and 500 meters. The dataset also contains $1,000$ pseudo-randomized treatments and their respective distance intervals, used in the original paper for randomization inference. However, the dataset lacks precise longitude and latitude data for the street segments, preventing me from extending randomization testing beyond the available $1,000$ random treatments.

Due to the large fraction of zero outcomes, I use both an indicator for any crime occurrence and the number of crimes as outcome variables, as shown in Table 1.9.

In the main specification, I use the PIRT with the difference-in-means estimator as the test statistic. A key advantage of this framework is that it remains valid regardless of the choice of test statistic. However, in practice, researchers may incorporate covariates to improve power or conduct heterogeneous analysis. See Appendix A.6.2 for a detailed discussion on how to incorporate covariates into the analysis and additional insights gained from doing so.

Tables 1.9 reveal a significant displacement effect for violent crimes but not for property crimes. This contrasts with the original study, which found no evidence of significant displacement effects for violent crime. After adjusting for multiple hypothesis testing using Algorithm A.5.1, PIRT detects a significant short-range displacement effect within 125 meters at the 10% level when using the difference-in-means estimator. Moreover, it could be significant at the 5% level if directly reject by level α, as suggested by the simulation study. On the other hand, there is no evidence of spillover effects at any distance for property crimes.

For violent crimes, however, there is evidence of additional spillover effects beyond 250 meters, with an unadjusted p-value of 0.045 for the $(250m, \infty)$ interval when using the number of crimes as the outcome variable in PIRT. This suggests the presence of two distinct types of offenders committing violent crimes: risk-averse criminals may relocate farther away rather than being displaced to nearby neighborhoods, while less risk-averse criminals are more likely to relocate to adjacent areas.

To test this hypothesis, I further disaggregate violent crimes into socially costly crimeshomicides and sexual assaultsand other violent crimes. Table 1.10 presents results that investigate heterogeneous displacement patterns within violent crimes, revealing distinct interference effects. Specifically, socially costly crimes contribute to spillover effects beyond 250 meters but not within 125 meters. High-risk offenders involved in homicides and sexual assaults tend to relocate farther away in response to intensive policing, rather than being displaced to nearby neighborhoods. This finding underscores the dynamic responses of different types of offenders to hot-spot policing.

To the best of my knowledge, this is the first causal evidence of a displacement effect extending to more distant areas rather than proximate neighborhoods.[17] However, after adjusting for multiple hypothesis testing using Algorithm A.5.1, these results are no longer statistically significant. Applied researchers should interpret these findings with caution in future studies.

These results not only highlight the general applicability of the PIRT method but also provide suggestive evidence for policy implications and potential criminal motives in Bogotá, following the insights of Blattman et al. (2021). From a policy perspective, it remains unclear whether reallocating state resources to these hotspots has led to an overall reduction in crime. Further investigation is needed to identify the specific locations most affected by displacement so that those areas can be directly targeted.

---

[17] This finding suggests that the distance interval $(500m, \infty)$ may serve as a more appropriate control group than the $(250m, \infty)$ interval used by Blattman et al. (2021).

Regarding criminal motives in Bogotá, a possible explanationconsistent with standard economic models of crimeis that violent crime in the city's hotspots is not purely expressive, as suggested by Blattman et al. (2021). Instead, some violent crimes, such as contract killings, may be driven by generally mobile criminal rents. By increasing the risk of detection, intensive policing deters criminals from committing crimes in specific locations but the crimes themselves may relocate rather than be entirely prevented. In contrast, property crimes, which are often instrumental and linked to immobile criminal rents, appear to be deterred without causing further spillover effects. As Blattman et al. (2021) noted, violent crimes are often considered more severe than property crimes, making displacement effects an essential consideration when evaluating the overall welfare impact of policy interventions.

Table 1.8: Descriptive Statistics During the Intervention

Stats	**Crime hotspots**		**Non-hotspots (distance to treated units)**			
	Treated	Non-treated	$(0m, 125m]$	$(125m, 250m]$	$(250m, 500m]$	$(500m, \infty)$
Obs.	756	1,163	24,571	32,034	45,147	33,313
# of total crimes						
Mean	0.935	1.311	0.378	0.294	0.242	0.180
SD	1.519	2.332	1.006	0.921	0.736	0.602
Max	12	43	33	40	25	31
% of > 0	44.84	53.22	23.17	19.01	16.69	13.13
t-stat of t-test		-3.93		10.34	8.60	12.62
# of property crimes						
Mean	0.712	1.035	0.262	0.195	0.158	0.111
SD	1.269	2.099	0.778	0.683	0.555	0.441
Max	12	40	32	36	21	27
% of > 0	38.36	45.66	17.44	13.96	11.90	8.86
t-stat of t-test		-3.81		10.93	8.26	12.83
# of violent crimes						
Mean	0.224	0.276	0.115	0.099	0.084	0.069
SD	0.593	0.650	0.467	0.473	0.376	0.334
Max	5	6	17	40	13	11
% of > 0	16.40	20.29	8.59	7.29	6.51	5.47
t-stat of t-test		-1.79		4.23	4.75	5.79

Notes: This table presents descriptive statistics for crime data during the intervention, divided into two categories: crime hotspots (treated and non-treated) and non-hotspot areas, which are further grouped by their distance from the treated units. The statistics cover three types of crimes: total crimes, property crimes, and violent crimes. For each group, the table provides the mean, standard deviation (SD), maximum (Max), and the percentage of units with positive crimes (% of > 0). The t-statistic values (t-stat of t-test) represent the results from t-tests comparing the difference in means between treated versus non-treated units within crime hotspots, non-hotspot units within $(0m, 125m]$ versus $(125m, 250m]$, non-hotspot units within $(125m, 250m]$ versus $(250m, 500m]$, and non-hotspot units within $(250m, 500m]$ versus $(500m, \infty)$.

Table 1.9: p-Values: Testing the Displacement Effect at Different Distances

	Unadjusted p-values		
	$(0m, \infty)$	$(125m, \infty)$	$(250m, \infty)$
Violent crime			
Indicator of > 0	*0.027*	0.812	0.060
# of crimes	*0.047*	0.546	*0.045*
Property crime			
Indicator of > 0	0.390	0.466	0.486
# of crimes	0.325	0.346	0.394

Notes: The table shows the impact of intensive policing on violent and property crime. "Indicator of > 0" refers to an indicator for any crime occurrence, while "# of crimes" represents the raw number of reported crimes. p-Values are constructed using the difference-in-means estimator as the test statistic.

Table 1.10: p-Values: Heterogeneous Patterns Within Violent Crimes

	Unadjusted p-values		
	$(0m, \infty)$	$(125m, \infty)$	$(250m, \infty)$
Homicides and sexual assaults			
Indicator of > 0	0.274	0.864	0.065
# of crimes	0.417	0.815	*0.043*
Not homicides or sexual assaults			
Indicator of > 0	*0.030*	0.752	0.097
# of crimes	*0.044*	0.491	0.057

Notes: The table reports the impact of intensive policing on two types of violent crimes. Indicator of > 0: indicator of any crime; # of crimes: raw number of reported crimes. p-Values constructed based on the difference-in-means estimator as the test statistic.

1.5 CONCLUSION

This paper introduces a practical testing framework for detecting interference in network settings. The proposed tests offer computational simplicity over previous methods while retaining strong power and size properties, making them highly applicable for empirical research.

Theoretically, I formalize unconditional randomization testing and PIRT, addressing two primary challenges in testing partially sharp null hypotheses: only a subset of potential outcomes is imputable, and the set of units with imputable potential outcomes varies across treatment assignments. The PIRT addresses the first challenge by employing pairwise imputable statistics and the second by constructing p-values through pairwise comparisons. I prove the PIRT maintains size control, and I propose a sequential testing procedure to estimate the "neighborhood of interference, ensuring control over the FWER.

Beyond network settings, the PIRT might hold broader applicability. For instance, Zhang & Zhao (2021) shows that partially sharp null hypotheses are relevant in time-staggered designs. This opens promising avenues for future research, including extending the framework to quasi-experimental settings and observational studies. In quasi-experimental designs, developing a unified framework that can be applied to time-staggered adoption, regression discontinuity, and network settings would be highly valuable (Borusyak & Hull, 2023; Kelly, 2021). For observational studies, incorporating propensity score weighting to create pseudo-random treatments and conducting sensitivity analyses would be crucial, as noted by Rosenbaum (2020).

While simulations suggest that the PIRT performs favorably compared to CRTs, their power properties remain unexplored. Insights from studies such as Puelz et al. (2021) on CRT power properties and Wen et al. (2023) on the near-minimax optimality of minimization-based p-values suggest that further exploration of PIRT's power could yield valuable insights. Additionally, power may increase when the PIRT is combined with CRTs in specific settings, making the construction of an optimal testing framework for interference an important question for future research.

CHAPTER 2

Convexity Not Required: Estimation of Smooth Moment Condition Models

2.1 INTRODUCTION

The Generalized and Simulated Method of Moments (GMM, SMM) are commonly used to estimate structural Economic models. To find these estimates, modern computer software provides researchers with a large set of free and non-free numerical optimizers, which, after inputting some tuning parameters, return a guess for the parameters of interest. While sampling properties of estimators are often derived, their practical implementation often receives a less detailed treatment. There is now a vast literature on statistical learning with a convex loss function. However, these results need not directly apply to GMM, as it often involves non-convex minimizations. A number of authors have pointed out the lack of robustness of off-the-shelf methods, Knittel & Metaxoglou (2014) illustrate this in the context of demand estimation.

In many empirical studies, the authors comment on the challenge of estimation due to the non-convexity of the sample GMM objective function. Methods like gradient-descent or quasi-Newton are seldom used, not too surprisingly as these are convex optimizers. An important concern is that, in general, non-convex optimization is particularly challenging when the number of parameters to be estimated is moderate or large, as generic non-linear optimization is subject to a curse of dimensionality (Andrews, 1997, Sec2).

The main contribution of the chapter is to show that *convexity is not required* for some methods to perform well in GMM estimation, specifically: some algorithms are globally convergent if the Jacobian of the moments satisfies a global rank condition. This defines a class of non-convex problems that is as hard as convex problems for optimization. Since this is perhaps surprising, the following gives some intuition behind the result. Suppose the sample moments $\overline{g}_n(\theta) = \partial_\theta \ell_n(\theta)$ correspond to the gradient of a sample log-likelihood function, say that of a Probit model. Then, their Jacobian $G_n(\theta) = \partial^2_{\theta,\theta'} \ell_n(\theta)$ is the Hessian of the log-likelihood. The Jacobian of the moments is the Hessian of the log-likelihood, so it is strictly negative definite everywhere when the log-likelihood, to be maximized, is strictly concave. Meanwhile, the GMM objective $Q_n(\theta) = \frac{1}{2}\|\overline{g}_n(\theta)\|^2$, to be minimized, here with identity weighting, need not be convex. Its Hessian $\partial^2_{\theta,\theta'} Q_n(\theta) = G_n(\theta)' G_n(\theta) + (\overline{g}_n(\theta)' \otimes I_d)\partial_\theta \text{vec}(G_n'(\theta))$ can be singular, or non-definite, depending on the last term.[1] When this is the case: ℓ_n is concave but Q_n is non-convex, even though they estimate the exact same quantity $\hat{\theta}_n$.

Method of moments could be solved as systems of non-linear equations, $\overline{g}_n(\theta) = 0$. Over-identified GMM is generally framed as an M-estimation, $\min_{\theta \in \Theta} Q_n(\theta)$, because the system of equations does not have an exact solution in finite samples. Yet, the Probit example shows information can be lost when minimizing Q_n directly. This paper shows that some algorithms are robust to the non-convexity of Q_n by implicitly solving for $\overline{g}_n(\theta) = 0$ rather than explicitly minimizing Q_n. Under a rank condition, gradient-descent and Gauss-Newton (GN) are globally convergent, with appropriate tuning. Newton-Raphson and quasi-Newton can be unstable as they require inverting the potentially singular Hessian of Q_n. The result applies to over-identified and moderately misspecified models by adapting the rank condition appropriately. For these models, there are no parameters for which $\overline{g}_n(\theta) = 0$ is feasible, even in large samples. As one may suspect, the

---

[1] $\otimes$ is the kronecker product, vec vectorizes the matrix to a column vector, $\partial_\theta \text{vec}(G_n'(\theta))$ is the Jacobian of the vectorized Jacobian.

rank condition precludes local optima. The rank condition is invariant to smooth one-to-one moderately non-linear reparameterization. When there is a single parameter and moment, the condition stated in this paper simply requires that the scalar moment be strictly monotone. In the particular case where the moments have an exact solution $\overline{g}_n(\theta) = 0$, our rank conditions imply the so-called Polyak-Łojasiewicz inequality; a popular relaxation of strong convexity in the machine learning literature.

A simple MA(1) estimation from Gourieroux & Monfort (1996) illustrates this analytically and numerically. The problem is non-convex: the scalar Hessian can be positive, negative, or zero; yet the rank condition holds. As predicted, the recommended Gauss-Newton algorithm converges. Newton-Raphson provably diverges, and off-the-shelf optimizers can be unstable. Then, two empirical applications further confirm the predictions. The first application revisits the numerical results of Knittel & Metaxoglou (2014) for estimating random coefficient demand models. The same GN algorithm systematically converges from a wide range of starting values. In contrast, R's more sophisticated built-in optimizers can be inaccurate and often crash without additional error-handling. The second application estimates a small New Keynesian model with endogenous total factor productivity by impulse response matching. Matlab's built-in optimizers have better error-handling so that crashes are less problematic. Nonetheless, these optimizers' performance can be mixed and sensitive to reparameterizations whereas GN performs well for nearly all starting values.

In all three applications, the GMM objective is non-convex at most values, whereas the rank conditions hold at all or most values. Given the results presented in the paper, this explains the good performance of GN relative to more commonly used methods. The main takeaway is that non-convexity need not be a deterrent to structural estimation: simple algorithms can converge quickly and globally under alternative conditions.

Structure of the chapter. Section 2.2 reviews optimization results for convex and non-convex loss function and then provides the main results, illustrated in Section 2.4 with two empirical applications. Appendix B.1 gives the proofs to the main results. Appendix B.4 gives R code to replicate the MA(1) example. Appendix B.5 provides additional simulation and empirical results. Appendix B.6 gives additional details about the methods found in the survey.

2.2 GMM ESTIMATION WITHOUT CONVEXITY

Let $\overline{g}_n(\theta) = 1/n \sum_{i=1}^n g(\theta; x_i)$ be the sample moments and $G_n(\theta) = \partial_\theta \overline{g}_n(\theta)$ their Jacobian. Their population counterparts are $g(\theta) = \mathbb{E}[g(\theta; x_i)]$ and $G(\theta) = \partial_\theta g(\theta)$. W_n is a weighting matrix which, for simplicity, does not depend on θ – this excludes continuously-updated estimations. The sample GMM objective function is:

$$Q_n(\theta) = \frac{1}{2}\overline{g}_n(\theta)' W_n \overline{g}_n(\theta),$$

and the goal is to find the global minimizer $\hat{\theta}_n$ of Q_n in Θ, a compact and convex subset of $\mathbb{R}^{d_\theta}$. The population objective $Q(\theta) = g(\theta)' W g(\theta)$, defined similarly using the limit W of W_n, has a global minimizer $\theta^\dagger$. Throughout, it will be assumed that the sample Q_n is continuously differentiable on Θ. More specifically, this paper considers derivative-based optimizers of the form:

$$\theta_{k+1} = \theta_k - \gamma_k P_k G_n(\theta_k)' W_n \overline{g}_n(\theta_k), \tag{2.1}$$

for $k = 0, 1, \dots$, some staring value $\theta_0 \in \Theta$ and a matrix P_k, called conditioning matrix, assumed to be symmetric. The tuning parameter $\gamma_k \in (0, 1]$ is called the learning rate. There are several ways to motivate (2.1) as a minimization algorithm in the context of GMM estimation. They are conceptually similar but implicitly rely on a different set of assumptions. The first is to consider a quadratic approximation of the GMM objective function Q_n:

$$Q_n(\theta) \simeq Q_n(\theta_k) + \partial_\theta Q_n(\theta_k)(\theta - \theta_k) + \frac{1}{2\gamma_k}(\theta - \theta_k)'\partial^2_{\theta,\theta'} Q_n(\theta_k)(\theta - \theta_k),$$

here γ_k penalizes the quality of the quadratic approximation. For linear models, such as OLS and IV regressions, Q_n is quadratic so that $\gamma_k = 1$ is feasible. For non-linear models, the approximation is inexact, and $\gamma_k < 1$ is generally required. Minimizing the right-hand-side with respect to θ yields a Newton-Raphson (NR) iteration: $\theta_{k+1} = \theta_k - \gamma_k[\partial^2_{\theta,\theta'} Q_n(\theta_k)]^{-1}\partial_\theta Q_n(\theta_k)$ with $\partial_\theta Q_n(\theta_k) = G_n(\theta_k)' W_n \overline{g}_n(\theta_k)$ and $P_k = [\partial^2_{\theta,\theta'} Q_n(\theta_k)]^{-1}$. A quasi-Newton (QN) iterations replaces the Hessian matrix $\partial^2_{\theta,\theta'} Q_n(\theta_k)$ with an approximation computed sequentially over k. The most popular QN software implementation is called BFGS. Importantly, the quadratic approximation implicitly requires that Q_n is strongly convex around θ_k that is $\partial^2_{\theta,\theta'} Q_n(\theta_k)$ strictly positive definite so that (2.1) minimizes the quadratic approximation. When the Hessian H_n is non-definite, (2.1) is not the minimizer.

Another way to motivate (2.1) is to consider a linear approximation of the moments and plug it into the GMM objective function:

$$\begin{aligned}
\overline{g}_n(\theta) &\simeq \ \overline{g}_n(\theta_k) + \frac{1}{\gamma_k} G_n(\theta_k)(\theta - \theta_k), \\
Q_n(\theta) &\simeq \Big[\overline{g}_n(\theta_k) + \frac{1}{\gamma_k} G_n(\theta_k)(\theta - \theta_k)\Big]' W_n \Big[\overline{g}_n(\theta_k) + \frac{1}{\gamma_k} G_n(\theta_k)(\theta - \theta_k)\Big],
\end{aligned}$$

where now γ_k penalizes the quality of the linear approximation. Take the first order condition in the last display to find (2.1) with $P_k = (G_n(\theta_k)'W_nG_n(\theta_k))^{-1}$, a Gauss-Newton (GN) iteration. The quadratic approximation requires the Hessian H_n of Q_n to be strictly positive definite at θ_k. A GN iteration minimizes the linear approximation as long as the Jacobian G_n of $\overline{g}_n$ has full rank at θ_k so that $G_n(\theta_k)'W_nG_n(\theta_k)$ is strictly positive definite. Convexity is more challenging to satisfy away from the solution since $\|\overline{g}_n(\theta_k)\| \gg 0$ can result in a non-definite Hessian $H_n(\theta_k) = G_n(\theta_k)'W_nG_n(\theta_k) + (\overline{g}_n(\theta_k)'W_n \otimes I_d)\partial_\theta \text{vec}[G_n(\theta_k)']$, depending on the last term. This is illustrated with a simple MA(1) example below. This suggests that quadratic-based methods (NR, BFGS) and linear-based methods (GN) can behave differently when Q_n is globally non-convex.

Gradient-Descent (GD) can be motivated by either a linear or a quadratic approximation. The following summarizes the choice of P_k for each algorithm:

Table 2.1: Optimizers considered in (2.1)

1.	Gradient-Descent (GD)	$P_k = I_d$,
2.	Newton-Raphson (NR)	$P_k = [\partial^2_{\theta,\theta'} Q_n(\theta_k)]^{-1}$,
3.	quasi-Newton (QN)	P_k approximates $[\partial^2_{\theta,\theta'} Q_n(\theta_k)]^{-1}$,
4.	Gauss-Newton (GN)	$P_k = [G_n(\theta_k)'W_nG_n(\theta_k)]^{-1}$.

The following gives assumptions on the population moments used to describe the large sample optimization properties. When these assumptions hold, the sample moments have similar properties, this is shown in Lemmas B.12, B.14.

Assumption 1. *Suppose the observations x_i are iid and: (i) $Q(\theta) = \|g(\theta)\|^2_W$ has a unique minimum $\theta^\dagger \in$ interior(Θ), (ii) $g(\theta; x_i)$ and $g(\theta)$ are continuously differentiable on Θ, (iii) $\mathbb{E}[\|g(\theta; x_i)\|^2] < \infty$, for all $\theta \in \Theta$, $\mathbb{E}[\|G(\theta; x_i)\|^2] < \infty$, for all $\theta \in \Theta$, there exist a $\bar{L}(\cdot) \geq 0$ such that for any $(\theta_1, \theta_2) \in \Theta^2$, $\|G(\theta_1; x_i) - G(\theta_2; x_i)\| \leq \bar{L}(x_i)\|\theta_1 - \theta_2\|$, where $\mathbb{E}[|\bar{L}(x_i)|^2] < \infty$, and $\mathbb{E}[\bar{L}(x_i)] < L$, $\sigma_{\max}[G(\theta)] < \overline{\sigma} < \infty$, for all $\theta \in \Theta$ (iv) for some*

$R_G > 0$ such that $\mathcal{B}_{R_G}(\theta^\dagger) \subset \Theta$, $\sigma_{\min}[G(\theta)] > \underline{\sigma} > 0$ for all $\|\theta - \theta^\dagger\| > R_G$, (v) Θ is convex and compact, and (vi) $W_n \xrightarrow{p} W$, $0 < \underline{\lambda}_W < \lambda_{\min}(W) \leq \lambda_{\max}(W) < \overline{\lambda}_W < \infty$.

Assumption 1 gives standard conditions for global and local identification as well as uniform convergence of the sample moments. The quantity $\sigma_{\min}[G(\theta)]$ in the local identification condition refers to the smallest singular value of $G(\theta)$.[2] The main Assumption 2 below will rely on the following quantities:

$$\overline{G}(\theta) = \int_0^1 G(\omega\theta + (1-\omega)\theta^\dagger)d\omega, \quad \overline{G}(\theta_1, \theta_2) = \int_0^1 G(\omega\theta_1 + (1-\omega)\theta_2)d\omega.$$

The matrix $\overline{G}(\theta)$ measures the average Jacobian between the solution $\theta^\dagger$ and θ. The main results of the paper rely on a mean-value identity, found in Lemma B.11, which states that $g(\theta_1) - g(\theta_2) = \overline{G}(\theta_1, \theta_2)(\theta_1 - \theta_2)$ for any $\theta_1, \theta_2 \in \Theta$.

Assumption 2. *There exists $0 < \rho < \underline{\sigma}\underline{\lambda}_W/2$ such that, for all $\theta \in \Theta$, (a) $\sigma_{\min}[G(\theta)'W\overline{G}(\theta)] > \rho\underline{\sigma}$, or (b) $\|G(\theta)'W\overline{G}(\theta)(\theta - \theta^\dagger)\| > \rho\underline{\sigma}\|\theta - \theta^\dagger\|$.*

Assumption 2 gives the main conditions used in this paper for global GMM estimation of just and over-identified models. The factor ρ is assumed to be set, without loss of generality, such that $\sigma_{\min}[\overline{G}(\theta)] > \underline{\sigma}$ under (a) and $\|\overline{G}(\theta)(\theta - \theta^\dagger)\| > \underline{\sigma}\|\theta - \theta^\dagger\|$ under (b) for $\underline{\sigma}$ found in Assumption 1, also set sufficiently small for these normalizations to hold. Assumption 2 (a) replaces the convexity condition that $0 < \underline{\lambda}_H \leq \lambda_{\min}[H_n(\theta)] \leq \lambda_{\max}[H_n(\theta)] < \overline{\lambda}_H < \infty$ used to derive convergence results for GD, NR and QN.[3], which may not hold for GMM, as the MA(1) example illustrates. Assumption 2 (a) implies Assumption 2 (b); the latter is the weaker condition. Assumption 2 (a) implies that $G(\theta)$ has full rank for all θ, Assumption 2 (b) only

---

[2] For a rectangular matrix G of size $n \times m$, $m < n$, the singular values are given by $\sigma_j[G] = \sqrt{\lambda_j(G'G)} \geq 0$, where λ_j are eigenvalues; $G'G$ is a square matrix of size $m \times m$.

[3] See Nesterov (2018, pp33-35), especially equations (1.2.25), (1.2.27) and Theorem 1.2.4 for GD.

requires $G(\theta)'W\overline{G}(\theta)$ to be non-singular in the relevant direction $(\theta - \theta^\dagger)$. For over-identified models, both conditions (a) and (b) depend on the choice of weighting matrix W. Indeed, unlike square matrices, the product of full rank rectangular matrices does not automatically have full rank, and the weighting matrix changes the way G and $\overline{G}$ are multiplied, the Assumption may or may not hold depending on the choice of W.[4] Assumption 2 is robust to one-to-one affine transformations of the parameters, i.e. $\vartheta = A + B\theta$ with B invertible. Assumption 2 (a) is also robust to moderately non-linear one-to-one transformations of the parameters $\vartheta = h(\theta)$. This will be shown in the next Section.

Assumption 1 implies that Assumption 2 (a) holds locally, i.e. over a neighborhood of $\theta^\dagger$. This is shown in Lemma B.13. The condition simply requires that it holds globally rather than locally. Assumption 2 is related to the global and local identification conditions. For instance, Assumption 2 (a) implies Assumption 1 (iv). Futher discussion of Assumption 2 will be provided in the next section. When both Assumptions 1 and 2 holds, Assumption 2 also holds for the sample analogs:

$$\overline{G}_n(\theta) = \int_0^1 G_n(\omega\theta + (1-\omega)\hat{\theta}_n)d\omega, \quad \overline{G}_n(\theta_1, \theta_2) = \int_0^1 G_n(\omega\theta_1 + (1-\omega)\theta_2)d\omega,$$

with probability approaching 1, this is shown in Lemma B.14. The conditions can be checked on the sample moments and their Jacobian numerically, when Assumption 2 cannot be verified analytically. This is discussed further below.

Assumption 3. *With probability approaching 1: P_k is such that:*

$$0 < \underline{\lambda}_P \leq \lambda_{\min}(P_k) \leq \lambda_{\max}(P_k) \leq \overline{\lambda}_P < \infty.$$

---

[4] Take $G(\theta_1)' = (1,0)$ and $G(\theta_2)' = (0,1)$, both have full rank and yet $G(\theta_1)'G(\theta_2) = 0$ is singular.

Assumption 3 requires P_k to be finite and strictly positive definite. This is always the case for GD since $P_k = I_d$, and holds for GN under Assumption 2 (a). Under Assumption 2 (b), Assumption 3 may not hold for GN but remains valid for GD. One can apply the Levenberg-Marquardt (LM) algorithm to GN by setting $P_k = (G_n(\theta_k)' W_n G_n(\theta_k) + \lambda I_d)^{-1}$, and Assumption 3 holds since $\underline{\lambda}_P \geq \lambda > 0$, by design.

2.2.1 Correctly-specified models

The first set of results concerns models that are correctly specified: the minimizer $\hat{\theta}_n$ is such that $\overline{g}_n(\hat{\theta}_n) = O_p(n^{-1/2})$. The following Proposition shows that for any tuning parameter γ, there exists a neighborhood where (2.1) is locally convergent.

Proposition 2 (Local Convergence). *If Assumptions 1, 3 hold, then for $\gamma \in (0,1)$ small enough, there exist $R_n \geq 0$ and $\tilde{\gamma} \in (0,1)$ such that with probability approaching 1:*

$$\|\theta_{k+1} - \hat{\theta}_n\| \leq (1-\tilde{\gamma})\|\theta_k - \hat{\theta}_n\| \leq \cdots \leq (1-\tilde{\gamma})^{k+1}\|\theta_0 - \hat{\theta}_n\| \tag{2.2}$$

for any $\|\theta_0 - \hat{\theta}_n\| \leq R_n$. For just-identified models, $\overline{g}_n(\hat{\theta}_n) = 0$, $R_n > 0$ with probability 1. For over-identified models, $\overline{g}_n(\hat{\theta}_n) = O_p(n^{-1/2})$, $R_n > 0$ with probability approaching 1.

The analysis differs from standard results in the literature for non-linear systems of equations (e.g. Dennis & Schnabel, 1996; Nocedal & Wright, 2006, Ch11). First, the system can be over-identified and is not required to have an exact solution. Both are particularly relevant to GMM estimations. When $\overline{g}_n(\theta) = 0$ does not have a solution, existing results do not apply. Second, the area of local convergence R_n is tied to a) the choice of tuning parameter γ, and b) the size of the moments at the solution $\overline{g}_n(\hat{\theta}_n)$. This will be important for global convergence with over-identified and misspecified models.

GN is often used for non-linear least squares estimations. Results also rely on a full rank condition for the Jacobian around the solution to show that θ_k with $\gamma_k = 1$ solves a first-order condition asymptotically as k increases (e.g. Nocedal & Wright, 2006, Th10.1). Interpreting those results as a global convergence property requires strong convexity, however.

For GN, the coefficient $\tilde{\gamma} \in (0, \gamma)$ is needed to have $R_n > 0$, as illustrated below. Because of scaling, other methods may allow $\tilde{\gamma} > \gamma$. A proof specialized to GN, and the general case are given in Appendix B.1. For GN, $R_n = \min(R_G, \tilde{R}_n)$ is the smallest of R_G and:

$$\tilde{R}_n = (1 - \tilde{\gamma}/\gamma)\frac{\underline{\sigma}}{L\sqrt{\kappa_W}} - \frac{1}{\underline{\sigma}\sqrt{\underline{\lambda}_W}}\|\overline{g}_n(\hat{\theta}_n)\|_{W_n},$$

where $\kappa_W = \overline{\lambda}_W/\underline{\lambda}_W$ bounds the condition number of the weighting matrix W_n. Having $\|\overline{g}_n(\hat{\theta}_n)\|_{W_n} \neq 0$ reduces the area of local convergence. For correctly specified models $\overline{g}_n(\hat{\theta}_n) = O_p(n^{-1/2})$ implies $\tilde{R}_n \xrightarrow{p} \tilde{R} = (1 - \tilde{\gamma}/\gamma)\underline{\sigma}/(\sqrt{\kappa_W}L) > 0$. Note that for GN, Proposition 2 holds for any choice of $\gamma \in (0, 1)$. This is typically not the case for other choices of P_k: GD is only locally convergent when γ is sufficiently small. GD and GN iterations require the same inputs G_n and $\overline{g}_n$, but the latter is preferred since it converges more quickly. As NR and QN iterations require an exact or approximate Hessian, they are more costly than GD, GN.

The expression for $\tilde{R}_n$ illustrates that the choice of weighting matrix W_n matters. An ill-conditioned W_n, $\kappa_W \gg 1$, can make local optimization challenging. When the sample moments are highly correlated, the optimal weighting matrix can be ill-conditioned. Using a diagonal weighting matrix, as commonly done in practice, may improve numerical stability.

2.2.1.1 *Just-Identified Models*

The Theorem below proves global convergence for $\gamma \in (0,1)$ sufficiently small.

Theorem 2 (Global Convergence, Just-Identified)**.** *Suppose* $\overline{g}_n(\hat{\theta}_n) = 0$, *Assumptions 1, 2, 3 hold, then for* γ *small enough, there exist a* $\overline{\gamma} \in (0,1)$, *and* $0 < \underline{\lambda} \leq \overline{\lambda} < \infty$ *such that:*

$$\|\theta_{k+1} - \hat{\theta}_n\| \leq (1-\overline{\gamma})^{k+1}\sqrt{\overline{\lambda}/\underline{\lambda}}\|\theta_0 - \hat{\theta}_n\|, \tag{2.3}$$

for any starting value $\theta_0 \in \Theta$, *with probability approaching* 1.

The proof is given in Appendix B.1. The main steps are to show that for γ sufficiently small, we have: i) $Q_n(\theta_{k+1}) \leq (1-\overline{\gamma})^2 Q_n(\theta_k)$ for some $\overline{\gamma} \in (0,1)$ under the assumptions. Iterating on this inequality implies convergence of the objective function: $Q_n(\theta_k) \leq (1-\overline{\gamma})^{2k} Q_n(\theta_0)$. The same assumptions also imply the norm equivalence: ii) $\underline{\lambda}\|\theta - \hat{\theta}_n\|^2 \leq Q_n(\theta) \leq \overline{\lambda}\|\theta - \hat{\theta}_n\|^2$ for some $0 < \underline{\lambda} \leq \overline{\lambda} < +\infty$. Together, these two properties imply convergence of θ_k to $\hat{\theta}_n$.

The main takeaway from Theorem 2 is that global convergence can be achieved using any Algorithm which has P_k strictly positive definite for any $k \geq 0$, with an adequate choice of $\gamma \in (0,1)$, if G_n is everywhere non-singular. This assumption does not imply the convexity of Q_n. Note that the choice of γ depends on the choice of algorithm, through P_k. Some methods are associated with faster convergence than others, which is measured by $\overline{\gamma}$.

2.2.1.2 Over-Identified Models

Theorem 3 (Global Convergence, Over-Identified). *Suppose* $\overline{g}_n(\hat{\theta}_n) = O_p(n^{-1/2})$*, Assumptions 1, 2, 3, then for* γ *small enough, there exist* $\overline{\gamma} \in (0,1)$*,* $C > 0$*,* $0 < \underline{\lambda} \leq \overline{\lambda} < \infty$*, and* $C_n = O_p(1)$ *such that with probability approaching* 1*:*

$$\|\theta_k - \hat{\theta}_n\|^2 \leq (1-\overline{\gamma})^{2k} \frac{\overline{\lambda} + C\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}}{\underline{\lambda} - C\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}} \|\theta_0 - \hat{\theta}_n\|^2 + C_n\|\overline{g}_n(\hat{\theta}_n)\|^2_{W_n}, \tag{2.4}$$

for any $\theta_0 \in \Theta$*. Given this choice of* γ*, take* R_n *from Proposition 2. Since* $C_n\|\overline{g}_n(\hat{\theta}_n)\|^2_{W_n} \leq R_n^2/2$ *with probability approaching* 1*, setting* $k = k_n + j$*,* $j \geq 0$ *implies:*

$$\|\theta_k - \hat{\theta}_n\| \leq (1-\tilde{\gamma})^j R_n,$$

where $\tilde{\gamma} \in (0,1)$ *is the local rate in Proposition 2 and* $k_n \geq \frac{2\log(R_n) - \log 2 - \log(d_{0n})}{2\log(1-\overline{\gamma})}$ *with* $d_{0n} = 2[\underline{\lambda} - C\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}]^{-1}[\|\overline{g}_n(\theta_0)\|^2_{W_n} - \|\overline{g}_n(\hat{\theta}_n)\|^2_{W_n}]$.

Explicit formula for $\underline{\lambda}, \overline{\lambda}, C$, and C_n are given in the proof of the Theorem (Appendix B.1). Notice that larger values of $\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}$ can degrade convergence. The results for misspecified models further investigate the case where $\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}$ does not vanish in the limit.

Pen and pencil example. Take $y_i \sim \mathcal{N}(\mu, \sigma^2)$, the parameters of interest are $\theta = (\mu, \sigma^2)$. Compute the sample moments $\hat{\mu}_n = (\hat{\mu}_{n1}, \hat{\mu}_{n2}, \hat{\mu}_{n4})'$, where $\hat{\mu}_{n1} = \overline{y}_n$, $\hat{\mu}_{n2} = \hat{\sigma}^2_n$, and $\hat{\mu}_{n4} = 1/n\sum_{i=1}^n (y_i - \overline{y}_n)^4$, and let: $\overline{g}_n(\theta) = \hat{\mu}_n - (\mu, \sigma^2, 3\sigma^4)$. Set $W_n = I_d$ and take $\theta^\dagger = (0,1)$. A quick numerical computation reveals the population objective function is non-convex: the eigenvalues of $\partial^2_{\theta,\theta'}Q(\theta)$ are $(74, 2)$ at $\theta = (0,1)$ and $(2, -7)$ at $\theta = (0, 1/2)$ – the Hessian is positive definite at the true value but not everywhere. For starting values such that $\partial^2_{\theta,\theta'}Q_n(\theta)$ is (near)-singular, NR and

QN iterations can be erratic, as in the MA(1) example below. Nonetheless, some calculations imply that:

$$G(\theta)' = \begin{pmatrix} -1 & 0 & 0 \\ \\ 0 & -1 & -6\sigma^2 \end{pmatrix}, \quad G(\theta)'\overline{G}(\theta) = \begin{pmatrix} 1 & 0 \\ \\ 0 & 1 + 18\sigma^2\{\sigma^2 + \sigma^{\dagger 2}\} \end{pmatrix}$$

is positive definite for any two $\theta = (\mu, \sigma^2)$, with $\theta^\dagger = (\mu^\dagger, \sigma^{\dagger 2})$. In this simple example, the Hessian of Q_n can be singular, yet Assumption 2 (a) holds.

2.2.2 Misspecified models

So far, the results imply that fast global convergence is feasible under a rank condition for correctly specified models. In applications, misspecification can be a concern so that understanding the robustness of the results above to non-negligible deviations from this baseline is empirically relevant. Recently, Hansen & Lee (2021) studied the properties of iterated GMM procedures. Here the focus is on computing a single GMM estimate. The following considers "moderate" amounts of misspecification in the sense that:

$$\text{plim}_{n\to\infty} Q_n(\hat{\theta}_n) := \varphi^2 \geq 0$$

exists and can be non-zero in the limit. When $\varphi > 0$, the degree of misspecification is non-negligible asymptotically and the statistic $n\|\overline{g}_n(\hat{\theta}_n)\|^2_{W_n} \to \infty$ can diverge. However, φ cannot be too large for the local and global convergence results to hold as shown below. For simplicity, only Gauss-Newton will be considered in the results. Also, since G_n cannot be full rank at $\theta = \hat{\theta}_n$ when the model is both just-identified and misspecified,[5] the results presented here solely consider over-identified models.

For correctly specified models, a test for over-identifying restrictions can diagnose global convergence (Andrews, 1997, Sec3.3). For misspecified models, such test would frequently reject in large samples. Then the issue is that, when the test rejects, either 1) the optimizer has not found valid estimates, or 2) the model fits the data poorly in some dimension(s). When Q_n is globally convex, a given value is the global solution if, and only if, it satisfies the first and second-order optimality conditions.[6] Without convexity, this only guarantees a local optimum. For moderately misspecified models, Assumption 2 provides an alternative to convexity in these settings.

Proposition 3 (Local Convergence, Misspecified)**.** *Suppose Assumptions 1, 2, 3 hold, and* φ *is such that:*

$$0 \leq \varphi < \frac{\sigma^2 \underline{\lambda}_W}{L \overline{\lambda}_W^{1/2}}. \tag{2.5}$$

For any $\gamma \in (0,1)$*, there exists* $\tilde{\gamma} \in (0,\gamma)$*, such that, with probability approaching* 1*, for some* $R_n > 0$*, strictly positive, any* $\|\theta_0 - \hat{\theta}_n\| \leq R_n$*, and all* $k \geq 0$*:*

$$\|\theta_{k+1} - \hat{\theta}_n\| \leq (1-\tilde{\gamma})\|\theta_k - \hat{\theta}_n\| \leq \cdots \leq (1-\tilde{\gamma})^{k+1}\|\theta_0 - \hat{\theta}_n\|. \tag{2.2'}$$

Also, $\text{plim}_{n\to\infty} R_n = R > 0$*.*

R_n in Proposition 3 takes the same form as in Proposition 2, (2.5) ensures that

---

[5] The solution $\hat{\theta}_n$ is s.t. $G_n(\hat{\theta}_n)' W_n \overline{g}_n(\hat{\theta}_n) = 0$, misspecification implies $\overline{g}_n(\hat{\theta}_n) \neq 0$, and since W_n has full rank, it must be that $G_n(\hat{\theta}_n)$ is singular for just-identified models. For over-identified models, $\overline{g}_n(\hat{\theta}_n)$ is in the null space of $G_n(\hat{\theta}_n)' W_n$, which allows $G_n(\hat{\theta}_n)$ to be full rank.

[6] The first is $\partial_\theta Q_n(\hat{\theta}_n) = 0$ and the second $\partial^2_{\theta,\theta'} Q_n(\hat{\theta}_n)$ positive semidefinite.

the corresponding R is strictly positive in the limit; the neighborhood of convergence is non-negligible asymptotically. Under identity weighting, $W_n = I_d$, (2.5) only depends on $\underline{\sigma}$ and L. For linear models, $L = 0$ implies that any $\varphi \in [0, \infty)$ is feasible. For non-linear models, a larger $L > 0$ requires a smaller φ: increased non-linearity requires milder misspecification. Smaller values of $\underline{\sigma}$, which measures local identification, also require a smaller φ. In Proposition 2 with GN, any rate of convergence $\overline{\gamma} \in (0, \gamma)$ can be used. When the model is misspecified, larger values of $\varphi \geq 0$ require $\tilde{\gamma} \in (0, \gamma)$ to be smaller, resulting in slower convergence.

Theorem 4 (Global Convergence, Misspecified)**.** *Suppose Assumptions 1, 2, 3 hold. If φ is such that:*

$$0 \leq \varphi < \frac{\underline{\sigma}^2 \underline{\lambda}_W}{2L\overline{\lambda}_W^{1/2}}. \tag{2.5'}$$

Then for γ small enough, there exist $\overline{\gamma} \in (0,1)$ and $C_n = O_p(1)$, $0 < C, \underline{\lambda}, \overline{\lambda} < \infty$, which do not depend on φ, such that with probability approaching 1:

$$\|\theta_k - \hat{\theta}_n\|^2 \leq (1-\overline{\gamma})^{2k} \frac{\overline{\lambda} + C\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}}{\underline{\lambda} - C\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}} \|\theta_0 - \hat{\theta}_n\|^2 + C_n \|\overline{g}_n(\hat{\theta}_n)\|_{W_n}^2. \tag{2.4'}$$

Let $\Delta = 1/2[\underline{\sigma}^2 \underline{\lambda}_W - 2L\overline{\lambda}_W^{1/2}\varphi] > 0$. Suppose $\gamma \in (0,1)$ and $\varphi \geq 0$ are such that:

$$\frac{\Delta\gamma^2 c_2 + 2\gamma \overline{c}_3^2/\underline{c}_1}{[\gamma \underline{c}_1/2 - \gamma^2 c_2]\Delta^2}\varphi^2 < \left((1-\varepsilon)\frac{\sigma}{L\sqrt{\kappa_W}} - \frac{\varphi}{\underline{\sigma}\sqrt{\underline{\lambda}_W}} \right)^2, \tag{2.6}$$

for some $\varepsilon \in (0,1)$, where $\underline{c}_1 = 2/3\rho^2([\underline{\sigma}/\overline{\sigma}]^2 \kappa_W^{-1})^2$, $c_2 = L_Q(\overline{\sigma}\overline{\lambda}_W^{1/2}/[\underline{\sigma}^2 \underline{\lambda}_W])^2$, $\overline{c}_3 = 2\overline{\sigma}\overline{\lambda}_W^{1/2}$, $\kappa_W = \overline{\lambda}_W / \underline{\lambda}_W$, and L_Q is the Lipschitz constant of $\partial_\theta Q_n$. Take any $\tilde{\gamma} \in (0, \varepsilon\gamma)$

and R_n *from Proposition 3, set* $k = k_n + j$, $j \geq 0$, *then with probability approaching* 1*:*

$$\|\theta_k - \hat{\theta}_n\| \leq (1 - \tilde{\gamma})^j R_n,$$

where $k_n \geq \frac{2\log(R_n)+\log(\delta)-\log(d_{0n})}{2\log(1-\overline{\gamma})}$, *with* d_{0n} *as in Theorem 3, for some small enough* $\delta \in (0,1)$.

Another way to read Theorem 4 is that global convergence is not guaranteed when the model is heavily misspecified and highly non-linear, i.e. both φ and L are very large. Again, the choice of weighting matrix matters, as φ depends on W.

2.3 ASSUMPTION 2 AND ITS RELATION TO THE LITERATURE

Convexity, monotonicity and the Polyak-ojasiewicz condition. The following briefly reviews some convexity conditions found in the literature and an important relaxation called the Polyak-ojasiewicz (PL) condition. The latter has gathered much attention in the machine learning literature in recent years. Because Assumption 2 is stated on population quantities, the following discussion will focus on Q.

For general minimization of an objective Q, GD, NR and QN are globally convergent for $\theta^\dagger$ if Q is *μ-strongly convex*, i.e. if for some $\mu > 0$:

$$Q(\theta_2) \geq Q(\theta_1) + \partial_\theta Q(\theta_1)(\theta_1 - \theta_2) + \frac{\mu}{2}\|\theta_1 - \theta_2\|^2,$$

for all $\theta_1, \theta_2 \in \Theta$. When Q is twice continuously differentiable it is strongly convex if its Hessian $H(\theta) = \partial^2_{\theta,\theta'} Q(\theta)$ is strictly positive definite everywhere. In particular, for $\gamma > 0$ sufficiently small, we have:

$$Q(\theta_{k+1}) - Q(\theta^\dagger) \leq (1-\eta)\left(Q(\theta_k) - Q(\theta^\dagger)\right),$$

for some $\eta \in (0,1)$ which depends on γ, the choice of algorithm, i.e. P_k, and the eigenvalues of H. Iterating on this inequality indicates that the fit improves rapidly from any starting value θ_0: $Q(\theta_{k+1}) - Q(\theta^\dagger) \leq (1-\eta)^{k+1}\left(Q(\theta_0) - Q(\theta^\dagger)\right)$. Under strong convexity, Q has a unique global minimizer and no local optima.

The literature has considered a number of relaxations of strong convexity under which GD is globally convergent. This includes the so-called *star convexity* condition due to Nesterov & Polyak (2006):

$$Q(\theta^\dagger) \geq Q(\theta) + \lambda \partial_\theta Q(\theta)(\theta^\dagger - \theta) + \frac{\mu}{2}\|\theta^\dagger - \theta\|^2$$

for some $\mu \geq 0$ and $\lambda = 1$. This is similar-looking to strong convexity but only involves the pairs $(\theta_1, \theta_2) = (\theta, \theta^\dagger)$. For these functions, the convexity property only holds on line segments toward $\theta^\dagger$. Star convexity implies that $\theta^\dagger$ is the unique global minimizer of Q. This condition can be further weakened to *quasar convexity*, which allows for $\lambda > 1$ in the inequality above. Hinder et al. (2020, Figure 1) plot examples of functions that satisfy these conditions but are not strongly convex. Note that Assumption 2 is similarly stated for averages $\overline{G}$ of G on line segments between θ and $\theta^\dagger$.

Karimi et al. (2016), Guminov et al. (2017) showed that a number of relaxations of strong convexity imply the so-called *Polyak-Łojasiewicz* (PL) inequality, after Polyak (1963) and Lojasiewicz (1963), which requires that:

$$\|\partial_\theta Q(\theta)\|^2 \geq \mu \left(Q(\theta) - Q(\theta^\dagger) \right), \tag{PL}$$

for all $\theta \in \Theta$ and some $\mu > 0$. When Q satisfies the PL inequality, $\partial_\theta Q(\theta) = 0$ implies θ is globally optimal, i.e. $Q(\theta) = Q(\theta^\dagger)$. The global minimizer may not be unique, however, unlike strong convexity. If the PL inequality holds and $\partial_\theta Q$ is Lipschitz continuous, it can be shown that for $\gamma > 0$ small enough: $Q(\theta_{k+1}) - Q(\theta^\dagger) \leq (1-\eta)\left(Q(\theta_k) - Q(\theta^\dagger)\right)$ for GD (Karimi et al., 2016, Th1). This does not imply that θ_{k+1} converges to $\theta^\dagger$, however, unless the minimizer is unique. Because strong convexity implies the PL inequality, Karimi et al. (2016) argue that the latter holds locally over a larger area than strong convexity, predicting better optimization performance. They also note that it is difficult to characterize which functions satisfy the PL inequality. They show that $Q(\theta) = h(A\theta)$, with h strongly convex and A a non-zero matrix, satisfies the PL inequality.

Closely related to the GMM setting, a smaller literature has considered conditions for solving non-linear systems of equations of the form: $g(\theta) = 0$, typically with g and θ of the same dimension. An important reference is Dennis & Schnabel (1996), who cast the problem as minimizing $Q(\theta) = \|g(\theta)\|^2$, similar to GMM, and derive global convergence results to a local minimum under convexity conditions (Theorems 6.3.3-6.3.4). Deuflhard (2005, Ch3) studies global convergence under different assumptions. His related global convergence result, Theorem 3.7, implicitly assumes linearity of g in the proof, however. For just and under-determined systems, several authors derived global convergence results under a *strong monotonicity condition*:

$$(g(\theta_1) - g(\theta_2))'(\theta_1 - \theta_2) \geq \mu\|\theta_1 - \theta_2\|^2,$$

with $\mu > 0$ e.g. Solodov & Svaiter (2000), Polyak & Tremba (2020), Heid (2023). Note that when $g = \partial_\theta F$, then g is strongly monotone if, and only if, F is strongly convex. Hence, global convergence under strong monotonicity is related to global convergence under strong convexity of F. In that case, g is said to be cyclically monotone (Rockafellar, 2015, p238). The results listed above do not consider problems where $g(\theta^\dagger) \neq 0$ which is particularly relevant to sample GMM estimations with overidentifying moment restrictions.

In a companion paper, Forneron (2023) considers GMM estimation with potentially non-smooth but correctly specified sample moments when only conditions analogous to Assumption 1 are assumed. There are two important differences that are inherant to that setting: 1) the sample Jacobian G_n is not defined, so that (2.1) is not directly applicable, and 2) even with infinite data, (2.1) may not minimize Q since the assumptions do not exclude non-global optima. Unlike here, the curse of dimensionality appears in the converge results for $\|\theta_k - \hat{\theta}_n\|$ but is much less pronounced than worse-case lower bounds found in the literature for general purpose global optimizers.

Relation between the different conditions. Narrowing to the GMM setting specifically, the following shows that the PL inequality holds in the *in the population for correctly specified models* under Assumption 2. A related result is derived under misspecification.

As discussed above, Assumption 2 (a) implies Assumption 2 (b). The latter confers most of the properties required for minimizing Q. It can be useful to rewrite the condition in terms of g: Assumption 2 (b) $\|G(\theta)'W[g(\theta) - g(\theta^\dagger)]\| > \rho\underline{\sigma}\|\theta - \theta^\dagger\|$. For correctly specified models, $g(\theta^\dagger) = 0$ and $G(\theta)'Wg(\theta) = \partial_\theta Q(\theta)$. The lower bound implies that the only critical point is $\theta = \theta^\dagger$. This excludes local minima, maxima, and saddle points.[7]

Proposition 4 (Correct Specification). *Suppose Assumptions 1 (ii), (iii), (vi), 2 (b) hold and* $Q(\theta^\dagger) = 0$*, then there exists strictly positive constants* C_1, C_2, C_3 *such that for all* $\theta \in \Theta$*:*

$$\begin{aligned} &(1) \quad \|\partial_\theta Q(\theta)\|^2 \geq C_1 \left(Q(\theta) - Q(\theta^\dagger)\right) \\ &(2) \quad C_2\|\theta - \theta^\dagger\|^2 \leq Q(\theta) - Q(\theta^\dagger) \leq C_3\|\theta - \theta^\dagger\|^2. \end{aligned}$$

Proposition 4 shows that Assumption 2 (b), together with bounds on W and Lipschitz continuity of G imply the PL inequality (1) for Q. In addition, (2) implies global identification and is needed to derive the convergence rate of θ_k. Strong convexity also implies (1) and (2).

Proposition 5. *Suppose* W *is invertible,* $Q(\theta^\dagger) = 0$*. The following holds: 1) If* Q *satisfies the PL inequality with* $\mu > 0$ *and* $C_2\|\theta - \theta^\dagger\|^2 \leq Q(\theta) - Q(\theta^\dagger)$ *for* $C_2 > 0$ *and all* $\theta \in \Theta$*, then Assumption 2 (b) holds. 2) If* Q *is quasar-convex with* $\mu > 0$*, then Assumption 2 (b) holds.*

Proposition 5 gives a condition under which quasar-convexity and the PL inequality imply Assumption 2 (b). On compact sets, Assumption 1 (i), (iii), (iv) together imply a $C_2 > 0$ exists for correctly specified models. Assumption 2 (b) does not imply quasar-convexity.[8] The following considers strong monotonicity and introduces a *strong injectivity* condition:

$$\|g(\theta_1) - g(\theta_2)\| \geq \mu\|\theta_1 - \theta_2\|.$$

---

[7] A critical point is a θ such that $\partial_\theta Q(\theta) = 0$. Assuming Q is twice differentiable, it is a local minimum if $\partial^2_{\theta,\theta'}Q(\theta)$ is positive semidefinite, maximum if $\partial^2_{\theta,\theta'}Q(\theta)$ is negative semidefinite, and a saddle point if $\partial^2_{\theta,\theta'}Q(\theta)$ is indefinite, i.e. has both positive and negative eigenvalues.

[8] Quasar-convexity implies $(\theta - \theta^\dagger)'G(\theta)'W\overline{G}(\theta)(\theta - \theta^\dagger) \geq \frac{\mu}{2\lambda}\|\theta - \theta^\dagger\|^2$. This is more restrictive than Assumption 2 (b).

It can be shown that the strong injectivity property holds on compact convex sets under the Gale-Nikaidô-Fisher-Rothenberg conditions: det$(G(\theta)) > 0$ and $G(\theta)$ is positive quasi-definite, for all $\theta \in \Theta$, where det is the determinant.[9]

Proposition 6 (Just-Identified). *1) If Ag is strongly monotone for some invertible matrix A and $\mu > 0$, then Assumption 2 (b) holds. 2) If g is strongly injective with $\mu > 0$, then Assumption 2 (b) holds.*

Figure 2.1: Relationship between conditions for correctly specified models

$$\begin{tikzcd}
\text{strong convexity} \arrow[r, Rightarrow] & \text{star convexity} \arrow[r, Rightarrow] & \text{quasar convexity} \arrow[d, Rightarrow] & \\
& \text{Assumption 2 (a)} \arrow[r, Rightarrow] & \text{Assumption 2 (b)} \arrow[d, Leftrightarrow] & \text{strong injectivity} \arrow[l, Rightarrow] \\
& & \text{(PL) + QLB} & \text{strong monotonicity} \arrow[u, Rightarrow]
\end{tikzcd}$$

Legend: Relations hold when $Q(\theta^\dagger) = 0$. QLB = Quadratic Lower Bound, i.e. $C_2\|\theta - \theta^\dagger\|^2 \leq Q(\theta) - Q(\theta^\dagger)$ for some $C_2 > 0$. Relations with strong monotonicity and strong injectivity are for just-identified models.

Figure 2.1 summarizes the results of Propositions 4, 5, 6. Since $Q_n(\hat{\theta}_n) = 0$ for just-identified models that are correctly specified, the relationship also applies in the finite samples problems where these conditions are met. When g and θ are scalar, Assumption 2 implies strict monotonicity, g is either increasing or decreasing, but does not imply convexity of Q, however, as the MA example below will illustrate. Also, Assumptions 2 (a) and (b) are equivalent in the scalar case so that strong, star, and quasar convexity imply Assumption 2 (a) in that particular setting.

---

[9] G is positive quasi-definite if, and only if, $G + G'$ is positive definite. See Fisher (1966), Rothenberg (1971); and Komunjer (2012) for a discussion and alternative conditions.

It remains to determine if Assumption 2 (b) is minimal for global convergence, or if can be weakened further. The following condition is *necessary* for GD and other gradient-based optimizers of the form (2.1) to be globally convergent:

$$\partial_\theta Q(\theta) = 0 \Leftrightarrow \theta = \theta^\dagger. \tag{N}$$

The following shows that, under smoothness and local identification conditions, (N) implies Assumption 2 (b).

Proposition 7. *Suppose condition (N) and Assumption 1 (ii)-(vi) hold, then Assumption 2 (b) holds on any compact convex set containing* $\theta^\dagger$*.*

Proposition 7 implies that given standard regularity conditions, Assumption 2 (b) is a necessary condition on compact sets. The case of overidentified and misspecified models, is more complicated as the following shows that the equivalence between the PL inequality and Assumption 2 (b) is not automatic.

Proposition 8 (Misspecification)**.** *Suppose Assumptions 1 (ii), (iii), (vi), 2 (b) hold and* $Q(\theta^\dagger) = \varphi > 0$*, then there exists strictly positive constants* C_2, C_3, C_4 *such that for all* $\theta \in \Theta$*:*

$$(1) \quad \|\partial_\theta Q(\theta)\| \geq \left(\rho\underline{\sigma} - \sqrt{\varphi}\overline{\lambda}_W^{1/2} L\right) \|\theta - \theta^\dagger\|$$

$$(2) \quad (C_2 - C_4\sqrt{\varphi})\|\theta - \theta^\dagger\|^2 \leq Q(\theta) - Q(\theta^\dagger) \leq (C_3 + C_4\sqrt{\varphi})\|\theta - \theta^\dagger\|^2,$$

where C_2, C_3 *are the same as in Proposition 4 and* L *is the Lipschitz constant of* G *from in Assumption 1 (iii). If in addition* $\rho\underline{\sigma} - \sqrt{\varphi\overline{\lambda}_W}L > 0$*, then for all* $\theta \neq \theta^\dagger$*:*

$$(1') \quad \|\partial_\theta Q(\theta)\|^2 \geq \frac{(\rho\underline{\sigma} - \sqrt{\varphi\overline{\lambda}_W}L)^2}{C_3 + C_4\sqrt{\varphi}} \left(Q(\theta) - Q(\theta^\dagger)\right).$$

Proposition 8 (1) is only informative when the amount of misspecification is moderate, i.e. $\varphi < \rho^2\underline{\sigma}^2/[\overline{\lambda}_W L^2]$. When this holds, there are no local optima besides $\theta^\dagger$. It also implies the PL inequality (1′) holds. To recover convergence for θ, the lower bound in (2) should be informative which further requires $\sqrt{\varphi} < C_2/C_4$.[10] The degree of non-linearity - measured by L - and the choice of weighting matrix - measured by $\overline{\lambda}_W, \underline{\lambda}_W$ and φ - constrain the amount of misspecification permitted to get informative bounds. For correctly specified overidentified models, $Q_n(\hat{\theta}_n) = o_p(1)$ implies that (1′) and (2) hold asymptotically and are informative.

Further characterization of Assumption 2 (Just-Identified). Like star-convexity, Assumption 2 is stated relative to the unknown $\theta^\dagger$. The following Proposition gives several conditions under which Assumption 2 (a) holds and properties implied by these conditions.

Proposition 9. *(Sufficient Conditions) Consider the following conditions:*
(a) $\sigma_{\min}[\overline{G}(\theta_1, \theta_2)] > \underline{\sigma} > 0$, *for all* $\theta_1, \theta_2 \in \Theta$, *(b) for all* $\theta \in \Theta$, $G(\theta) = US(\theta)V$ *for* U, V *invertible and* $S(\theta)$ *symmetric with* $0 < \underline{\lambda}_S < \lambda_{\min}[S(\theta)] < \overline{\lambda}_S < \infty$, *for all* θ, *(c)* $g(\theta) = \partial_\theta F(\theta)$, *for all* $\theta \in \Theta$, *where* $F : \Theta \to \mathbb{R}$ *is twice continuously differentiable, strongly convex.*
The following holds: (1) (c) $\Rightarrow$ *(b)* $\Rightarrow$ *(a)* $\Rightarrow$ *Assumption 2 (a) holds; (2) (a) implies* $g(\cdot)$ *is one-to-one; (3) if (a) holds, there exists a reparameterization* $h(\cdot) = \psi \circ g \circ \phi(\cdot)$ *with* ϕ *one-to-one and* ψ *affine, such that* $1/2h(\theta)'Wh(\theta)$ *is strongly convex.*

Condition (a) does not require knowledge of $\theta^\dagger$ and implies that $g(\cdot)$ is one-to-one. The latter is often assumed for indirect inference.[11] Condition (a) also

---

[10] The derivations give the following bounds $C_2 = 1/2\frac{\rho^2\underline{\sigma}^2}{\overline{\sigma}^2\overline{\lambda}_W}$ and $C_4 = \overline{\lambda}_W^{1/2} L$ so that the condition reads $\sqrt{\varphi} < 1/2\rho^2\underline{\sigma}^2[\overline{\sigma}^2\overline{\lambda}_W^{3/2} L]^{-1}$. It is possible to relax this condition at the cost of more complicated derivations using a combination of global and local convergence arguments.

[11] See e.g. Gourieroux et al. (1993), Assumption (A4).

implies strongly injectivity with $\mu = \underline{\sigma}$. When the Jacobian can be linearly rearranged into a symmetric positive definite matrix $S(\theta) = U^{-1}G(\theta)V^{-1}$, then condition (a) holds. These problems can be thought of as *implicitly convex* in the special case where where S is the second derivative of a convex function. For a given $\theta \in \Theta$, decomposition (c) always exists: the singular value decomposition gives $G(\theta) = U(\theta)S(\theta)V(\theta)$ where $U(\theta), V(\theta)$ are unitary and $S(\theta)$ is diagonal with positive entries. A lesser known results, due to Frobenius (1910) shows that any square matrix can be written as the product of two real symmetric matrices; here $G(\theta) = S_1(\theta)S_2(\theta)$. The Jordan normal form of $G(\theta)$ can be used to compute this factorization (Bosch, 1986). If $G(\theta)$ is invertible, for all $\theta \in \Theta$, and U, V or one of S_1, S_2 do not vary with θ, in the singular value or Frobenius decomposition, then (c) holds. Under condition (b), g is cyclically monotone, and thus strongly monotone.

Proposition 10. *(Reparameterization) Take $h : \mathcal{U} \to \Theta$, one-to-one, continuously differentiable on $\mathcal{U}$ compact and convex, with $0 < \underline{\sigma}_h \leq \min_{u\in\mathcal{U}} \sigma_{\min}[\partial_u h(u)] \leq \max_{u\in\mathcal{U}} \sigma_{\max}[\partial_u h(u)] \leq \overline{\sigma}_h < \infty$. Let $u^\dagger = h^{-1}(\theta^\dagger)$, the minimizer of $Q \circ h$. Let $\overline{\sigma} = \sup_{\theta\in\Theta} \sigma_{\max}[G(\theta)]$ and:*

$$L_{1,h} = \sup_{u\in\mathcal{U}} \|\partial_u h(u) - \partial_u h(u^\dagger)\|$$

$$L_{2,h} = \sup_{u\in\mathcal{U},\omega\in[0,1]} \|h(\omega u + (1-\omega)u^\dagger) - \omega h(u) - (1-\omega)h(u^\dagger)\|.$$

If Assumption 2 (a) holds for g and $\underline{\sigma} > [L_{1,h}\overline{\sigma} + L_{2,h}L\overline{\sigma}_h]/\underline{\sigma}_h$, where L is the Lipschitz constant of G, then Assumption 2 (a) holds for $g \circ h$. In particular, if $h = Au + b$ is affine with A invertible then $L_{1,h} = L_{2,h} = 0$ and Assumption 2 (a) holds for $g \circ h$.

Strong convexity is preserved by affine transformations and reparameterization that satisfy particular component-wise monotonicity constraints on the reparameterization (e.g. Boyd & Vandenberghe, 2004, Sec3.2). Proposition 10 shows that Assumption 2 is also preserved by affine transformations and moderately nonlinear one-to-one reparameterizations h. Hencre, optimization should be locally robust to the choice of parameterization. Statements for overidentified models can be found in Propositions B.216, B.217.

Iteration depedent choice of learning rate γ_k. The results are stated for a fixed globally convergent choice of learning rate. In practice, adaptive choices of γ_k are common, using a line search for instance. If the adaptive algorithm is tuned to satisfy the requirements for global convergence, then it is also globally convergent. To preserve convergence properties, additional tuning parameters are typically involved (Nocedal & Wright, 2006, Ch3.1). A backtracking line search, a simple and popular way to set the learning rate (Nocedal & Wright, 2006, Ch3.1), is used as a benchmark comparison for the fixed learning rate used in the applications. It is described below.

Algorithm 2 Backtracking Line Search for Gauss-Newton

```
Tuning Parameters: Initial $\gamma_{\text{init}}$, $\rho \in (0,1)$, $c \in (0,1)$.
Inputs            : Previous iterate $\theta_k$, moments $\overline{g}_n(\theta_k)$, Jacobian $G_n(\theta_k)$
Compute           : Search direction: $p_k = (G_n(\theta_k)' W_n G_n(\theta_k))^{-1} G_n(\theta_k)' W_n \overline{g}_n(\theta_k)$,
                    $J_k = G_n(\theta_k)' W_n \overline{g}_n(\theta_k)$.
Set               : $\gamma_k = \gamma_{\text{init}}$ and $\theta_{k+1} = \theta_k - \gamma_k p_k$
while $Q_n(\theta_{k+1}) > Q_n(\theta_k) - c\gamma_k J_k' p_k$ do
    Set               : $\gamma_k = \rho\gamma_k$ and $\theta_{k+1} = \theta_k - \gamma_k p_k$
end
Output            : New iterate $\theta_{k+1}$, Learning Rate $\gamma_k$.
```

Setting $Q_n(\theta_{k+1}) = +\infty$ if $\theta_{k+1} \notin \Theta$ is outside the bounds.[12] This can occur when γ_{init} is too large to be globally convergent. By construction, $J_k' p_k \geq 0$ so that

---

[12] Another approach is to project θ_{k+1} inside Θ when γ_k is too large.

the final γ_k decreases the value of the objective function. The while loop terminates once the so-called *Armijo condition* is met:[13] $Q_n(\theta_{k+1}) \leq Q_n(\theta_k) - c\gamma_k J_k' p_k$. When the rank conditions below hold, the termination criterion is feasible for any $\theta_k \neq \hat{\theta}_n$ if c is *sufficiently small*.[14] Having $\theta_k = \hat{\theta}_n$ implies $p_k = 0$; the condition holds for any $\gamma_k \in (0, 1]$. A common choice is $c = 10^{-4}$, $\gamma_{\text{init}} = 1$, $\rho = 0.8$. These were used in all examples.

2.4 APPLICATIONS

2.4.1 A pen and pencil example: the MA(1) model

Now, to build intuition, consider a simple MA(1) process:

$$y_t = e_t - \theta^\dagger e_{t-1}, \quad e_t \overset{iid}{\sim} \mathcal{N}(0, 1), \quad \theta^\dagger \in (-1, 1),$$

for $t = 1, \dots, n$. $\theta^\dagger$ is the parameter of interest. Set $p \geq 1$, following Gourieroux & Monfort (1996, Ch4.3), $\theta^\dagger$ is estimated by matching coefficients from an auxiliary AR(p) model:

$$y_t = \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + u_t.$$

For $p = 1$, $\hat{\beta}_1 \overset{p}{\to} -\theta^\dagger/(1 + \theta^{\dagger 2})$ defines the moment condition:

$$\overline{g}_n(\theta) = \hat{\beta}_1 + \frac{\theta}{1 + \theta^2},$$

---

[13] See Nocedal & Wright (2006, p33), Nesterov (2018, pp28-29) for discussions.

[14] Consider, for instance, just-identified models: in Theorem 2 below, it is shown that $Q_n(\theta_{k+1}) \leq (1 - \overline{\gamma})^2 Q_n(\theta_k)$, for any $\theta_k \in \Theta$, when $\gamma \in (0, 1)$ small enough for some $\overline{\gamma} \in (0, \gamma)$. The rank conditions further imply that $Q_n(\theta_k)$ and $J_k' p_k$ are proportional. So the Armijo condition is feasible if c is small enough.

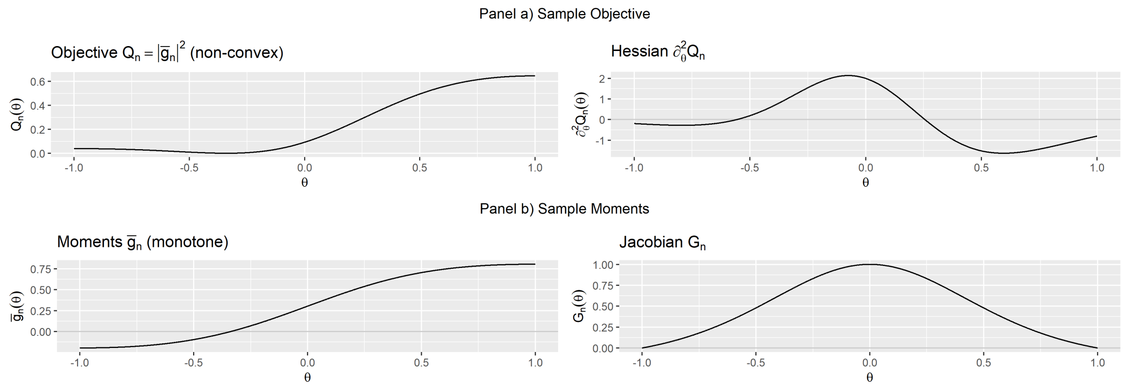
with Jacobian $G_n(\theta) = (1-\theta^2)/(1+\theta^2)^2 > 0$ for any $\theta \in (-1,1)$ and $G_n(\theta) = 0$ for $\theta \in \{-1,1\}$. It has full rank on any interval of the form $[-1+\varepsilon, 1-\varepsilon]$, $\varepsilon \in (0,1)$. However, Figure 2.2 shows that the Hessian $\partial^2_{\theta,\theta} Q_n(\theta)$ can be positive, negative, or equal to zero depending on the value of θ – Q_n is non-convex, especially when $\overline{g}_n(\theta)$ is large. Now notice that:

$$\overline{g}_n(\theta) = \partial_\theta F_n(\theta) \text{ where } F_n(\theta) = \hat{\beta}_1\theta + \frac{1}{2}\log(1+\theta^2)$$

which not a GMM objective but is nevertheless convex on $[-1,1]$, strongly convex on any $[-1+\varepsilon, 1-\varepsilon]$, $\varepsilon \in (0,1)$. The two, F_n and Q_n, are minimized at the same solution $\hat{\theta}_n$. From a statistical perspective, Q_n and F_n define identical M-estimates. However, one involves a convex minimization while the other does not. Notice that because the gradient of F_n is $\overline{g}_n$, and its Hessian is G_n, a NR update for F_n coincides with a GN update for $\overline{g}_n$. Implicitly, GN minimizes the convex F_n – whereas NR explicitly minimizes the non-convex Q_n. This change of loss function from Q_n to F_n is only illustrative of the connection between the two sets of conditions in the scalar case. It would be difficult to implement with multiple coefficients and is generally not feasible for overidentified models.

Table 2.2 shows the search paths for NR and GN with a fixed $\gamma = 0.1$ as well as R's built-in *optim*'s BFGS implementation and the bound-constrained L-BFGS-B. NR diverges, because the objective is locally concave at $\theta_0 = -0.6$. This is surprising given how close θ_0 is to the true value $\theta^\dagger$. GN converges steadily from the same staring value to $\hat{\theta}_n$. Although the GMM objective Q_n is locally convex around $\hat{\theta}_n$ which is useful for local optimization, the corresponding neighborhood can be fairly small from a practical standpoint. BFGS is more erratic, especially when $\theta_k \simeq -0.5$, i.e. $k = 1$, leading to a search outside the unit circle ($k = 2$), before reaching an area where the iterations are better behaved ($k = 3$ onwards). While here this is not too problematic, the objective function is well defined outside the bounds, this is more concerning in applications where the model cannot be solved outside the bounds – this is illustrated in Section 2.4.2. A natural solution is to introduce bounds using L-BFGS-B. The search, however, remains somewhat erratic as seen in the Table. Compare these to BFGS* and L-BFGS-B* which minimize F_n, instead of Q_n, using the same *optim*. Like GN, they steadily converge to $\hat\theta_n$.

Figure 2.2: MA(1): illustration of non-convexity and the rank condition



Legend: simulated sample of size $n = 200$, $\theta^\dagger = -1/2$, $\bar g_n(\theta) = \hat\beta_1 - \theta/(1+\theta^2)$, $W_n = I_d$. The GMM objective (panel a) is non-convex but the sample moments (panel b) satisfy the rank condition.

For $p = 12$, the model becomes over-identified, and the condition for global convergence requires $G_n(\theta_1)' W_n G_n(\theta_2)$ to be non-singular for all pairs $(\theta_1, \theta_2) \in \Theta \times \Theta$. For just-identified models, this amounts to $G_n(\theta)$ non-singular for all $\theta \in \Theta$. Figure B.5.2 in Appendix B.5 illustrates, similar to Figure 2.2, that Q_n is non-convex and that the rank condition holds for $\Theta = [-1+\varepsilon, 1-\varepsilon]$. F_n is no longer defined because of over-identification. Table 2.2 shows that NR, BFGS and L-BFGS-B all fail to converge from $\theta_0 = 0.95$, a starting value with negative curvature.[15] Compare with GN, which steadily converges to $\hat{\theta}_n$. Starting closer to the solution, BFGS and L-BFGS-B also fail to converge using $\theta_0 = 0.6$; GN remains accurate (not reported). R codes using $p = 12$, $W_n = I_d$ can be found in Appendix B.4.

Table 2.2: MA(1): search paths for NR, GN, BFGS, and L-BFGS-B

k	0	1	2	3	4	5	6	7	…	99	$Q_n(\theta_{99})$
						$p = 1$					
NR	-0.600	-0.689	-0.722	-0.749	-0.772	-0.793	-0.811	-0.828	…	-0.993	0.038
GN	-0.600	-0.560	-0.529	-0.504	-0.484	-0.466	-0.451	-0.438	…	-0.338	$7 \cdot 10^{-8}$
GN-BACK	-0.600	-0.202	-0.326	-0.338	-0.338	-0.338	-0.338	-0.338	…	-0.338	$7 \cdot 10^{-8}$
BFGS	-0.600	-0.505	4.425	-0.307	-0.359	-0.338	-0.337	-0.337	…	-0.337	$7 \cdot 10^{-8}$
L-BFGS-B	-0.600	-0.505	1.000	-0.455	-0.375	-0.318	-0.341	-0.339	…	-0.338	$7 \cdot 10^{-8}$
BFGS*	-0.600	-0.462	-0.286	-0.345	-0.340	-0.338	-0.338	-0.338	…	-0.338	$7 \cdot 10^{-8}$
L-BFGS-B*	-0.600	-0.462	-0.286	-0.345	-0.339	-0.338	-0.338	-0.338	…	-0.338	$7 \cdot 10^{-8}$
						$p = 12$					
NR	0.950	0.956	0.961	0.965	0.969	0.972	0.975	0.978	…	1.000	4.786
GN	0.950	0.890	0.860	0.834	0.810	0.787	0.763	0.740	…	-0.623	0.101
GN-BACK	0.950	0.350	-0.089	-0.478	-0.591	-0.616	-0.616	-0.623	…	-0.626	0.101
BFGS	0.950	-8.290	-8.279	-8.267	-8.256	-8.244	-8.233	-8.221	…	-6.979	0.397
L-BFGS-B	0.950	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	…	-1.000	1.7

Legend: simulated data with sample size $n = 200$, $\theta^\dagger = -1/2$. For $p = 1$, $\overline{g}_n(\theta) = \hat{\beta}_1 - \theta/(1+\theta^2)$. For $p = 12$, $\overline{g}_n(\theta) = \hat{\beta}_n - \beta(\theta)$ where $\beta(\theta)$ is the p-limit of the AR(p) coefficients, evaluated at θ. $W_n = I_d$. The solutions are $\hat{\theta}_n = -0.339$ ($p = 1$) and $\hat{\theta}_n = -0.626$ ($p = 12$). NR = Newton-Raphson, GN = Gauss-Newton, GN-BACK = Gauss-Newton with backtracking line search (Algorithm 2). The learning rate is $\gamma = 0.1$ for NR and GN. BFGS = R's *optim*, L-BFGS-B = R's *optim* with bound constraints $\theta \in [-1, 1]$. BFGS* and L-BFGS-B* apply the same optimizers to F_n instead of Q_n. Additional results for GN using a range of values $\gamma \in \{0.1, 0.2, 0.4, 0.6, 0.8, 1\}$ can be found in Appendix B.5.1, Figures B.5.1, B.5.4.

2.4.2 Estimation of a Random Coefficient Demand Model Revisited

The following revisits the results for random coefficient demand estimation in Knittel & Metaxoglou (2014) with the cereal data from Nevo (2001).[16] This is

---

[15] L-BFGS-B relies on projection descent which maps search directions outside the unit circle to -1 or 1 where $\partial_\theta Q_n(-1) = \partial_\theta Q_n(1) = 0$, a stationary point for (2.1).

[16] It available in the R package BLPestimatoR (Brunner et al., 2017). The data consists of 2,256 observations for 24 products (brands) in 47 cities over two quarters in 94 markets. The specification is identical to Nevo's, with cereal brand dummies, price, sugar content (sugar), a mushy dummy indicating whether the cereal gets soggy in milk (mushy), and 20 IV variables.

a non-linear instrumental variable regression with sample moment conditions: $\overline{g}_n(\theta,\beta) = \frac{1}{n}\sum_{j,t} z_{jt}[\delta_{jt}(\theta) - x'_{tj}\beta]$, where z_{jt} are the instruments, x_{jt} the linear regressors in market j at time period t. The 8 parameters of interest are the random coefficients θ,[17] which enter δ_{jt}, recovered from market shares s_{jt} using the fixed point algorithm of Berry et al. (1995). The 25 linear coefficients β are nuisance parameters concentrated out by two-stage least squares for each θ. The replication sets the maximum number of iterations for the contraction mapping to 20000 and the tolerance level for convergence to 10^{-12}. This is important for the optimization to be well-behaved; see e.g. Brunner et al. (2017), Conlon & Gortmaker (2020). The range of starting values used here is much wider than in these papers,[18] which explains why optimizers are more prone to crashing here than in their replications.

Table 2.3 and Figure 2.3 compare the performance of quasi-Newton (BFGS), Nelder-Mead (NM), Simulated-Annealing (SA), and Nelder-Mead after Simulated-Annealing (SA+NM), using R's default optimizer *optim*, with Gauss-Newton (GN) for 50 different starting values.[19] As reported in Knittel & Metaxoglou (2014), optimization can crash often.[20] Crashes could be avoided using error handling (try-catch statements). However, this may not be enough to produce accurate estimates as the next application will illustrate.[21] Only GN and BFGS systematically produce accurate estimates, but BFGS crashes 60% of the time. Derivative-free optimizers,

---

[17] 8 parameters are the unobserved standard deviation and the income coefficient on the constant term, price, sugar, and mushy.

[18] Conlon & Gortmaker (2020, p25) draw "starting values from a uniform distribution with support 50% above and below the true parameter value."

[19] The solution of the contraction mapping is not well defined for all values in Θ, so we use the first 50 values produced by the Sobol sequence such that δ_{jt} is finite for all j, t.

[20] The optimizers will crash when the fixed point algorithms fail to return finite values. This is typically the case when the search direction was poorly chosen at the previous iteration.

[21] Conlon & Gortmaker (2020) illustrate that modifications to the fixed-point algorithm and specific optimizer implementations to handle near-singularity of the Hessian can also improve performance for BFGS.

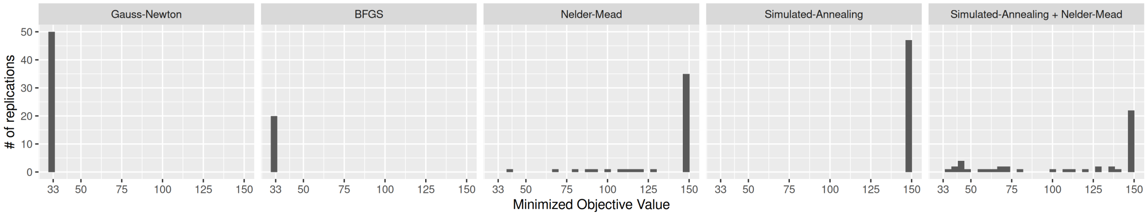
Table 2.3: Demand for Cereal: performance comparison

		STDEV				INCOME				objs	crash
		const.	price	sugar	mushy	const.	price	sugar	mushy		
TRUE	est	0.28	2.03	-0.01	-0.08	3.58	0.47	-0.17	0.69	33.84	-
	se	0.11	0.76	0.01	0.15	0.56	3.06	0.02	0.26	-	
GN	avg	0.28	2.03	-0.01	-0.08	3.58	0.47	-0.17	0.69	33.84	0
	std	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
GN-BACK	avg	0.28	2.03	-0.01	-0.08	3.58	0.47	-0.17	0.69	33.84	0
	std	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	
BFGS	avg	0.28	2.03	-0.01	-0.08	3.58	0.47	-0.17	0.69	33.84	30
	std	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	
NM	avg	0.32	0.35	-0.08	-0.88	3.94	-2.64	-0.10	1.26	628.44	4
	std	1.37	8.91	0.09	3.08	3.63	10.74	0.23	5.13	772.23	
SA	avg	0.87	-0.58	-0.72	-0.00	0.01	0.33	1.64	-1.16	$1.46 \cdot 10^5$	3
	std	7.68	8.66	3.58	7.88	6.67	6.97	3.65	7.92	$2.36 \cdot 10^5$	
SA+NM	avg	0.43	-0.88	-0.06	-0.84	4.15	-2.18	-0.15	0.71	506.44	3
	std	0.61	9.45	0.12	2.25	3.56	11.48	0.19	5.06	1250.65	

Legend: Comparison for 50 starting values in $[-10, 10] \times \cdots \times [-10, 10]$. Avg, Std: sample average and standard deviation of optimizer outputs. TRUE: full sample estimate (est) and standard errors (se). Objs: avg and std of minimized objective value. crash: optimization terminated because the objective function returned an error. GN Gauss-Newton with $\gamma = 0.1$, $k = 150$ iterations for all starting values. GN-BACK Gauss-Newton with backtracking line search, $k = 150$. Additional results for GN using a range of values $\gamma \in \{0.1, 0.2, 0.4, 0.6, 0.8, 1\}$ can be found in Appendix B.5.2, Table B.5.1.

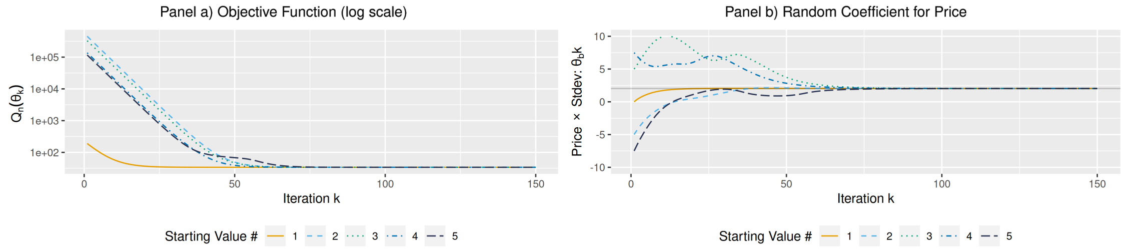
NM, SA, and SA+NM, can produce inaccurate estimates.

Figure 2.3: Demand for Cereal: distribution of minimized objective values



Legend: Comparison for 50 starting values. Minimized objective values for non-crashed optimizations. Objective values are truncated from above at $Q_n(\theta) = 150$.

Figure 2.4: Demand for Cereal: Gauss-Newton iterations for 5 starting values



Legend: 150 GN iterations for 5 starting values in $[-10, 10] \times \cdots \times [-10, 10]$. Panel b) horizontal grey line = full sample estimate.

Figure 2.4, illustrates the convergence of GN for the first 5 starting values. In line with the predictions of Theorem 3, though Q_n is non-convex, GN iterations steadily converge to the solution. This type of "Gauss-Newton regression" is related to Salanié & Wolak (2022) who compute two-stage least-squares for linearized BLP.

2.4.3 Innovation, Productivity, and Monetary Policy

The second application revisits Moran & Queralto (2018)'s estimation of a model with endogenous total factor productivity (TFP) growth (see Moran & Queralto, 2018, Sec2, for details about the model). They estimate parameters related to Research and Development (R&D) by matching the impulse response function (IRF) of an identified R&D shock to R&D and TFP in a small-scale Vector Auto-Regression (VAR) estimated on U.S. data.

The parameters of interest are $\theta = (\eta, \nu, \rho_s, \sigma_s)$ which measure, respectively, the elasticity of technology creation to R&D, R&D spillover to adoption, the persistence coefficient and size of impulse to the R&D wedge. The sample moments are $\overline{g}_n(\theta) = \hat{\psi}_n - \psi(\theta)$, $\hat{\psi}_n$ and $\psi(\theta)$ are the sample and predicted IRFs, respectively. The latter is computed using Dynare in Matlab. To minimize Q_n, the authors use Sims's CSMINWEL[22] algorithm with a reparameterization which bounds the coefficients.[23] Although this type of reparameterization is commonly used, the Jacobian is singular at the boundary; this matters for both local and global convergence, according to the results.

Table 2.4: Impulse Response Matching: performance comparison

		η	ν	ρ_s	σ_s	objs	crash	η	ν	ρ_s	σ_s	objs	crash
TRUE	est	0.30	0.29	0.39	0.17	4.65	-	0.30	0.29	0.39	0.17	4.65	-
		WITHOUT REPARAMETERIZATION						WITH REPARAMETERIZATION					
GN	avg	0.30	0.29	0.39	0.17	4.65	2	0.30	0.29	0.39	0.17	4.65	5
	std	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.00	
GN-BACK	avg	0.30	0.29	0.39	0.17	4.65	1	0.30	0.29	0.39	0.17	4.65	2
	std	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.00	
BFGS	avg	0.12	0.10	-0.34	5.42	$2 \cdot 10^4$	0	0.37	0.21	0.07	0.14	104.08	0
	std	0.56	0.20	0.47	4.77	$2 \cdot 10^4$		0.32	0.14	0.65	0.06	136.63	
CSMINWEL	avg	0.36	-0.00	0.27	0.15	46.42	0	0.62	0.20	0.07	0.14	133.76	0
	std	0.24	1.54	0.33	0.19	183.74		0.39	0.22	0.76	0.08	123.32	
NM	avg	0.47	-5.27	0.43	0.16	14.81	0	0.48	0.26	0.37	0.39	$1 \cdot 10^3$	0
	std	0.54	37.28	0.11	0.05	34.32		0.33	0.16	0.34	1.68	$9 \cdot 10^3$	
SA	avg	1.39	-2.08	0.48	0.09	75.21	0	0.60	0.21	0.44	1.26	$7 \cdot 10^3$	0
	std	2.23	3.59	0.19	0.09	91.35		0.46	0.30	0.74	3.62	$2 \cdot 10^4$	
SA+NM	avg	0.97	-84.27	0.41	0.09	66.53	0	0.61	0.21	0.43	1.08	$5 \cdot 10^3$	2
	std	2.01	124.00	0.22	0.09	79.78		0.45	0.29	0.71	3.33	$2 \cdot 10^4$	
lower bound		0.05	0.01	-0.95	0.01	-	-	0.05	0.01	-0.95	0.01	-	-
upper bound		0.99	0.90	0.95	12	-	-	0.99	0.90	0.95	12	-	-

Legend: Comparison for 50 starting values. TRUE: full sample estimate (est). Objs: avg and std of minimized objective value. crash: optimization terminated because objective returned error. Lower/upper bound used for the reparameterization. GN run with $\gamma = 0.1$ for $k = 150$ iterations for all starting values. Standard errors were not computed in the original study. GN-BACK Gauss-Newton with backtracking line search, $k = 150$. Additional results for GN, using a range of values $\gamma \in \{0.1, 0.2, 0.4, 0.6, 0.8, 1\}$ can be found in Appendix B.5.3, Tables B.5.2, B.5.3

In the original paper, the authors initialize the estimation at $\theta_0 =$

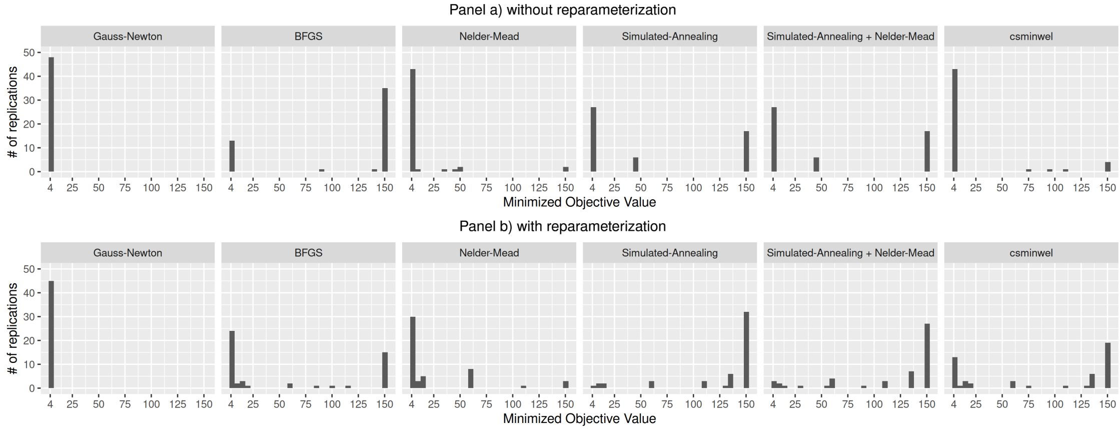
---

[22] Details about CSMINWEL and code can be found at: http://sims.princeton.edu/yftp/optimize/.

[23] The replication uses the mapping $\theta_j = \underline{\theta}_j + \frac{\overline{\theta}_j - \underline{\theta}_j}{1+\exp(-\vartheta_j)}$, where each ϑ_j is unconstrained. The original study relied on $\theta_j = 1/2(\overline{\theta}_j + \underline{\theta}_j) + 1/2(\overline{\theta}_j - \underline{\theta}_j)\frac{\vartheta_j}{\sqrt{1+\vartheta_j^2}}$, which we found to make optimizers very unstable.

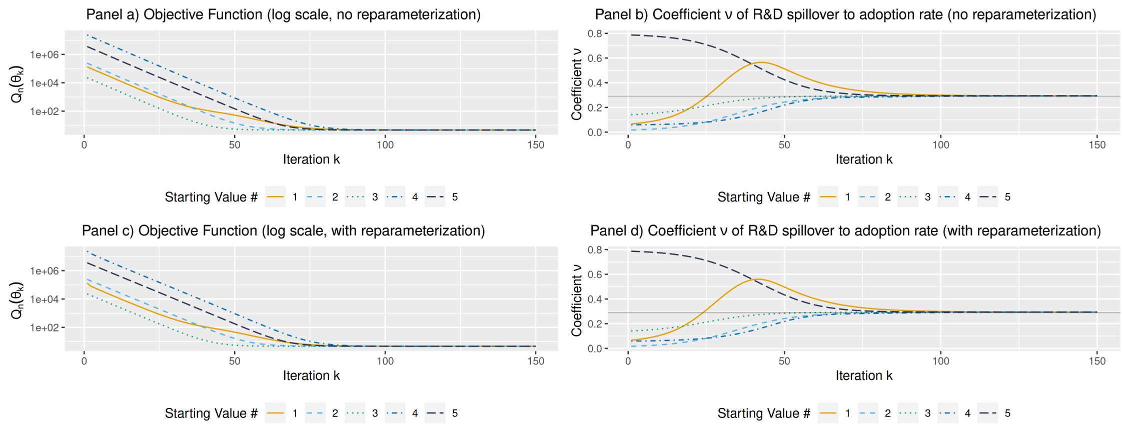
$(\eta_0, \nu_0, \rho_{s0}, \sigma_{s0}) = (0.20, 0.20, 0.30, 0.10)$, very close to $\hat{\theta}_n$. Here, 50 starting values are generated within the bounds in Table 2.4. The model is estimated using CSMINWEL and the same set of optimizers used in the previous replication. Table 2.4 reports the results with and without the non-linear reparameterization. Similar to the MA(1) model with $p = 12$, without the reparameterization, several optimizers return values outside the parameter bounds, which motivates the constraints in these cases. GN correctly estimates the parameters for all starting values but crashes twice for starting values for which both η and ν are close to their lower bounds where the Jacobian is nearly singular. With the reparameterization, GN crashed more often, five times in total, but is otherwise accurate. With backtracking, crashes are fewer. The crashes occur at a value strictly within the parameter bounds for which Dynare cannot solve the model and returns an error. There is no obvious way to modify GN to avoid this problem.

Figure 2.5: Impulse Response Matching: distribution of minimized objective values



Legend: Comparison for 50 starting values. Minimized objective values for non-crashed optimizations. Objective values are truncated from above at $Q_n(\theta) = 150$.

Figure 2.6: Impulse Response Matching: Gauss-Newton iterations for 5 starting values



Legend: 150 GN iterations for 5 non-crashing starting values. Panels a,c) value of the objective function at each iteration, Panels b,d) coefficient η at each iteration; horizontal grey line = full sample estimate.

The other two gradient-based optimizers, BFGS and CSMINWEL, never crash because of better error handling in Matlab. They produce valid estimates less often than GN. Figure 2.5 illustrates that CSMINWEL is sensitive to reparameterization. Likewise, derivative-free methods can be inaccurate, as illustrated in Table 2.4 and Figure 2.5; some crashes occur despite Matlab's error handling. Finally, Figure 2.6 shows 5 optimization paths for which GN does not crash with and without the reparameterization. They are nearly identical. Tables B.5.2, B.5.3 in Appendix B.5.3 gives additional results for larger values of $\gamma \in (0, 1]$, plus results with error handling and the global step from Forneron (2023).

2.5 CONCLUSION

Non-convexity of the GMM objective function is an important challenge for structural estimation, and the survey highlights how practitioners approach this issue. This chapter considers an alternative condition under which there are globally convergent algorithms. The results are robust to non-convexity, one-to-one non-linear reparameterizations, and moderate misspecification. Though off-the-shelf methods might fail to converge due to the non-convexity of the optimization problem, the chapter has shown that this does not necessarily imply that it will be difficult in practice. Econometric theory emphasizes the role of the weighting matrix W_n on the statistical efficiency of the estimator $\hat{\theta}_n$. Also, Hall & Inoue (2003), Hansen & Lee (2021) showed it can alter the pseudo-true value of the parameter under misspecification. Here, the rank condition may or may not hold, depending on W_n. The condition number κ_W also affects local convergence. This highlights another role for the weighting matrix: it may facilitate or hinder the estimation itself. Two empirical applications illustrate the performance of the preferred Gauss-Newton algorithm.

CHAPTER 3

Racial Screening on the Big Screen? Evidence from the Motion Picture Industry

3.1 INTRODUCTION

An employer must decide whether to hire a job applicant. An admission committee must decide whether to admit a candidate to its entering freshman class. A journal editor must decide whether to accept an article for publication. All these settings are characterized by a *decision maker* who must make an in-or-out decision about an *applicant*, having only imperfect information about the applicant's quality. The decision maker may use information about the applicant's race or gender to guide their decision, which may result in discrimination, i.e., the unequal treatment of applicants with otherwise identical characteristics. The econometrician, however, can typically observe only the ex-post outcomes of these decisions: the worker's productivity, the student's grades, or the number of citations received by an article. If we observe differences by race or gender in outcomes, what can we infer about the extent and nature of discrimination by the decision maker?

In this chapter, we address this question in the context of the U.S. motion picture industry, where the producer is the decision maker. There are two main advantages to studying discrimination in the motion picture industry. First, this setting is of intrinsic interest because of the widespread perception of bias in the industry. For example, in the 2010s, only 7% of the nominees for the Academy Awards were African Americans, which is approximately half of their proportion in the population.[1] Does this underrepresentation reflect racial bias? Second, we

---

[1] https://www.washingtonpost.com/news/arts-and-entertainment/wp/2016/02/26/these-charts-explain-how-oscars-diversity-is-way-more-complicated-than-you-think/, accessed on October 26, 2021.

can accurately measure productivity using box office revenue. This is an essential requirement to understand the nature of discrimination. The existence of discrimination in this industry can also have wider ranging implications, because actors can also serve as role models and impact students' educational attainment (Riley, 2024). Therefore, racial discrimination in movie production may differentially affect young viewers of different backgrounds. Understanding whether and to what extent discrimination can be reduced (e.g., via information; Chan, 2024) may guide the design of corrective policies.[2]

We develop a model of discrimination that allows us to interpret differences in box-office revenue, conditional on production. In the model, a producer[3] receives an offer to produce a movie (a "script," similar to the applicant in the examples above). They observe the expected racial composition of the cast based on the script and receive a noisy signal of the movie's expected box-office revenue. Based on the information, they must choose whether to produce the movie and release it to the public or not. We define a "white" movie as a movie in which the leading roles are solely played by whites and a "non-white" movie as a movie in which the leading roles include non-whites. Our model nests different forms of discrimination within it and delivers a rich set of predictions regarding the extent and nature of discrimination. We distinguish between three types of discrimination: a) *customer discrimination*, whereby moviegoers have a preference for white movies over non-white movies; b) *employer* or *taste-based discrimination*, where the pro-

---

[2] Recently, the Academy of Motion Picture Arts and Sciences has announced a multitude of diversity-oriented changes, including diversity requirements for movies that wish to be nominated for the Academy Award in the Best Picture category. In this chapter, we analyze a time period that precedes the inclusion of such standards.

[3] Throughout the chapter we refer for simplicity to the agent deciding on whether to produce the movie as the "producer." This could be a studio executive or other decision maker, and does not necessarily have to coincide with the producer listed in the movie's credits.

ducer suffers a negative utility from producing a non-white movie (Becker, 1957); and c) *statistical discrimination*, where the signal conveyed by non-white movies is less informative about the movie's true quality (Phelps, 1972; Arrow, 1973). We show that the moments of the distribution of box-office revenue of movies *that are produced* allow one to distinguish between the three types of discrimination.

To test the model's predictions, we construct a novel data set with racial identifiers for the cast of more than 7,000 motion pictures released in the United States between 1997 and 2017. We obtained the data by scraping the popular website IMDB,[4] and combined it with extensive information from OpusData, a private company specialized in providing data and information on the movie industry.[5] The racial identifiers are constructed by combining human raters' classifications and a machine learning architecture that integrates a convolutional neural network (CNN) and support vector machine (SVM; Anwar & Islam, 2017).[6]

In our main analysis, we define a movie as "non-white" if two of the four top-billed performers are classified as non-white.[7] We document the following findings. First, the average box-office revenue of non-white movies is substantially *higher* than that of white movies. The raw non-white/white revenue gap is about 91 log points (150%). The inclusion of a standard set of control variables for other movie characteristics and the cast reduces the gap to between 43 and 34 log points (between 54% and 40%), still large and highly statistically significant. Second, the

---

[4] http://www.imdb.com

[5] www.opusdata.com

[6] We rely on the machine learning algorithm to classify the 8% of actors in our data whose racial classification was an object of disagreement for more than two of our eight (sometimes nine) human raters. We find that the algorithm obtains a classification accuracy of more than 95% in our validation data set, which is considered excellent in the image classification literature. See section 3.4 for further details.

[7] The non-white category includes mostly African-Americans but may also include Asians, Hispanics, and other ethnicities.

box office premium of non-white movies is driven primarily by movies in the bottom half of the distribution. Quantile regressions show that the adjusted gap is around 54 log points (about 72%) at the bottom quantiles of the distribution, but the gap at the upper end of the distribution shrinks to about 28 log points (about 33%). These results are robust to different definitions of non-white movies or different dependent variables (e.g., profit margins or profits). Third, we create a measure of the extent to which a movie's box-office revenue overperforms relative to expectations. Following Moretti (2011), we calculate this as the residual in a regression of opening weekend box-office revenue on the number of opening-weekend theaters. We find that relative to white movies, non-white movies substantially overperform relative to expectations.

These results are not consistent with either customer discrimination or statistical discrimination. Instead, we argue that the results are consistent with taste-based discrimination:[8] Non-white movies are held to a higher standard, i.e., they are produced only if the expected revenue surpasses a threshold that is higher than the one set for white movies. This pattern may result from either pure producer taste or a systematic underestimation of the box-office potential of non-white movies.

This chapter is situated within a broad and interdisciplinary literature documenting and exploring discrimination in a variety of settings. While our goal in the next few paragraphs is to focus on the streams of this work to which we directly contribute, we refer the reader to several excellent surveys in economics, including Fang & Moro (2011), Lang & Lehmann (2012), Bertrand & Duflo (2017),

---

[8] Our identification argument relies on the ordering of the means of the (observed) white and non-white box-office revenue distributions, as well as on the ordering of the variances. Therefore, our results may be interpreted as taste-based discrimination quantitatively dominating any other forms of discrimination that may be at play.

Lang & Spitzer (2020) and Onuchic (2022) for a broader overview.

We see our work contribute to the stream of the literature that aims to understand the nature of unequal treatments by either distinguishing between statistical and taste-based discrimination in the data,[9] or testing for the presence of one of the two in a specific market or context. These goals have been pursued experimentally (List, 2004; Zussman, 2013; Doleac & Stein, 2013; Agan & Starr, 2018; Cui et al., 2020; Bohren et al., 2023; Gallen & Wasserman, 2023; Chan, 2024)[10], as well as by testing theoretical predictions on secondary data (Altonji & Pierret, 2001; Knowles et al., 2001; Charles & Guryan, 2008). We contribute to this literature by focusing on a market where some salient interactions exist between employees and final customers, and customer demand drives profit maximization. We propose a simple theoretical framework that nests not only employer taste-based and statistical discrimination but also customer racial animus, and delivers testable predictions for each source of unequal treatment.

This chapter is also related to the literature comparing outcomes between groups to detect the presence of taste-based discrimination. The overarching problem at the heart of average comparisons (or average-based outcome tests; Becker 1957) is that of *infra-marginality*, i.e., in the racial setting, differences in averages might mask both unequal treatment for candidates that are identical but for their race, as well as racial differences in the distributions of unobserved characteristics. Canay et al. (2023) present an extensive discussion on the conditions required for such tests to be valid. The existing literature has dealt with this problem via either random assignments of candidates to decision makers (Arnold et al., 2022);

---

[9] For a review of the literature on the topic, see Guryan & Charles (2013) and Lippens et al. (2020).

[10] Chan (2024)'s field evidence and framework are particularly broad as they expand the focus beyond taste-based and statistical discrimination to include behavioral mechanisms such as biased beliefs and deniable prejudice.

exploiting the timing of release decisions made by parole boards (Anwar & Fang, 2015); specifying equilibrium models (Knowles et al., 2001); or adding distributional assumptions (Simoiu et al., 2017; Pierson et al., 2018; Pierson, 2020). We contribute to the third stream by proposing a parametric approach that is suitable for describing a relatively broad class of screening problems and only relies on the first and second moments of the observed outcome distribution (in our case, box office revenue) for identification. Although we do rely on distributional assumptions to separately identify the different sources of discrimination, we argue that the identification results extend to a set of alternative parameterizations with which empirical researchers might feel comfortable in a variety of settings.

In using higher order moments of the outcome distribution, our test has a similar flavor to those recently proposed by Bharadwaj et al. (2024) and Benson et al. (2024). Bharadwaj et al.'s test is based on the comparison (in the sense of first-order stochastic dominance) between the entire wage distributions of different groups under the implicit assumption that all workers are employed, and there is no screening of workers based on expected productivity. While our approach nests within Bharadwaj et al.'s insight that studying a non-binary outcome (over a binary outcome, e.g., callback) adds margin to separately identify different sources of discrimination, our model explicitly considers the effect of different forms of discrimination on the (continuous) distribution of outcomes *conditional* on production. Moving away from exploiting the entire outcome distributions, in parallel work developed independently, Benson et al. propose a model of racial bias in hiring that nests taste-based discrimination, screening discrimination, and complementary production. They achieve separate identification through testable implications that rely on the mean and variance of workers' productivity under managers of different pairs of races, which they test within the retail context. While the modeling and identification approaches in the two papers are similar, our work departs from Benson et al. in that we do not require the race of the decision maker to be observable. We argue that this is an important contribution to study discrimination in contexts where decisions are likely made by groups rather than single individuals (e.g., admission committees, parole boards, lending organizations, grant review panels); or the decision maker in charge might be influenced by other layers of the organization or industry actors (e.g., media and artistic production, health care treatment approvals, charging decisions, regulatory or legal compliance decisions); or, as is probably quite common, the identity of the decision maker is not observed.

Through its empirical application, the chapter also adds to the line of research that documents the presence of racial discrimination in the motion picture industry (Weaver, 2011; Fowdur et al., 2012). Closest to our work is the paper by Kuppuswamy & Younkin (2020), who find that movies with multiple African-American actors enjoy a box office premium. They rule out customer racial tastes as a discrimination mechanism through an experimental approach. We confirm their conclusion in a more comprehensive data set and provide an analytical framework that can be used to interpret racial differences in the mean and variance of the observed revenue distributions as a function of different forms of discrimination. Our application is also related to a broad empirical literature on labor market and recruiting discrimination, which among the most recent contributions include Åslund et al. (2014); Dustmann et al. (2016); Hedegaard & Tyran (2018); Kline et al. (2022)[11], as well as customer discrimination more broadly (Neumark et al. 1996;

---

[11] See Benson et al. (2024) for a more comprehensive list.

Bar & Zussman 2017; Combes et al. 2016; Leonard et al. 2010 in traditional labor market and service settings; Kahn & Sherer 1988; Nardinelli & Simon 1990; Stone & Warren 1999; Burdekin & Idson 1991 in sport contexts.)

The rest of the chapter proceeds as follows. Section 3.2 describes the institutional background of the motion picture industry. Section 3.3 presents our theoretical model and discusses its empirical implications. Section 3.4 describes the data and the process used to classify performers by race. Section 3.5 presents the main empirical findings and assesses the robustness of the results to different definitions of race or dependent variables. Section 3.6 presents suggestive evidence of incorrect beliefs on the revenue potential of non-white movies within the industry. Section 3.7 discusses and concludes.

3.2 INSTITUTIONAL BACKGROUND OF FILM PRODUCTION

Filmmaking is a complex industry that involves a multiplicity of skills, targets, and decision makers. Each movie displayed on the screen has been through three articulated macro-phases: script writing, production, and distribution. This chapter studies racial discrimination at the production stage.

A key decision maker in the production phase is the producer.[12] They decide whether a script is worth being turned into a movie and, if so, raise the money (sometimes supported by one or more executive producers.) The producer is then responsible for the financial and logistic aspects of the movie.[13] The producer oversees the hiring of the director, who is the creative soul of the movie, the cast, and

---

[12] See for reference Crimson Engine (2018).

[13] The Producers Guild of America (P.G.A.) has established that the producer's name in the film credits can be followed by the *p.g.a.* certification mark only if the producer has performed a significant portion of the producing duties, which includes being physically present on set for a substantial fraction of the production time (P.G.A., n.d.).

the crew, and decides on the budget allocation.[14]

In our conceptual framework, we assume that the movie script itself determines the racial composition of the leading characters in a movie. Although the producer and the casting team[15] may have some latitude in choosing the supporting characters, we think it is plausible that the race of the main characters can be inferred directly from the script. In fact, casting notices for actors typically specify features such as race and ethnicity (and other aspects of physical appearance) for specific roles.

Our model, presented below, describes the producer's decision about whether to produce the movie after they have seen the movie's script and observed the racial composition of the cast and a signal of the movie's quality.

3.3 A MODEL OF THE SCREENING PROCESS

We present here a theoretical framework that helps us understand how observed box-office revenue can inform us about the extent and nature of discrimination in the industry. We assume that the movie production process has the following timeline.

Step 1: Script arrival

There are two types of movies: white movies, denoted by w, and non-white

---

[14] In describing our model in Section 3.3, we will therefore refer to the decision maker as the "producer." It is likely more accurate, however, to think of the decision as made by several agents along a more complex chain of command, as illustrated in the following quote by popular American filmmaker Ed Zwick (Zwick, 2024): "When the creative executive says ''we're going to make this movie", it means she'll try to get the VP to read it. When the VP says he'll make it, it means he's read positive coverage. When the EVP says it, it means she'll take credit for finding it if the president of production likes it. When the president of production says it, it means he needs to tell the CEO which actor is starring in it. And at last, when the CEO says we're going to make this movie, it means it'll get made if he still has his job in six months."

[15] While the producer can be correctly thought of as the primary decision maker in the production process, casting decisions are typically shared among multiple roles.

movies, denoted by b. A risk-averse producer wishes to maximize *log* revenue, denoted by π. The producer receives a script and perfectly observes its type t. However, box office revenues are not observed. We assume that ex-ante box-office revenues of a movie of type t follow a log-normal distribution[16] with type-specific parameters μ_t and $\sigma^2_{\pi t}$:

$$\pi \mid t \sim N(\mu_t, \sigma^2_{\pi t}), \qquad \forall t \in \{w, b\}$$

.

Step 2: Signal and prior updating

Based on the script, the producer updates her prior about the movie's success. Formally, we can think of the producer observing a signal (y) of the movie's box-office revenue. The signal is normally distributed and is well-calibrated, meaning that in expectation, it is equal to the movie's actual (log) box-office revenue, but it is noisy. Critically, we assume that the precision of the signal may differ by movie racial type. Therefore:

$$y \mid \pi, t \sim N(\pi_t, \sigma^2_{yt}).$$

Given this setup, it is straightforward to calculate the posterior mean of log box-office revenue, conditional on the signal and the movie's type:

$$E(\pi \mid y, t) = \frac{\sigma^2_{\pi t}}{\sigma^2_{\pi t} + \sigma^2_{yt}} y + \frac{\sigma^2_{yt}}{\sigma^2_{\pi t} + \sigma^2_{yt}} \mu_t. \tag{3.1}$$

---

[16] The log-normal assumption is made for analytical convenience. In Appendix C.2, we explore alternative distributional assumptions. Most of our results are not sensitive to the specific distributional assumptions. Later in the chapter, we highlight which results depend on the log-normal distribution.

Step 3: Production decision

Producers produce a movie and release it to the public if the expected log box-office revenue, conditional on the movie's type and signal, exceeds a given threshold. This threshold (the *revenue threshold*) is exogenously given. We can think of it as the reservation revenue from a sequential search model, i.e., the value of the revenue that makes the producer indifferent between producing the movie or waiting for a better script.[17] We denote this revenue threshold π_{0t}, making the critical assumption that the threshold is type-specific. For example, this could result from the producer having a taste for producing movies of a given type.

The movie is produced if

$$E(\pi|y,t) > \pi_{0t}, \tag{3.2}$$

This is equivalent to saying that the movie is produced only if the signal y exceeds a given threshold (the *signal threshold*). Based on equation (3.1) and condition (3.2), it is easy to show that the signal threshold is

$$\bar{y}_t = \pi_{0t} + (\pi_{0t} - \mu_t)\frac{\sigma_{yt}^2}{\sigma_{\pi t}^2}. \tag{3.3}$$

In other words, the signal threshold is type-specific and depends on the revenue threshold, the parameters of the prior distribution, and the precision of the signal.

This threshold, together with the statistical features of the ex-ante distribution of box-office revenue and the distribution of revenue conditional on the signal, determines the ex-post distribution of box-office revenue. The following proposi-

---

[17] We think of the race-specific revenue threshold as capturing the disutility cost associated with producing a movie of a given racial type. In our model, the producer does not explicitly internalize the production cost. See Section 3.5.4 for a more extensive discussion of this assumption.

tion establishes the comparative statics of the signal threshold with respect to the parameters of the model.

Proposition 11. *The following comparative statics results hold:*

(a) $\bar{y}_t$ *decreases in* μ_t*.*

(b) $\bar{y}_t$ *increases in* π_{0t}*.*

(c) If $\pi_{0t} > \mu_t$, $\bar{y}_t$ *increases in* σ^2_{yt}*.*

(d) If $\pi_{0t} < \mu_t$, $\bar{y}_t$ *decreases in* σ^2_{yt}*.*

Proof. See Appendix C.1 □

The first two statements in Proposition 11 are straightforward and intuitive. If the ex-ante expected (log) revenue is higher (a high μ_t), the movie is produced even if the signal is not very good. Similarly, when the revenue threshold (π_{0t}) is high, the signal must be excellent to produce the movie. The third and fourth items in the Proposition are more involved but are familiar from the literature on statistical discrimination (Aigner and Cain, 1977; Lundberg and Startz, 1983; Neumark, 2012). Intuitively, if the signal is less precise (a high value of σ^2_{yt},) and the producer wants to produce only high-revenue movies, she will have to set a high signal threshold to make sure she only picks the right tail of the revenue distribution (item (c) in Proposition 11); on the other hand, if the producer only wants to cull out very low revenue movies and the signal is uninformative, the threshold must be set at a low value to ensure that only the very worst (i.e., lowest-revenue) movies are weeded out (item (d) in the proposition).[18]

---

[18] In Appendix C.2, we explore two departures from the normal-normal model. First, we consider

3.3.1 Predictions for empirical work

Proposition 11 characterizes the properties of the signal threshold that determines whether a movie is produced. In practice, we do not observe the signal threshold, so the results are not useful for empirical analysis. However, we do observe the box office revenue of movies that are actually produced and released to the public. The mean and variance of log box-office revenue, conditional on production, are:[19]

$$E(\pi \mid y > \bar{y}_t) = \mu_t + \sigma \frac{\phi(\frac{\pi_0 - \mu_t}{\sigma})}{1 - \Phi(\frac{\pi_0 - \mu_t}{\sigma})} \tag{3.4}$$

$$Var(\pi \mid y > \bar{y}_t) = \sigma^2 \left(1 + \sigma_{yt}^2 + \lambda(\frac{\pi_0 - \mu_t}{\sigma}) \left(\frac{\pi_0 - \mu_t}{\sigma} - \lambda(\frac{\pi_0 - \mu_t}{\sigma})\right)\right), \tag{3.5}$$

where $\sigma = \frac{\sigma_{\pi t}^2}{\sqrt{\sigma_{\pi t}^2 + \sigma_{yt}^2}}$ and $\lambda(x) = \frac{\phi(x)}{(1-\Phi(x))}$.

We can then formulate our central proposition, which enables us to predict how different types of discrimination affect box-office revenues of white and non-white movies produced.

Proposition 12. *Let* $E_t \equiv E(\pi | y > \bar{y}_t)$ *and* $Var_t \equiv Var(\pi | y > \bar{y}_t)$ *be the mean and variance of log box-office revenue conditional on production, as defined in equations (3.4)*

---

a case where producers care only about the binary outcome "whether a movie is a hit" and decide to produce the script only if the posterior probability exceeds a certain threshold (we dub this the Beta-Binomial model). The comparative statics for the signal threshold in this model match exactly those of the normal-normal model, and so do the testable predictions. Second, we consider the case where the prior distribution of revenue is Pareto rather than log-normal (the Pareto model). The comparative statics for the signal threshold in the Pareto model match exactly those of the normal-normal model for the cases of customer and taste-based discrimination. The predictions are somewhat different, as in the Pareto model a) under taste-based discrimination, both the mean and the variance of log revenue conditional on production are predicted to be higher for non-white movies; and b) under statistical discrimination, both the mean and the variance appear to have a U-shaped relationship with the noise of the signal. Importantly, in the Pareto model, none of the three forms of discrimination can match the observed patterns that the mean log revenue is *higher* for non-white movies, while the variance of log-revenue is *lower* for non-white movies (see Section 3.5.5).

[19] Rosenbaum (1961).

and (3.5). Then, the following comparative statics results hold:

(a) E_t *and* Var_t *increase in* μ_t.

(b) E_t *increases in* π_0, Var_t *decreases in* π_0.

(c) E_t *decreases in* σ^2_{yt}, Var_t *increases in* σ^2_{yt}.

Proof. See Appendix C.1 □

We first focus on the intuition behind the comparative statics of E_t with respect to the parameters. The intuition for the first two results is straightforward: Expected revenue conditional on production is higher, the more to the right lies the prior distribution of revenue (result (a)), and the higher is the revenue threshold (result (b)). The third result implies that expected revenue conditional on production increases with signal precision. This result may seem counter-intuitive, as the signal threshold can either increase or decrease with σ^2_{yt} (Proposition 11, results from (c) and (d)). To gain intuition, it is useful to consider the extreme cases of a perfectly informative ($\sigma^2_{yt} = 0$) vis-à-vis perfectly uninformative signal ($\sigma^2_{yt} \to \infty$). If the signal is perfectly informative, the movie is produced only if the signal (which is exactly equal to box-office revenue) is above the revenue threshold. This implies that expected revenue conditional on production is strictly greater than μ_t because some movies will be below the threshold and are not produced. On the other hand, if the signal is perfectly uninformative, whether a movie exceeds the signal threshold conveys no information about its revenue – the expected revenue conditional on production is, therefore, μ_t.

For the variance results, it is helpful to consider the case of a perfectly informative signal. The distribution of revenue conditional on production is a truncated normal distribution, with the truncation point equal to the revenue threshold π_0.

If the whole distribution is shifted to the right and the threshold remains the same, it is easy to see that the variance also increases (result (a)). If the revenue threshold π_0 increases, the truncation point shifts to the right and the distribution variance decreases (result (b)). As for the third result, it is again useful to consider the two polar cases of a perfectly informative vs. a perfectly uninformative signal: With a perfectly informative signal, the distribution of box-office revenue is a truncated normal distribution, which necessarily has a smaller variance than the untruncated distribution that results from a perfectly uninformative signal.

We can now use Proposition 12 to characterize the mean and variance of observed box-office revenues for white and non-white movies under different types of discrimination.

Case 1: Customer discrimination. Customer discrimination implies that the viewing public has a preference for white movies over non-white ones. In terms of our model, this means that the entire distribution of log box-office revenue for white movies is shifted to the right relative to the distribution for non-white movies, or $\mu_b < \mu_w$.

Then, by result 1, it follows that $E_b < E_w$, and $V_b < V_w$. We can, therefore, state the following prediction:

Prediction 1. *Under customer discrimination, the mean log box-office revenue for non-white movies is lower than for white movies, and the variance of log box-office revenue for non-white movies is lower than for white movies.*

Case 2: Taste-based discrimination. We can think of taste-based discrimination as the producer suffering a utility loss from producing non-white movies. Holding everything else constant, the producer will produce a non-white movie only if the expected log revenue exceeds a higher threshold than the one she sets for white movies to compensate her for the disutility of producing a non-white movie. In this case, $\pi_{0b} > \pi_{0w}$. By result 2, we have that $E_b > E_w$ and $V_b < V_w$. We can, therefore, state Prediction 2:

Prediction 2. *Under taste-based discrimination, the mean log box-office revenue for non-white movies is higher than for white movies. The variance of log box-office revenue for non-white movies is lower than for white movies.*

Case 3: Statistical discrimination. We classify under statistical discrimination the case where the informativeness of the signal for non-white movies is smaller than the one for white movies. We believe this assumption is plausible as historically there have been fewer movies with non-white characters, and the (mostly white) producers may find it more difficult to evaluate how successful a movie with non-white characters will be. In this case, $\sigma^2_{yb} > \sigma^2_{yw}$. By result 3, we have that $E_b < E_w$ and $V_b > V_w$. We can, therefore, state prediction 3:

Prediction 3. *Under statistical discrimination, the mean log box-office revenue for non-white movies is lower than for white movies, and the variance of log box-office revenue for non-white movies is higher than for white movies.*

Table 3.1 summarizes our model predictions. In the remainder of the chapter, we use the above predictions to assess the extent and nature of discrimination in the motion picture industry.[20]

---

[20] Throughout, we have assumed that a movie's script is not race-neutral. In fact, our empirical results are driven by genres in which the assumption of non-race-neutral scripts is more likely to hold (see Table 3.7). If scripts are race-neutral, the producer may decide both whether to produce the movie and the racial composition of the cast. However, if hiring a non-white cast is (at least on average) cheaper than hiring a white cast (Appendix Figure D.2), then a mere comparison of our race-specific revenue thresholds will likely *understate* the extent of taste-based discrimination in the market.

3.4 DATA

3.4.1 Facial classification

A key ingredient of our chapter is creating a data set with racial identifiers for movie casts. Some recent papers have used machine learning tools to classify images based on skin tone (Adukia et al., 2023; Colella, 2021). We note that these methods are only partially adequate for our purposes. First, we are interested in classifying images of all non-white actors, including those of Asian, Native American, and other ethnicities that are hard to classify based on skin tone alone. Second, even the most accurate machine-learning algorithm will yield some error rate and, most importantly, may not be able to fully capture all the shades of human perceptions, which is likely the most important dimension for classification in an entertainment context. Therefore, we relied on a team of ten human raters to assign racial identifiers to more than 7000 performer images downloaded from the popular website IMDB.[21]

Each rater was assigned 8 blocks of about 800 performers[22] and was asked to assess whether they thought the person in the image was *White/Caucasian*, *Black/African-American*, *Hispanic*, *Asian*, *Native American/Pacific Islander*, *South Asian*, or *Other*. The option *Unable to Tell* was also made available to the respondents. Raters were specifically instructed not to consult the internet for any information about the performer and to classify the image based on their perception alone. This procedure resulted in between 8 and 9 human ratings for each of the performers in our data set. We assigned to each image the modal classification as long as no more than two raters disagreed on that image's classification.[23] This

---

[21] www.imdb.com

[22] One rater completed only four blocks.

[23] We grouped together the White/Caucasian and Hispanic categories, as we realized that it was

allowed us to classify about 92% of the performers in our sample as either White (79.8%), Black (9.1%) or Asian (2.6%). For the remaining performers in the sample, we used the machine learning algorithm proposed by Anwar & Islam (2017)[24], described in more detail in Appendix C.3.

3.4.2 Additional variables

Our analysis is based on a sample of more than 7000 motion pictures released in the United States between 1997 and 2017. We obtain this information from Opus Data,[25] a private company that collects information on the industry, and rely on IMDb for the approximately 5% of observations in our sample for which OPUS revenues are unavailable. We gather aggregate financial data (box office revenue, production budget, opening weekend revenue, etc.) and metadata (genre, production method) for all movies in our sample.

The main variables of interest in our data set include the gross domestic box-office revenue,[26] production costs,[27] movie run time, Metacritic score, release date,

---

difficult to accurately distinguish between the two. None of the substantive results in the chapter are meaningfully affected if we do not impose this grouping.

[24] Link: https://arxiv.org/ftp/arxiv/papers/1709/1709.07429.pdf.

[25] www.opusdata.com

[26] Our baseline definition of a movie's revenue includes domestic box-office revenues and excludes international box-office sales as well as DVD and Blu-ray revenues. While the information for revenues other than from domestic theaters is available in OPUS, it is not in IMDB, which is the data source we use for revenues whenever the information in OPUS is missing. Reassuringly, we note that, as we restrict the analysis to non-missing OPUS data, the progressive inclusion of DVD, Blu-ray, and international sales does not qualitatively alter our main results. The results are available upon request. Our definition of box-office revenues also excludes streaming revenues, which are instead not available in our data. In 2021, the digital market (which includes video streaming) accounted for 72% of the industry revenue composition, with online video subscription becoming the second largest subscription revenue market as a result of a 26% surge (Motion Picture Association, 2022). While we cannot directly test whether non-white movies account for a similar share of revenues across the streaming vs. non-streaming sectors, we note that our main result is robust (and even larger in magnitude; see Table 3.7) as we restrict our sample to the years before 2007, when streaming accounted for a negligible share of spending on entertainment (see Appendix Figure D.1). Also, no geographic breakdown of revenues is available in our data.

[27] All monetary values are expressed in 2005 dollars.

MPAA rating, number of theaters in which the movie was released, and number of weeks in which the movie was in theaters. We also collect information on the gender and age of the four top-billed performers. We create a variable called "star power," equal to the cumulative box-office revenue of all movies in which each performer appeared up to the release date of the current movie.

3.4.3 Summary statistics

Summary statistics are shown in Table 3.2. The top panel shows that about 12 percent of the top-billed performers in our sample are non-white. About three-quarters of the movies have zero non-white performers, and about 18 percent have only one non-white performer. Our baseline analysis defines a movie as non-white if at least two of the four top-billed performers are non-white. Based on this definition, about eight percent of the movies in our sample are non-white. We also assess the robustness of the results to different definitions of non-white movies.

As for the other variables, the distribution of box office revenue is heavily skewed to the right. Therefore, we use its logarithm as the main dependent variable in our baseline analysis. We collapse the "niche" genres into broader categories so that all movies fall into one of five broad genres. For some of the variables, we only have incomplete data: For example, production costs are available for only about 56 percent of the sample,[28] while the Metacritic score is available only for 71 percent of the sample. To maximize sample size, in the empirical analysis, we replace missing values with zeros and add a dummy variable indicating

---

[28] Probit and logit regressions suggest that movies with higher revenue have a significantly (in the statistical sense) higher probability of non-missing cost information, while the conditional difference between white and non-white movies is not statistically distinguishable from zero. The main result is robust to restricting the sample to observations with non-missing cost information: see columns 5 and 6 in Table 3.3.

that the variable is missing if the missing value is not central to the analysis.

3.5 RESULTS

3.5.1 Non-parametric analysis

Figure 3.1 presents a box-whisker plot of box-office revenue by the number of non-white performers in the movie (out of the four top-billed actors.) The mean box-office revenue increases markedly with the number of non-white performers, while the dispersion of the distribution decreases. Also, the 25th percentile (and lower adjacent value) visibly increases with the number of non-white performers. On the other hand, the 75th percentile is quite stable across cast racial compositions, and the upper adjacent value reduces slightly. We interpret these patterns as the left tail of the non-white movie distribution being missing, which is consistent with the notion that non-white movies are held to a higher standard for production.

Of course, this analysis does not take into account other observable differences that may exist between white and non-white movies. In the following sections, we assess whether the non-white premium in box-office revenue is robust to the inclusion of a broad set of other movie and cast characteristics.

3.5.2 OLS regressions

The main regression model is the following:

$$\ln y_{it} = \beta_0 + \beta_1 Nonwhite_{it} + \beta_2 X_{it} + \delta_t + \varepsilon_{it}, \qquad (3.6)$$

where y_{it} denotes domestic box-office revenue, in 2005 U.S. dollars, of movie i released in year t; $Nonwhite_{it}$, the key explanatory variable of interest, is a dummy variable indicating whether at least two of the four top-billed performers are non-white; X_{it} is a vector of additional control variables, including both cast (average age, gender composition, the "star power" variable described previously) and movie (production budget, MPAA rating, Metacritic score, run time, genre dummies) characteristics; δ_t is a year-of-release fixed effect, and ε_{it} is the robust standard error clustered by distributor.[29]

The results are presented in Table 3.3. The first column of the table shows the unadjusted difference in mean log revenue between white and non-white movies without any controls. The mean box-office revenue of non-white movies is almost 2.5 times as high as that of white movies ($\exp(0.910) \approx 2.5$). In column 2, we include controls for other characteristics of the cast (average age, gender composition, and star power), and the coefficient remains almost unchanged. In column 3, we add controls for the production budget, a dummy for whether the production budget is missing, and all other movie characteristics, including genre and year-of-release fixed effects. The coefficient on the non-white indicator drops to 0.433, implying that non-white movies earn about 54 percent more than white movies at the box office. In column 4, we further add controls for the distributor-level fixed effects, and the coefficient drops to 0.336 (40% revenue gap) while remaining highly significant. For both column 5 and column 6, we restrict the analysis to movies with non-missing data on production costs. In column 5, we replicate column 2, and the coefficient drops from 0.926 to 0.488, which explains what drives the decrease of the coefficient from column 2 to column 3.[30] Finally, column 6 repli-

---

[29] We collapse all distributors with only one movie in our data set into one single distributor category (Other/Unknown).

[30] Conditional on being available in our data, the Metacritic score does not differ significantly on average across white and non-white movies, and a Kolmogorov-Smirnov test fails to reject that the distributions are the same. The Metacritic score is missing for 30% of white movies and 18% of non-white movies in our sample. The correlation coefficient between log revenues and the Metacritic

cates column 3, and the results in this restricted sample are mostly unchanged – the coefficient on the non-white indicator rises to 0.522, implying that non-white movies earn on average about 69 percent more than white movies.[31]

These initial results on the differences between white and non-white movies are not consistent with either a model of customer discrimination, where audiences prefer white movies to non-white movies nor a model of statistical discrimination, where the signal conveyed by non-white scripts is less informative about future box-office revenue. Both models predict that white movies should have, on average, higher box-office revenue than non-white movies, in contrast to our findings. Instead, the results are consistent with a model of taste-based discrimination, where non-white movies are held to a higher standard, i.e., they are only produced if the revenue exceeds a higher threshold than the one required of white movies. In what follows, we look at how other features of the distribution differ between white and non-white movies.

3.5.3 Quantile regressions

The model described in Section 3.3 derives predictions for not only the mean but also the variance of box-office revenues. In this subsection, we analyze other measures of dispersion, namely the white-nonwhite gap at different percentiles of the revenue distribution. Specifically, we estimate a series of quantile regressions of

---

score is equal to .11 and statistically significant at the 1% level.

[31] We have data on the script languages for approximately 70% of the working sample. Within this sample, approximately 86% percent of the movies in our sample have English as the only language on file, and this is a subset of the 95% that have English among the languages to which they are associated. The main result is robust to restricting the sample to English-language movies, indicating that our findings are not driven by foreign-language movies.

the following type:

$$Q_\tau(\ln y_{it}|Nonwhite, X) = \gamma_{0\tau} + \gamma_{1\tau}Nonwhite_{it} + \gamma_{2\tau}X_{it} + \delta_t,$$

where $Q_\tau(\ln y_{it}|Nonwhite, X)$ denotes the τth conditional quantile of the distribution of log box-office revenue, and $\tau \in \{0.05, 0.10, ...0.95\}$. The main coefficients of interest are the $\gamma_{1\tau}$'s, which measure the gap in conditional quantiles across the white and non-white box-office revenue distributions..

Figure 3.2 plots the quantile regression coefficients against the quantiles. As was already apparent from the box-whisker plots in Figure 3.1, from the 20th quantile onwards there is a clear downward trend in the quantile coefficients: The white-nonwhite gap at the lower quantiles is around 60 log points, while it is only about 20 log points at the upper quantiles. This finding reinforces the interpretation that there is “missing mass” in the left-tail of the non-white revenue distribution, or in other words, that non-white movies at the low end of the distribution of box-office revenue are not produced, while comparable white movies are.

3.5.4 Robustness

We next investigate the robustness of our results to different definitions of movie type and different dependent variables.

Classification of non-white movies. In Table 3.4, we consider additional definitions of “non-white” movies. The first column in the table reproduces the results using our baseline classification of non-white movies as those in which at least two of the four top-billed performers are non-white. The first row in the table shows the OLS. results from Table 3.3, while the remaining rows present the quantile regression coefficients at selected quantiles. All specifications include the full set of control variables.

In column 2, we change the definition of non-white movies to include all movies in which at least *one* of the four top-billed performers is non-white. We view this as a noisier indicator of the movie type, as a non-white actor may be cast in a supporting role in a movie that is mainly about white characters and storylines (a form of *tokenism*). Using this definition, the OLS coefficient is substantially reduced (about 23 log points) but still large and highly statistically significant. The pattern of quantile regression coefficients is also clearly downward sloping, with the gap going from about 26 log points at the 25th to about 19 log points at the 90th percentile. In column 3, we replace the dummy indicator for non-white movies with the share of non-whites among the four top-billed performers. The results are quantitatively and qualitatively similar to those of the baseline specification. Finally, in column 4, we classify a movie as non-white only if the top-billed performer is non-white. According to this definition, the average white-nonwhite premium is slightly smaller than in the baseline (46 log points), and the pattern of the quantile regression coefficients is also downward sloping.[32]

On the whole, Table 3.4 shows that the main conclusions regarding the white-nonwhite premium and the nature of discrimination in the industry are not sensitive to the exact definition of non-white movies.

Choice of the dependent variable. In all the analyses so far, we have looked at the logarithm of box-office revenue as the primary dependent variable of interest. The main reason for this choice is that box-office revenue is readily available for almost all movies, and it has been traditionally used as the primary metric for

---

[32] Our main result is robust to restricting the sample to movies with cast popularity (see section 3.4 for a definition of “star power”) below the median, ruling out that the non-white premium that we find is driven by “superstar” non-white movies exclusively.

assessing the commercial success of a movie. However, producers also consider the expected cost of a movie when making production decisions. While we have addressed this in part by including production costs as an explanatory variable in Table 3.3, one may also want to work with profits directly. In the Opus data set, we observe a movie's production budget for about 56% of all movies so that we can calculate various measures of profit.[33] We report the results of this analysis in Table 3.5. The sample includes only those movies for which we observe the production budget. All specifications include the full set of control variables.

In column 1, we use the logarithm of the gross profit margin as a dependent variable, defined as the ratio of domestic box-office revenue to the production budget. The results are broadly consistent with those in the previous sections: Non-white movies have on average a substantially higher profit margin, and the white-nonwhite gap becomes smaller as we move from the low to the high end of the distribution.

In column 2, we focus on the total profit, calculated simply as the difference between box-office revenue and the production budget. It is still the case that the average non-white movie earns a higher profit than the average white movie (by about $8.7 million), holding other characteristics fixed. However, we no longer observe a clear declining pattern in the white-nonwhite gap as we move from lower to upper quantiles in the profit distribution. In fact, the gap appears to be fairly stable (at least in the statistical sense) at all quantiles of the distribution. This could

---

[33] Our measure of profits should only be viewed as a coarse estimate. First, the production budget does not represent the entirety of a movie's production costs, which typically also include marketing costs. Marketing costs are rarely disclosed. Second, the producer typically does not collect all of the box-office revenue, as theaters also receive a cut depending on bilateral negotiations as well as other factors such as the length of time that the movie has been in theaters. Third, cost sharing – a common practice in the movie industry (Weinstein, 1998) – is likely to reduce the extent to which producers internalize costs in their decision making.

be partly due to the shape of the profit distribution, which tends to be quite right-skewed. We confirm this in column 3, where we use the level of box-office revenue (rather than the logarithm) as the dependent variable. We find a positive premium favoring white movies, but now the pattern of quantile regression coefficients shows that the gap becomes larger as we move from the low to the high end of the distribution. We note that, given the substantial right skewness in the revenue distribution, the predictions regarding the variance of box-office revenues *in levels* conditional on production derived from a model that assumes normal distributions no longer hold necessarily.

3.5.5 The white-nonwhite gap in residual variance

An alternative approach to verify our dispersion predictions is to explore how the residual variance differs across white and non-white movies. Borrowing from the heteroskedasticity literature, we posit that the squared residuals from the OLS regression in equation 3.6 have the form:

$$u_{it}^2 = \exp(Z_{it}'\alpha),$$

where the vector Z_{it} contains a subset of the variables included in the main regression (potentially, all of them); We then estimate regressions of $\ln \hat{u}_{it}^2$ on the racial indicator and additional control variables. The results are reported in Table 3.6.

In column 1, the residual variance is assumed to depend only on the racial indicator. Consistent with the results of the box-whisker plot and quantile regressions, we find that non-white movies have a substantially lower residual variance than white movies. In columns 2 and 3, we progressively add additional controls to the variance regression. The results are essentially unchanged – the residual variance of non-white movies is lower than that of white movies.

In the remaining three columns, we experiment with different definitions of non-white movies. The coefficients on the racial variable in the residual regressions are somewhat smaller in absolute values, but still highly statistically significant.[34]

Overall, our results are consistent with what our theoretical model defines as taste-based discrimination, i.e., non-white movies being held to a higher standard, which results in a higher mean and lower variance of box-office revenue for the produced non-white movies.[35]

3.5.6 Heterogeneity Analysis

In Table 3.7, we explore the heterogeneity of our results along a number of different dimensions. First, we look at whether our results are driven by movies produced and distributed by specific segments of the industry. One concern is that our results may capture differences between movies produced by the major studios (the so-called "Big-Six")[36] vs. those produced by smaller studios. It could be that the smaller box-office revenue of white movies reflects the fact that these are often produced by small independent studios, while non-white movies are passed over by these studios altogether. Columns 1 and 2 of the table, however, show that this is not the case: the non-white revenue premium is present among movies distributed

---

[34] The coefficient on the racial variable remains negative and statistically significant when the outcome variable is the log of the profit margin (column 1 of Table 3.5,) but it is imprecisely estimated for profits and revenues in levels (columns 2 and 3 of Table 3.5.) Results are available upon request.

[35] These observed patterns stand in stark contrast to the concept of mean-variance trade-off in the rational asset pricing literature, which traditionally assumes perfectly informed mean-variance utility-maximizing agents (Cochrane, 2005).

[36] These Big-Six studios are: Warner Bros., Paramount Pictures, Walt Disney, Sony / Columbia Pictures, Universal Studios, and 20th Century Fox. These six studios accounted for almost 90% of the US/Canadian market as of 2007. In 2020, Disney acquired 20th Century Fox, and the group is now commonly referred to as the "Big Five". The Big-Six control is included in our regression analysis as a control.

by both types of studios.

We next look at differences across genres (columns 3-5 of the table). The non-white premium is more pronounced among comedies and dramas, where the script is more likely to convey information about the racial composition of the cast. By contrast, the non-white premium is small and not statistically significant in action/adventure movies.

Columns 6 and 7 examine heterogeneity by time period. We look separately at movies produced before and after 2007, the median year in our sample. If taste-based discrimination declines over time, either because of a change in attitudes or because of a change in the competitive landscape, we would expect the non-white premium to shrink. There is some evidence in support of these hypotheses: The non-white premium is 57 log points in the pre-2007 period, but falls to 26 log points in the the post-2008 period.

Finally, in columns 8 and 9, we look at whether the results differ by the gender composition of the cast. We define "female" movies as those in which (strictly) more than 50% of the leading actors are women. The non-white premium is considerably larger among female movies, suggesting that non-white movies must pass an even higher threshold if the cast is predominantly female.

3.5.7 Producer analysis

In this section, we explore the role of the producer's race in explaining the non-white box-office premium. Neither IMDB nor OpusData contains demographic information on movie producers. We, therefore, rely on a human rater to code the producers' race for a subsample of our films. The first step is matching films to producer names, which are available in the *Credits* section of the Opus data set.

We have information on producers for 3,878 out of the 6,943 movies in our sample: 9,842 distinct names are associated with those movies in the capacity of *Producer* or *Executive Producer.* Of these producers, fewer than 1% can be racially categorized via Wikipedia. For the remainder, we then randomly drew approximately 8% of the remaining producers associated with either white or non-white movies[37] and asked a human rater to racially classify these producers based on photos and text resources available online. We then matched the producer's racial information to our main data set. We end up with a working sample of 1,955 movies with racial information for at least one producer. Of these, 261 (13%) display more than two non-white actors, while 403 (21% – 257 of the white movies and 146 of the non-white movies) are associated with at least one non-white producer.

Table 3.8 shows the results of our producer analysis. Column 1 reports the estimated non-white premium in the sub-sample of interest. The coefficient is positive and statistically significant like the one obtained in the full sample (Table 3.3, column 3) but approximately half in size. Adding the producer's race to the controls (Column 2) does not change the coefficient of interest in any significant way, and the producer control itself is statistically insignificant.[38]

Column 3 reports the results obtained from interacting the racial indicator for the cast with the racial indicator for the producer. Our findings reveal that the non-white revenue premium is driven by movies with at least one non-white producer, while on average, films associated with white producers do not display a non-

---

[37] In our sample, the average number of producers and executive producers (and co-producers) associated with a film is 8 (9), and the median is 7 (8). Therefore, to guarantee a large enough working sample for our producer analysis, we randomize at the producer level and not at the movie level. This implies that in our exercise, we are comparing movies with at least one non-white producer to movies that may or may not have any non-white producers. We stratified our randomization by the movie racial type to end up with a reasonably balanced data set.

[38] The findings are robust to the inclusion of studio fixed effects. Standard errors are clustered by the studio. Results are available upon request.

white revenue premium. Taken at face value, these findings suggest that taste-based discrimination may be more concentrated among non-white producers. This evidence should be interpreted with caution, however, given the limited scope of our analysis. This pattern can be rationalized through the observation, discussed in Section 3.2, that producers may not be the pivotal decision makers in the film production decision and may be held to different standards themselves, depending on their racial group. An alternative interpretation is that non-white producers have a comparative advantage in producing non-white movies, and, in particular, may obtain a more precise signal of revenue when evaluating scripts. In the context of our model, such an informational advantage would indeed translate into higher revenues for non-white movies produced by non-white producers.

3.6 ALTERNATIVE EXPLANATION: IS THE INDUSTRY SURPRISED?

The empirical results so far suggest that non-white movies are held to higher production standards than white movies. A candidate interpretation of these patterns is that producers dislike producing non-white movies and face a disutility cost every time they produce one. As a result, the expected revenue for producing non-white movies needs to be higher than the expected revenue for producing white movies (in the context of the model, $\pi_{0b} > \pi_{0w}$).

An alternative, non-mutually exclusive, interpretation is that the industry systematically underestimates the revenue potential of non-white movies relative to white movies.[39,40] In other words, actual box-office revenue for non-white movies

---

[39] The film industry is known for having a hard time forecasting movies' success, as well as analyzing past results: "Why was ''The Hunger Games" such a big hit? Because it had a built-in audience? Because it starred Jennifer Lawrence? Because it was released around spring break? The business is filled with analysts who claim to have predictive powers, but the fact that a vast majority of films fail to break even proves that nobody knows anything for sure" (Davidson, 2012.)

[40] See Chan (2024); Bohren et al. (2023); Esponda et al. (2022); Bordalo et al. (2016); Fong & Luttmer

is π_b, but producers perceive it to be $\hat{\pi}_b = \pi_b - e_b$, with $e_b > 0$. This explanation would yield similar predictions to the ones derived from taste-based discrimination, even if the nature of discrimination in the industry is quite different.

Our model intrinsically cannot identify taste-based disutility costs and biased beliefs separately. Nevertheless, we can make some progress on this front by exploiting the decision that distributors make on the number of theaters at which the movie is displayed on the opening weekend. We argue that this is a proxy of the market's rational expectation of the movie's potential after production, as distributors' decision-making is less likely to be affected by taste-based or statistical discrimination: While producers "sign" a movie as a creation of theirs and create a permanent bond with the film, studios, and theater owners are more likely to make distribution choices based on purely profit-maximizing considerations once the movie has been produced. Moreover, statistical discrimination should also be of relatively less importance at the distribution stage, because distributors also observe the ex-post quality of the movie rather than just the script.

We conjecture that distributors choose the number of theaters based on expected customer demand. If non-white movies are displayed in fewer theaters than white movies, this indicates that distributors expect relatively smaller revenue from the non-white movies. Therefore, if non-white movies have the same level of customer demand but are displayed in fewer theaters, we conclude that distributors underestimate their revenue potential.

Using data on the number of screens in which a movie is shown, we test the hypothesis that the industry systematically underestimates the revenue potential of non-white movies. Specifically, we first regress first-weekend box office revenues

---

(2011) for evidence of inaccurate beliefs in other contexts.

on the number of theaters in which movies are projected over the first weekend upon their release. Following Moretti (2011), we interpret this as a proxy for the industry expectation of a movie's box-office revenue. The residuals from this regression can then be viewed as a measure of the industry's underestimation or overestimation of a movie's revenue potential. If the non-white mean residual is significantly larger than the white mean residual, this suggests that the industry systematically underestimates non-white movies' revenue potential relative to white movies.

We start by running a simple bivariate regression of log first-weekend revenues on the log number of theaters. The R-squared of this regression is 0.89 (column 1 of Table 3.9), and it remains relatively stable as further controls are added (columns 2 and 3). The number of theaters is, hence, a good predictor of first-weekend revenues.

We then test whether the residuals obtained from the regressions in Table are on average between non-white and white movies. The results are presented in the bottom panel of the table. We find that the mean residual for non-white movies is positive across specifications, while the mean residual for white movies is close to zero. That is, the industry underestimates the first-weekend success of non-white movies relative to white movies. The difference between the white and the non-white residual is always statistically significant. We conclude that our results might be at least partly explained by a systematic underestimation of non-white movies' box-office potential within the industry.

3.7 CONCLUSION

This chapter presents a framework for detecting the extent and nature of discrimination in contexts in which decision-makers screen applicants. The econometrician can only observe the outcomes of applicants who successfully pass the screening process. The framework nests several leading theories of discrimination and derives a rich set of testable empirical predictions.

We apply these tests in the context of racial representation in the U.S. motion picture industry. We show that non-white movies earn a box-office premium. The gap is particularly pronounced at low quantiles of the distribution, suggesting that non-white movies with low box-office potential are never produced; in other words, non-white movies are held to a higher standard in the production decision. In the context of our model, this evidence is consistent with taste-based discrimination, i.e., producers suffering a utility loss from producing non-white movies. The evidence is also consistent with producers and distributors having inaccurate beliefs and systematically underestimating the revenue potential of non-white movies. On the other hand, the evidence is not consistent with simple customer discrimination against non-white movies, nor with a statistical discrimination story in which the signal sent by non-white movies is less precise.

These results may appear puzzling to the extent that they hint at lost profits and relatively slow learning in the industry for non-white movies' potential. While our results indicate that the non-white revenue premium has more than halved between 1997-2007 and 2008-2017 (and become harder to distinguish from zero in a statistical sense, despite the larger sample size,) the point estimate for the latter period is far from zero in an economically significant way. Some of the *a priori* plausible explanations do not seem applicable to our setting: it is unlikely that learning is hindered by the non-white premium being too small to be detected or consequential, or that too few non-white movies are produced. The fact that – conditional on production – non-white movies are relatively more successful rules out customers' attitudes and pre-market discrimination as leading explanations (Becker, 1957). We argue that other industry-specific forces might be at play, including the high concentration of the motion picture industry, with the "Big Six" studios typically accounting for more than 80% of the industry's total market share. In non-competitive industries, firms may have more latitude to indulge their discriminatory taste. The introduction and growth of streaming services and smartphone applications in the 2010s appear to have increased the amount of competition in the industry (Kuehn & Lampe, 2023), which is consistent with the declining non-white premium in the latter part of our sample. There are also documented challenges and uncertainty that industry actors face in predicting movies' revenues and profitability, even when data are available (Lash & Zhao, 2016).[41] We leave the investigation of this puzzle to future research, along with the analysis of the consequences of the recent diversity-promoting rules that the Academy of Motion Picture Arts and Sciences set for those aspiring to best-picture qualifications (Sperling, 2020b,a).

While the specific application in this chapter looked at the motion picture industry, our model can be readily applied to other contexts in which decision makers can use group identifiers to screen applicants, and one can observe the outcome or productivity of successful applicants: the output of workers hired for a particular job, the academic performance of students admitted to a freshman class, or the number of citations accumulated by a published journal article. These other

---

[41] For discussions and examples in the public press, see Yahr (2016); Gladwell (2006); Thompson (2013); Snee (2016).

contexts are promising avenues for future research.

TABLES AND FIGURES

Table 3.1: Model Predictions

Discrimination Source	**Mathematical Definition**	**Comparative Statics: Expected Value**	**Comparative Statics: Variance**
Taste-based The producer bears a utility loss producing non-white movies.	$\pi_{0b} > \pi_{0w}$ The production threshold is relatively higher for non-white movies.	$E_b > E_w$	$Var_b < Var_w$
Customer The viewing public has a preference for white movies over non-white movies.	$\mu_b < \mu_w$ The distribution of box-office revenues for white movies is shifted to the right, relative to that of non-white movies.	$E_b < E_w$	$Var_b < Var_w$
Statistical The producer has “less” or “worse” information on non-white movies’ potential.	$\sigma^2_{yb} > \sigma^2_{yw}$ The signal for non-white movies is less informative.	$E_b < E_w$	$Var_b > Var_w$

Legend: Summary of the model predictions. In our notation, w (b) denotes white (non-white) movies; π_{0b}, π_{0w} are the type-specific production thresholds; μ_b, μ_w denote the type-specific means of the box-office revenue distributions; $\sigma^2_{yb}, \sigma^2_{yw}$ stand for the type-specific signal variances. See Section 3.3 for the derivations.

Table 3.2: Summary Statistics

	(1)	(2)	(3)	(4)	(5)
VARIABLES	N	mean	sd	min	max
PANEL A: Classification of movies by type					
Share of non-white performers	7,840	0.12	0.24	0	1
At least one non-white	7,840	0.26	0.44	0	1
At least two non-whites	7,840	0.08	0.27	0	1
Distribution of the number of non-white performers (percentages):					
0	74.3				
1	17.9				
2	4.5				
3	2.1				
4	1.3				
PANEL B: Other variables					
Gross revenue(in Millions of 2005 Dollars)	7,205	25.9	54.1	$1.94 * 10^{-5}$	804
Ln (Gross revenue)	7,205	14.14	3.39	2.97	20.50
Cost(in Millions of 2005 Dollars)	3,955	37.4	44.7	$1.1 * 10^{-3}$	907
Ln(Cost)	3,955	16.72	1.45	7.00	20.63
Run time(minutes)	6,804	103.51	18.49	38	600
IMDB score	4,915	6.25	0.97	1.50	9
Metacritic score	4,915	51.58	17.10	1	100
Average age of billed performers	7,715	41.92	10.42	10	99
Star power(in Millions)	7,840	262	303	0	2,350
Ln(Star power)	7,840	17.50	4.54	0	21.58
Number of weeks	6,491	11.62	14.66	1	476
Ln(Number of screens)	6,078	4.88	2.95	0.69	8.43
Distribution of movies by genre (percentages):					
Action	16.62				
Animation	0.17				
Comedy	26.27				
Drama	36.57				
Other	20.37				

Legend: Source: authors' calculations. Data sources are described in Section 3.4.

Table 3.3: The non-white revenue premium

	(1)	(2)	(3)	(4)	(5)	(6)
Sample:	Full	Full	Full	Full	Non-missing cost	Non-missing cost
	Ln(Gross Rev)	Ln(Gross Rev)	Ln(Gross Rev)	Ln(Gross Rev)	Ln(Gross Rev)	Ln(Gross Rev)
Race: at least	0.914^{***}	0.926^{***}	0.433^{***}	0.336^{***}	0.488^{***}	0.522^{***}
two non-white	(0.200)	(0.183)	(0.093)	(0.076)	(0.157)	(0.096)
Share of		-1.059^{***}	0.180^{**}	0.096	-0.732^{***}	0.130
female		(0.194)	(0.084)	(0.074)	(0.203)	(0.111)
ln(Star Power)		0.235^{***}	0.023^{**}	0.006	0.176^{***}	-0.033^{***}
		(0.022)	(0.010)	(0.008)	(0.020)	(0.011)
Average age		-0.065^{***}	-0.004	0.003	-0.036^{***}	-0.012^{***}
		(0.006)	(0.004)	(0.003)	(0.007)	(0.004)
ln(Cost)			0.554^{***}	0.403^{***}		0.728^{***}
			(0.051)	(0.034)		(0.044)
= 1 if ln(Cost)			6.249^{***}	4.804^{***}		
			(0.736)	(0.541)		
Movie controls			Y	Y		Y
Distributor FEs				Y		
N	6943	6943	6943	6943	3856	3856
R^2	0.006	0.125	0.698	0.341	0.071	0.596

Legend: Data sources and specification are described in Sections 3.4 and 3.5. Cast control variables include the share of females, the average age of the four top-billed performers, and "star power" (defined as the log of performers' cumulative box office revenues up to the movie release date). Movie control variables include indicators for movie genre, indicator of whether the movie is from the “Big 6”, run time, Metacritic score, MPAA rating, year fixed effects, and indicators for missing run time, Metacritic score, or MPAA rating. The movie budget cost (in the log) is included among the control variables when revenue is the dependent variable. Standard errors clustered by distributor in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.4: Robustness: different definitions of "non-white" movies

	(1)	(2)	(3)	(4)
Race:	At least two non-white	At least one non-white	Share of non-white	Leading role is non-white
	Ln(Gross Revenue)	Ln(Gross Revenue)	Ln(Gross revenue)	Ln(Gross revenue)
Race	0.433^{***}	0.227^{***}	0.628^{***}	0.464^{***}
	(0.093)	(0.061)	(0.121)	(0.081)
Q10	0.542^{***}	0.237^{**}	0.664^{***}	0.502^{***}
	(0.104)	(0.113)	(0.229)	(0.100)
Q25	0.517^{***}	0.263^{***}	0.813^{***}	0.520^{***}
	(0.109)	(0.059)	(0.171)	(0.122)
Q50	0.375^{***}	0.215^{***}	0.572^{***}	0.334^{***}
	(0.078)	(0.060)	(0.120)	(0.089)
Q75	0.281^{**}	0.189^{***}	0.469^{***}	0.327^{***}
	(0.112)	(0.049)	(0.122)	(0.099)
Q90	0.283^{***}	0.188^{***}	0.442^{***}	0.302^{***}
	(0.058)	(0.061)	(0.120)	(0.062)
Cast controls	Y	Y	Y	Y
Movie controls	Y	Y	Y	Y
N	6943	6943	6943	6943

Legend: Data sources and specification are described in Sections 3.4 and 3.5. Cast control variables include the share of females, the average age of the four top-billed performers, and "star power" (defined as the log of performers' cumulative box office revenues up to the movie release date). Movie control variables include indicators for movie genre, indicator of whether the movie is from the "Big 6", run time, Metacritic score, MPAA rating, year fixed effects, and indicators for missing run time, Metacritic score, or MPAA rating. The movie budget cost (in the log) is included among the control variables when revenue is the dependent variable. Standard errors clustered by distributor in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.5: Robustness to different dependent variables

	(1)	(2)	(3)
Sample:	Non-missing cost variable	Non-missing cost variable	Non-missing cost variable
	Ln(Profit Margin+1)	Profit(in million)	Revenue(in million)
Race: At least two non-white	0.553***	8.722***	1.746
	(0.095)	(2.156)	(3.004)
Q10	0.469***	5.849***	3.233***
	(0.114)	(1.569)	(0.919)
Q25	0.339***	6.934***	5.372***
	(0.095)	(1.250)	(1.192)
Q50	0.460***	8.352***	5.323***
	(0.104)	(1.287)	(1.654)
Q75	0.352***	9.494***	6.907**
	(0.085)	(3.679)	(3.368)
Q90	0.314***	11.883***	6.278
	(0.072)	(3.707)	(6.384)
Cast controls	Y	Y	Y
Movie controls	Y	Y	Y
N	3856	3856	3856

Legend: Data sources and specification are described in Sections 3.4 and 3.5. Cast control variables include the share of females, the average age of the four top-billed performers, and "star power" (defined as the log of performers' cumulative box office revenues up to the movie release date). Movie control variables include indicators for movie genre, indicator of whether the movie is from the "Big 6", run time, Metacritic score, MPAA rating, year fixed effects, and indicators for missing run time, Metacritic score, or MPAA rating. The movie budget cost (in logs) is included among the control variables when revenue is the dependent variable. Standard errors clustered by distributor in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.6: Conditional residual variance regressions: robustness with respect to different definitions of non-white movies

	(1)	(2)	(3)	(4)	(5)	(6)
Race definition	At least two	At least two	At least two	At least one	Share	Leading role
	Dependent variable: Ln(residual square)					
Race	-0.463***	-0.471***	-0.324***	-0.117**	-0.379***	-0.212**
	(0.099)	(0.098)	(0.094)	(0.059)	(0.109)	(0.083)
Cast Controls		Y	Y	Y	Y	Y
Movie Controls			Y	Y	Y	Y
N	6943	6943	6943	6943	6943	6943
R^2	0.003	0.012	0.121	0.123	0.122	0.117

Legend: Data sources and specification are described in Sections 3.4 and 3.5.5. Cast control variables include the share of females, the average age of the four top-billed performers, and "star power" (defined as the log of performers' cumulative box office revenues up to the movie release date). Movie control variables include indicators for movie genre, indicator of whether the movie is from the "Big 6", run time, Metacritic score, MPAA rating, year fixed effects, and indicators for missing run time, Metacritic score, or MPAA rating. The movie budget cost (in the log) is included among the control variables when revenue is the dependent variable. Standard errors are clustered by distributor in the main regressions only. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.7: Heterogeneity Analysis

	(1) Distributor Not Big-6	(2) Distributor Big-6	(3) Genre: Action/Adventure	(4) Genre: Comedy	(5) Genre: Drama
Race: At least two non-white	0.508***	0.355***	0.068	0.823***	0.395***
	(0.138)	(0.089)	(0.201)	(0.178)	(0.128)
Cast Controls	Y	Y	Y	Y	Y
Movie Controls	Y	Y	Y	Y	Y
P-Value of the difference	0.397		0.014		
N	4766	2177	1135	1880	2590
	(6) Period: Pre-2007	(7) Period: Post-2008	(8) Gender: ≤50% female	(9) Gender: >50% female	
Race: At least two non-white	0.567***	0.262*	0.408***	0.674**	
	(0.092)	(0.133)	(0.087)	(0.307)	
Cast Controls	Y	Y	Y	Y	
Movie Controls	Y	Y	Y	Y	
P-Value of the difference	0.041		0.365		
N	2774	4169	5933	1010	

Legend: Data sources and specification are described in Sections 3.4 and 3.5.5. In all specifications, the sample is restricted to observations with non-missing data on production costs. Cast control variables include the share of females, the average age of the four top-billed performers, and "star power" (defined as the log of performers' cumulative box office revenues up to the movie release date). The movie budget cost (in the log) is included among the control variables when revenue is the dependent variable. Standard errors clustered by distributor in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.8: Producer Analysis

	(1) Producer Sub-Sample	(2) Producer Sub-Sample	(3) Producer Sub-Sample
	Dependent variable: Ln(gross revenue)		
Cast: more than two non-white	0.271**	0.253*	-0.004
	(0.132)	(0.132)	(0.138)
Producer: more than one non-white		0.045	-0.081
		(0.121)	(0.126)
Cast x Producer			0.556**
			(0.228)
Observations	1,955	1,955	1,955
R-squared	0.733	0.733	0.734
Baseline controls	Y	Y	Y

Legend: Data sources and specification are described in Sections 3.4 and 3.5.7. In all specifications, the sample is restricted to observations with some information on the producer race. Cast control variables include the share of females, the average age of the four top-billed performers, and "star power" (defined as the log of performers' cumulative box office revenues up to the movie release date). Movie control variables include indicators for movie genre, indicator of whether the movie is from the "Big 6", run time, Metacritic score, MPAA rating, movie budget cost (in the log), year fixed effects, and indicators for missing run time, Metacritic score, MPAA rating, or budget cost. Standard errors clustered by distributor in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.9: Regressions of first-weekend theaters on number of theaters

VARIABLES	(1) First-weekend revenues, log	(2) First-weekend revenues, log	(3) First-weekend revenues, log
First-weekend # theaters, log	0.990***	0.980***	0.839***
	(0.016)	(0.017)	(0.018)
Cast controls	N	Y	Y
Movie controls	N	N	Y
Residuals: white vs non-white			
Average white	-0.014	-0.014	-0.016
Average non-white	0.151	0.152	0.178
Average difference	-0.165	-0.166	-0.194
p-value of t-test (two-sided)	0.001	0.001	0.000
N	6,276	6,276	6,276
R^2	0.889	0.890	0.927

Legend: Data sources and specification are described in Sections 3.4 and 3.6. Cast control variables include the share of females, the average age of the four top-billed performers, and "star power" (defined as the log of performers' cumulative box office revenues up to the movie release date). Movie control variables include indicators for movie genre, indicator of whether the movie is from the "Big 6", run time, Metacritic score, MPAA rating, year fixed effects, movie budget cost (in the log), and indicators for missing run time, Metacritic score, MPAA rating, and movie budget cost. Relative to the baseline sample used in Table 3.3, 633 observations are excluded due to missing data on first-weekend revenues or theaters, while 55 are lost due do taking logs (zero values.) Standard errors clustered by distributor in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

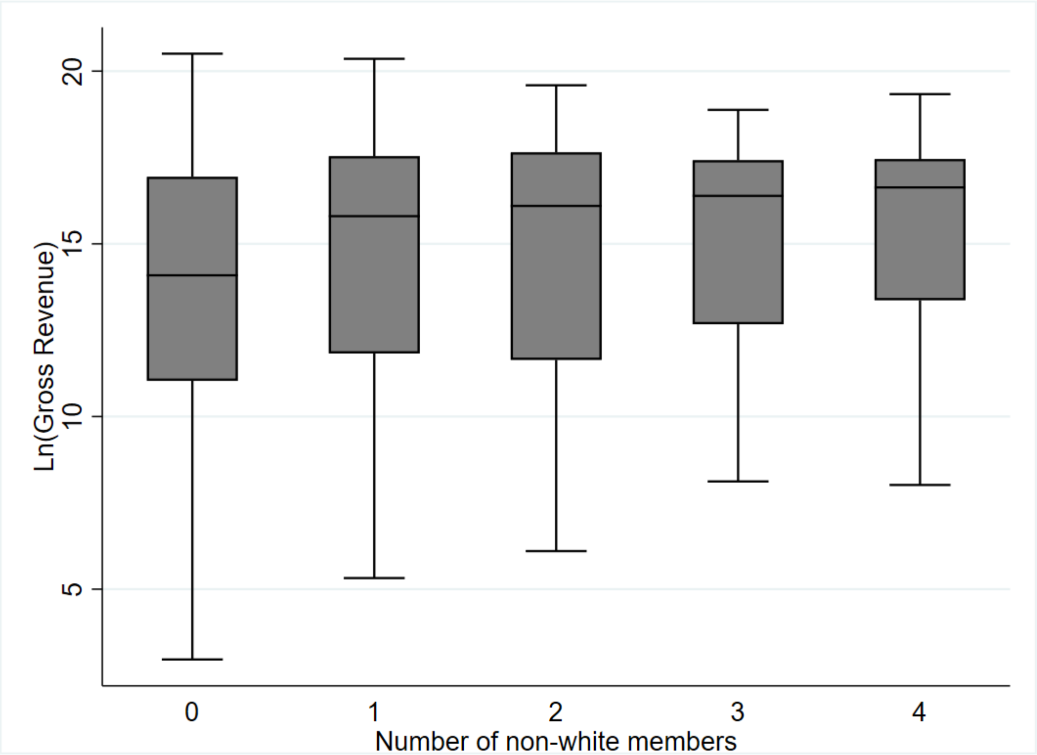


Figure 3.1: Revenue distribution by number of non-white members

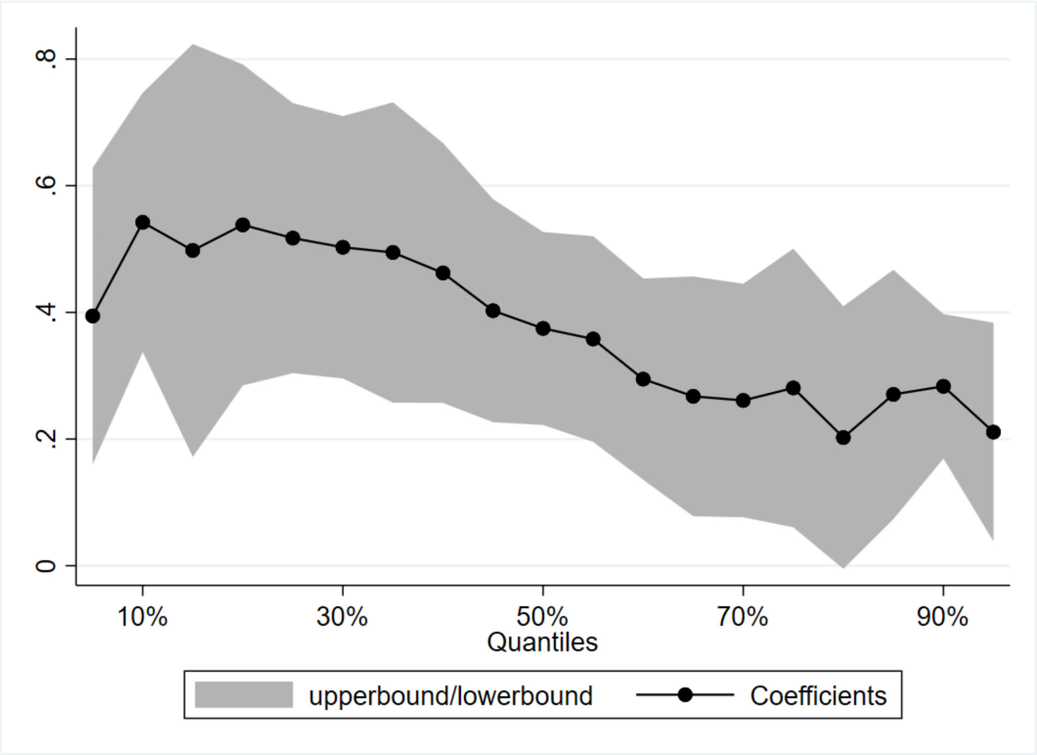


Figure 3.2: Coefficients are decreasing over quantiles

APPENDIX A

Supplementary Materials for Chapter One

A.1 PROOF OF THE THEOREMS

Proof of Proposition 1. For any $d, d' \in \{0,1\}^N$. Consider any $i \in \mathbb{I}(d) \cap \mathbb{I}(d')$. By Definition 4 of imputable units, under $H_0^{\epsilon_s}$, we have $Y_i(d) = Y_i(d')$. Hence, by Definition 6 of pairwise imputable statistics, $T(Y_{\mathbb{I}(d)}(d), d') = T(Y_{\mathbb{I}(d)}(d'), d')$. □

Proof of Theorem 1. Given any $\alpha > 0$, consider the subset of assignment

$$\mathbb{D} \equiv \{D^{obs} | pval^{pair}(D^{obs}) \leq \alpha/2\}.$$

Therefore, we can denote $P(pval^{pair}(D^{obs}) \leq \alpha/2) = \sum_{D^{obs} \in \mathbb{D}} P(D^{obs}) = w$. Since $E_P(\phi(D^{obs})) = P(pval^{pair}(D^{obs}) \leq \alpha/2)$, to prove the theorem, we want to show $w < \alpha$.

Denote $H(D^{obs}, D) = 1\{T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) \geq T(Y_{\mathbb{I}(D)}(D), D^{obs})\}$. Then, by construction, $H(D^{obs}, D) + H(D, D^{obs}) \geq 1$.

Under $H_0^{\epsilon_s}$, by Proposition 1 and Definition 9 of *p*-value,

$$pval^{pair}(D^{obs}) = \sum_{D \in \{0,1\}^N} H(D^{obs}, D) P(D).$$

Now, consider the term

$$\sum_{D^{obs} \in \mathbb{D}} \sum_{D \in \{0,1\}^N} H(D^{obs}, D) P(D) P(D^{obs}).$$

On the one hand, it equals

$$\sum_{D^{obs}\in\mathbb{D}} pval^{pair}(D^{obs})P(D^{obs}) \leq (\alpha/2)(\sum_{D^{obs}\in\mathbb{D}} P(D^{obs})) = w\alpha/2.$$

On the other hand, by flipping D and D^{obs} in the same set $\mathbb{D}$,

$$\begin{aligned}\sum_{D^{obs}\in\mathbb{D}}\sum_{D\in\mathbb{D}} H(D^{obs},D)P(D)P(D^{obs}) &= \sum_{D\in\mathbb{D}}\sum_{D^{obs}\in\mathbb{D}} H(D,D^{obs})P(D^{obs})P(D)\\ &= \sum_{D\in\mathbb{D}}\sum_{D^{obs}\in\mathbb{D}} H(D,D^{obs})P(D)P(D^{obs})\\ &= \sum_{D^{obs}\in\mathbb{D}}\sum_{D\in\mathbb{D}} H(D,D^{obs})P(D)P(D^{obs}).\end{aligned}$$

Hence, we would have

$$\begin{aligned}&\sum_{D^{obs}\in\mathbb{D}}\sum_{D\in\{0,1\}^N} H(D^{obs},D)P(D)P(D^{obs}) \geq \sum_{D^{obs}\in\mathbb{D}}\sum_{D\in\mathbb{D}} H(D^{obs},D)P(D)P(D^{obs})\\ &\quad= \sum_{D^{obs}\in\mathbb{D}}\sum_{D\in\mathbb{D}}(H(D,D^{obs})+H(D^{obs},D))P(D)P(D^{obs})/2\\ &\quad(\text{By } H(D^{obs},D^{obs})+H(D^{obs},D^{obs})=2)\\ &\quad> \sum_{D^{obs}\in\mathbb{D}}\sum_{D\in\mathbb{D}} P(D)P(D^{obs})/2 = w^2/2.\end{aligned}$$

Hence, $w^2/2 < w\alpha/2$, implying $w < \alpha$. As previously mentioned, using 1/2 to discount the number of equalities does not affect the test's validity because $H(D^{obs},D)+H(D,D^{obs}) \geq 1$ would still hold.

□

Too Many Potential Treatment Assignments. When the number of units N is large, there would be 2^N potential treatment assignments, which is a large number in practice. In such cases, given D^{obs} and Algorithm 1, we can show that $\|\hat{pval}^{pair} - pval^{pair}(D^{obs})\| = O_p(R^{-1/2})$. Specifically, by $\hat{pval}^{pair} = (1+\sum_{r=1}^{R} 1\{T_r \geq T_r^{obs}\})/(1+R)$ and $d^r \sim P$ independently, we have $E_{d^r}\hat{pval}^{pair} = pval^{pair}(D^{obs})$ and

$$
\begin{aligned}
Var(\hat{pval}^{pair}) &= Var(1\{T_r \geq T_r^{obs}\})/(1+R) \\
&= pval^{pair}(D^{obs})(1 - pval^{pair}(D^{obs}))/(1+R).
\end{aligned}
$$

Hence, by Chebyshev's inequality, $\|\hat{pval}^{pair} - pval^{pair}(D^{obs})\| = O_p(R^{-1/2})$.

A.2 FRAMEWORK FOR INTERSECTION OF NULL HYPOTHESES

Some hypotheses of interest can be expressed as the intersection of the partially sharp null hypotheses discussed in the main text. For example, Athey et al. (2018) and Puelz et al. (2021) define the following null hypothesis regarding the extent of interference at distance k:

Definition A.2.1 (Extent of Interference for Distance k in Puelz et al. (2021))**.** *In a social network, the null hypothesis at distance k states that for all $i = 1, \ldots, N$,*

$$
Y_i(d) = Y_i(d') \quad \textit{for any } d, d' \in \{0,1\}^N \textit{ such that } d_j = d'_j \textit{ for all } j \textit{ with } d(i,j) \leq k.
$$

This hypothesis asserts that a unit's outcome depends only on treatments within k-hops, not beyond. Unlike the partially sharp null in Definition 1, which requires potential outcomes to remain the same across one subset of assignments, this hypothesis allows unit i's outcome to change whenever a nearby unit j (with $d(i,j) \leq k$) switches treatment status. Nevertheless, for each combination of assignment statuses within k-distance, there is a subset of assignments yielding the same outcome for i.

Moreover, the null hypothesis in Definition A.2.1 differs from that in the main text: each combination of treatment statuses within k-distance can be viewed as a separate partially sharp null hypothesis. The main text, by contrast, focuses on a particular partially sharp null in which all units within k-distance are untreated.

Nonetheless, the framework presented in this paper naturally extends to intersections of partially sharp null hypotheses with minor modifications. We redefine the partially sharp null hypothesis in terms of a collection $\mathcal{D}^a = \{\mathcal{D}_i^a\}_{i=1}^N$:

Definition A.2.2 (Partially sharp null Hypothesis for $\mathcal{D}^a$)**.** *A partially sharp null hypothesis holds if there exists a collection of subsets $\mathcal{D}^a = \{\mathcal{D}_i^a\}_{i=1}^N$, where each $\mathcal{D}_i^a \subsetneq \{0,1\}^N$, such that*

$$H_0^{\mathcal{D}^a} : Y_i(d) = Y_i(d') \quad \textit{for all } i \in \{1, \ldots, N\}, \quad \textit{and any } d, d' \in \mathcal{D}_i^a.$$

Definition A.2.3 (Intersection of partially sharp null Hypotheses)**.** *For each $a \in \mathbb{F} = \{1, \ldots, F\}$ and the given $\mathcal{D}^a$, the intersection of partially sharp null hypotheses is defined as*

$$\bigcap_{a \in \mathbb{F}} H_0^{\mathcal{D}^a} := H_0^{\mathbb{F}},$$

which is equivalent to

$$H_0^{\mathbb{F}} : Y_i(d) = Y_i(d') \quad \textit{for all } i \in \{1, \ldots, N\}, \quad \textit{and any } d, d' \textit{ s.t. } \exists a \in \mathbb{F} \textit{ with } d, d' \in \mathcal{D}_i^a.$$

With slight adjustments to the definitions of imputable unit sets and pairwise imputable statistics, our main procedure can also test intersection null hypotheses $H_0^{\mathbb{F}}$.

A.2.1 Testing the Intersection of partially sharp null Hypotheses

Definition A.2.4 (Imputable Units (Intersection)). *Given two treatment assignments* $d, d' \in \{0,1\}^N$ *and an intersection of partially sharp null hypotheses* $H_0^{\mathbb{F}}$*, define*

$$\mathbb{I}(d,d') \equiv \big\{ i \in \{1,\dots,N\} : \exists\, a \in \mathbb{F} \text{ s.t. } d, d' \in \mathcal{D}_i^a \big\} \subseteq \{1,\dots,N\}.$$

as the imputable units set *under treatment assignments* d *and* d'*.*

Definition A.2.5 (Imputable Outcome Vector (Intersection)). *For* $d, d' \in \{0,1\}^N$ *and an intersection of partially sharp null hypotheses* $H_0^{\mathbb{F}}$*, the vector*

$$Y_{\mathbb{I}(d,d')} \equiv \{Y_i\}_{i \in \mathbb{I}(d,d')}.$$

is called the imputable outcome vector. If Y *is evaluated under a third assignment* d''*, then*

$$Y_{\mathbb{I}(d,d')}(d'') \equiv \big\{Y_i(d'')\big\}_{i \in \mathbb{I}(d,d')}.$$

Definition A.2.6 (Pairwise Imputable Statistic (Intersection)). *Let* $T : \mathbb{R}^N \times \{0,1\}^N \times \{0,1\}^N \longrightarrow \mathbb{R} \cup \{\infty\}$ *be a measurable function, and let* $Y_{\mathbb{I}(d,d')}$ *be the imputable outcome vector. We say* T *is a* pairwise imputable statistic *if, for any* $d, d' \in \{0,1\}^N$*, the following holds:*

$$\text{whenever } Y_i = Y_i' \text{ for all } i \in \mathbb{I}(d,d'), \quad \text{then} \quad T\big(Y_{\mathbb{I}(d,d')},\, d'\big) = T\big(Y'_{\mathbb{I}(d,d')},\, d'\big).$$

Proposition A.213. *Suppose the intersection of partially sharp null hypotheses* $H_0^{\mathbb{F}}$ *holds. Let* $T(Y_{\mathbb{I}(d,d')}, d')$ *be a pairwise imputable statistic. Then*

$$T\big(Y_{\mathbb{I}(d,d')}(d),\, d'\big) = T\big(Y_{\mathbb{I}(d,d')}(d'),\, d'\big)$$

for any $d, d' \in \{0,1\}^N$.

Definition A.2.7 (PIRT (Intersection)). *A PIRT is an unconditional randomization test defined by*

$$\phi^{\text{pair}}(D^{\text{obs}}) \;=\; 1\big\{pval^{\text{pair}}(D^{\text{obs}}) \le \alpha/2\big\},$$

where the p-*value function* $pval^{\text{pair}}(D^{\text{obs}})$ *is given by*

$$pval^{\text{pair}}(D^{\text{obs}}) \;=\; P\Big(T\big(Y_{\mathbb{I}(D^{\text{obs}},D)}(D^{\text{obs}}),\, D\big) \ge T\big(Y_{\mathbb{I}(D^{\text{obs}},D)}(D^{\text{obs}}),\, D^{\text{obs}}\big)\Big) \text{for } D \sim P,$$

and $T(Y_{\mathbb{I}(d,d')}, d')$ *is a pairwise imputable statistic.*

Theorem A.2.1. *Suppose the intersection of partially sharp null hypotheses* $H_0^{\mathbb{F}}$ *holds. Then the PIRT in Definition A.2.7 satisfies*

$$\mathbb{E}_P\big[\phi^{\text{pair}}(D^{\text{obs}})\big] \;<\; \alpha \quad \textit{for any } \alpha \in (0,1),$$

where the expectation is taken over $D^{\text{obs}} \sim P$.

The proofs follow the main text but use two treatment assignments when defining imputable units.

A.3 THE MINIMIZATION-BASED PIRT

The main limitation of the PIRT is that when rejecting the null hypothesis at significance level α, the probability of a false rejection can be as high as 2α instead of α. While one way to address this is to reject the null hypothesis when the *p*-value is below $\alpha/2$, a more conservative testing procedure inspired by Wen et al. (2023) can be considered. The core idea behind this minimization-based PIRT is to compute a test statistic that reflects the worst-case scenario across all possible treatment assignments. Specifically, I define the test statistic as

$$\tilde{T}(D^{obs}) = \min_{d \in \{0,1\}^N} T(Y_{\mathbb{I}(d)}(D^{obs}), D^{obs}),$$

where the test statistic T is evaluated for each potential treatment assignment d. Based on this, I define the *p*-value as follows.

Definition A.3.1 (Minimization-based PIRT)**.** *The minimization-based PIRT is an unconditional randomization test defined by* $\phi^{min}(D^{obs}) = 1\{pval^{min}(D^{obs}) \leq \alpha\}$*, where* $pval^{min}(D^{obs}) : \{0,1\}^N \to [0,1]$ *is the p-value function:*

$$pval^{min}(D^{obs}) = P(T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) \geq \tilde{T}(D^{obs})) \textit{ for } D \sim P.$$

Here, $T(Y_{\mathbb{I}(d)}(d), d')$ *represents the pairwise imputable statistic used to evaluate the hypothesis.*

To calculate this *p*-value in practice, Algorithm A.3.1 is applied. It computes the mean of $1 + R$ draws, where $r = 0$ corresponds to $d = D^{obs}$.

Algorithm A.3.1 Minimization-Based PIRT Procedure

```
Inputs   : Test statistic $T = T(Y(d), d)$, observed assignment $D^{obs}$, observed out-
           come $Y^{obs}$, treatment assignment mechanism $P$, and size $\alpha$.
for $r = 1$ to $R$ do
    Randomly sample $d^r \sim P$, and store $T_r \equiv T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), d^r)$.
    Store $T_r^{obs} \equiv T(Y_{\mathbb{I}(d^r)}(D^{obs}), D^{obs})$.
end
Compute: $\tilde{T}^\star(D^{obs}) = \min_{r=1,\ldots,R}(T_r^{obs})$
Output   : p-value: $\widehat{pval}^{min} = \frac{1+\sum_{r=1}^{R} 1\{T_r \geq \tilde{T}^\star(D^{obs})\}}{1+R}$.
           Reject if $\widehat{pval}^{min} \leq \alpha$.
```

In the toy example, shown in Table 1.6, $\tilde{T}(D^{obs}) = 1$; thus, $pval^{min} = 1/2$. The crucial distinction between minimization-based PIRT and PIRT is that minimization ensures size control, as demonstrated by Theorem A.3.1.

Theorem A.3.1. *Suppose the partially sharp null hypothesis* $H_0^{\epsilon_s}$ *holds. Then, the minimization-based PIRT, as defined in Definition A.3.1, satisfies* $\mathbb{E}_P[\phi^{min}(D^{obs})] \leq \alpha$ *for any* $\alpha \in (0, 1)$*, where the expectation is taken with respect to* $D^{obs} \sim P$*.*

Proof of Theorem A.3.1. To avoid confusion, denote $P_{D^{obs}}$ as probability respect to D^{obs} and P_D as probability respect to D.

Under the null $H_0^{\epsilon_s}$, by Proposition 1 and setting $d = D$, $d' = D^{obs}$, we have $T(Y_{\mathbb{I}(D)}(D), D^{obs}) = T(Y_{\mathbb{I}(D)}(D^{obs}), D^{obs})$. Hence, we have $\tilde{T}(D^{obs}) = min_{d \in \{0,1\}^N}(T(Y_{\mathbb{I}(d)}(d), D^{obs}))$.

Then, by construction, $\tilde{T}(D^{obs}) \sim \tilde{T}(D) \leq T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D)$, and

$$pval^{min}(D^{obs}) = P_D(T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) \geq \tilde{T}(D^{obs})) \geq P_D(\tilde{T}(D) \geq \tilde{T}(D^{obs})).$$

Therefore,

$$P_{D^{obs}}(pval^{min}(D^{obs}) \leq \alpha) \leq P_{D^{obs}}(P_D(\tilde{T}(D) \geq \tilde{T}(D^{obs})) \leq \alpha).$$

Let U be a random variable with the same distribution as $\tilde{T}(D)$, induced by P. Denote its cumulative distribution function by F_U. We then have $P_D(\tilde{T}(D) \geq \tilde{T}(D^{obs})) = 1 - F_U\{\tilde{T}(D^{obs})\}$, which is a random variable induced by $D^{obs} \sim P(D^{obs})$. Hence, $P_D(\tilde{T}(D) \geq \tilde{T}(D^{obs})) = 1 - F_U(U)$, and by the probability integral transformation, $P_D(\tilde{T}(D) \geq \tilde{T}(D^{obs}))$ respect to D^{obs} has a uniform $[0, 1]$ distribution under $H_0^{\epsilon_s}$. Thus, for any $\alpha \in [0, 1]$,

$$P_{D^{obs}}(pval^{min}(D^{obs}) \leq \alpha) \leq P_{D^{obs}}(P_D(\tilde{T}(D) \geq \tilde{T}(D^{obs})) \leq \alpha) \leq \alpha.$$

□

Handling a Large Number of Potential Treatment Assignments. When N is large, finding the minimum $\tilde{T}(D^{obs})$ across all possible treatment assignments can be computationally intensive. To ensure the validity of Algorithm A.3.1 when dealing with a large number of units, optimization methods can be used to approximate $\tilde{T}^R(D^{obs})$ such that $\tilde{T}(D^{obs}) \geq \tilde{T}^R(D^{obs}) - \eta_R$ with probability $1-\eta$. The rejection level can then be adjusted to $\tilde{\alpha}$ such that $\alpha = \tilde{\alpha}(1-\eta)+\eta$, thereby ensuring validity. However, this approach introduces additional computational complexity.

An alternative strategy is to combine CRTs with PIRTs to reduce the space of potential treatment assignments. As discussed by Athey et al. (2018) and Zhang & Zhao (2023), researchers often limit the assignment space to only those assignments with the same number of treated units as in the observed assignment. This two-stage approach–first defining the number of treated units and then performing testing within the reduced assignment space–remains valid. Using PIRTs in this context increases the set of focal units, potentially improving test power.

A.4 DISCUSSION ON SOME EXTREME CASES

Emptiness of Imputable Units Set. The emptiness of the imputable units set depends on three factors: the distance being tested, the network structure, and the randomization design.

First, the target distance interacts with the network structure. If the distance $\epsilon_s > \max_{i,j} G_{i,j}$, meaning it exceeds any existing distance in the network, then there will be no units in the imputable units set. In this case, additional data may be required to gain sufficient power for the test, or the target distance ϵ_s could be reduced. For clarity in the following discussion, we focus on the case where $\epsilon_s = 0$.

Second, with $\epsilon_s = 0$, if all units in the sample are treated, the imputable units set will still be empty. To detect the existence of interference, a sufficient number of units beyond our target distance across various treatment assignments is necessary to achieve reasonable power.

Cases with an Undefined Comparison Group. The distance being tested, network structure, and randomization design also influence whether one of the comparison groups is undefined. To highlight the core intuitions, we focus on the case where $\epsilon_s = 0$, implying that we are testing for the existence of interference and not all units are treated. Thus, some untreated units remain to conduct the test.

A general example is a network of couples, where exactly one unit in each pair is treated. In the example from the main text, with treatment assignments rotating across couples, the neighborhood units set may be empty. In practice, the test statistic must then assume a very high value for implementation.

More generally, for each assignment d and given ϵ_c, let $\|\{i : d_i = 1\}\|$ denote the number of treated units, $\|\{i : d \in \mathcal{D}_i(0)/\mathcal{D}_i(\epsilon_c)\}\|$ the number of units in the neighborhood set, and $\|\{i : d \in \mathcal{D}_i(\epsilon_c)\}\|$ the number in the control set. Whenever the number of non-imputable units is equal to or exceeds the number of neighborhood units, there may exist a pair of assignments (D^{obs}, D) such that the neighborhood units set is empty. This principle also applies to the control units set. Therefore, to ensure that both neighborhood and control sets are defined, we impose Assumption A.44.

Assumption A.44 (Regularization when $\epsilon_s = 0$)**.**

$$\min\left\{\min_{d\in\{0,1\}^N} \|\{i : d \in \mathcal{D}_i(0)/\mathcal{D}_i(\epsilon_c)\}\|, \min_{d\in\{0,1\}^N} \|\{i : d \in \mathcal{D}_i(\epsilon_c)\}\|\right\}$$

$$> \max_{d\in\{0,1\}^N} \|\{i : d_i = 1\}\|$$

It is worth noting that $\|\{i : d_i = 1\}\| = N - \|\mathbb{I}(d)\|$ when $\epsilon_s = 0$. Therefore, Assumption A.44 implies that the groups of interest occupy a large proportion of the population across all treatment assignments. This condition depends on ϵ_c, the network structure, and the experimental design. With Assumption A.44, we ensure all comparison groups remain non-empty across different potential assignments.

Proposition A.414. *Suppose Assumption A.44 holds. For any $D^{obs} \in \{0,1\}^N$, the pairwise imputable statistic $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), D) \neq \infty$ across different D.*

Proof of Proposition A.414. We proceed by contradiction. Assume there exists a $d \in \{0,1\}^N$ such that $T(Y_{\mathbb{I}(D^{obs})}(D^{obs}), d) = \infty$. Without loss of generality, suppose for any $i \in \mathbb{I}(D^{obs})$, $d \notin \mathcal{D}_i(\epsilon_c)$. Then, for any unit $j \in \{i : d \in \mathcal{D}_i(\epsilon_c)\}$, it must be the case that $j \notin \mathbb{I}(D^{obs})$, and hence $j \in \{i : D_i^{obs} = 1\}$.

Thus, we have

$$\|\{i : d \in \mathcal{D}_i(\epsilon_c)\}\| \leq \|\{i : D_i^{obs} = 1\}\|$$

However, we have

$$\|\{i : d \in \mathcal{D}_i(\epsilon_c)\}\| \geq \min_{d\in\{0,1\}^N} \|\{i : d \in \mathcal{D}_i(\epsilon_c)\}\|$$

and

$$\|\{i : D_i^{obs} = 1\}\| \leq \max_{d \in \{0,1\}^N} \|\{i : d_i = 1\}\|.$$

This implies that

$$\min_{d \in \{0,1\}^N} \|\{i : d \in \mathcal{D}_i(\epsilon_c)\}\| \leq \max_{d \in \{0,1\}^N} \|\{i : d_i = 1\}\|,$$

which contradicts Assumption A.44.

□

A.5 FRAMEWORK TO DETERMINE THE BOUNDARY OF INTERFERENCE

Building on the PIRT framework, we can determine the boundary of interference by estimating a sequence of partially sharp null hypotheses at varying distances ϵ_s. This approach is useful for selecting a pure control distance or assessing the extent of interference based on distance. To this end, I consider a sequence of distance thresholds:

$$\epsilon_0 < \epsilon_1 < \epsilon_2 < \cdots < \epsilon_K < \infty,$$

where $K \geq 1$ is chosen to include the settings introduced in previous sections. For instance, if the goal is to test for the existence of interference, one could set $K = 1$ with $\epsilon_0 = \epsilon_s = 0$ and $\epsilon_1 = \epsilon_c$.

Using this sequence of distances, I can test a series of null hypotheses as defined in Definition 3, where $\epsilon_s \in \{\epsilon_0, \ldots, \epsilon_K\}$. However, it is important to note that not all distance levels will yield non-trivial power. First, there is a trade-off between the number of thresholds tested and the power of each test. While testing more thresholds provides a richer understanding of how interference varies with distance, it can reduce the power to detect interference, especially if certain threshold groups lack sufficient units. Based on simulation results, I recommend ensuring that each exposure level includes at least 20 units to maintain sufficient power at a significance level of $\alpha = 0.05$.

Second, in some cases, ϵ_K may represent the maximum distance in the network, leaving no further room for ϵ_c. Although it remains possible to test $H_0^{\epsilon_K}$, alternative approaches—such as adjusting for the number of nearby treated units, as suggested by Hoshino & Yanagi (2023)—may be needed to construct a test statistic with non-trivial power. For simplicity, this section will focus on testing $H_0^{\epsilon_k}$ for $k \leq K - 1$.

Following Definition 3 of $H_0^{\epsilon_s}$, the multiple hypotheses under consideration exhibit a nested structure:

Proposition A.515. *Suppose there exists an index $\bar{K} \geq 0$ such that for any $k \leq \bar{K} - 1$, the partially sharp null hypothesis $H_0^{\epsilon_k}$ is false and $H_0^{\epsilon_{\bar{K}}}$ is true. Then, $H_0^{\epsilon_k}$ is true for any $k \geq \bar{K}$.*

Proof of Proposition A.515. By Definition 3, if $H_0^{\epsilon_{\bar{K}}}$ is true, then $Y_i(d) = Y_i(d')$ for all $i \in \{1, \ldots, N\}$ and any $d, d' \in \mathcal{D}_i(\epsilon^{\bar{K}})$.

Observe that for any $i \in \{1, \ldots, N\}$, by Definition 2,

$$\mathcal{D}_i(\epsilon_0) \supset \mathcal{D}_i(\epsilon_1) \supset \cdots \supset \mathcal{D}_i(\epsilon_K).$$

Thus, for any $k \geq \bar{K}$ and any $d, d' \in \mathcal{D}_i(\epsilon_k) \subseteq \mathcal{D}_i(\epsilon_{\bar{K}})$, it follows that $Y_i(d) = Y_i(d')$ for all $i \in \{1, \ldots, N\}$. By Definition 3, $H_0^{\epsilon_k}$ is true for any $k \geq \bar{K}$. $\square$

Proposition A.515 implies that interference is bounded within a certain distance. Given this nested structure, I aim to develop an inference method that determines such boundaries by rejecting the null hypothesis up to a certain distance and failing to reject it beyond that point. However, in practice, situations may arise where $H_0^{\epsilon_k}$ cannot be rejected but $H_0^{\epsilon_{k+1}}$ is rejected. This could happen either because the test lacks power to reject the false null $H_0^{\epsilon_k}$ or due to multiple hypothesis testing errors, which lead to an erroneous rejection of the true null $H_0^{\epsilon_{k+1}}$. To mitigate the risk of over-rejecting true null hypotheses, I propose controlling the FWER.

Definition A.5.1 (FWER over all $H_0^{\epsilon_k}$ for $k = 0, \ldots, K-1$)**.** *Given a test* $\varphi : \{0,1\}^N \to \{0,1\}^K$*, which maps the data to decisions for each hypothesis* $H_0^{\epsilon_k}$*, the family-wise error rate (FWER) is defined as*

$$\mathit{FWER} = P\left(\exists k \geq \bar{K} \text{ such that } \varphi_k(D^{obs}) = 1, \text{ meaning that } H_0^{\epsilon_k} \text{ is rejected}\right),$$

where $\bar{K} \geq 0$ *is such that for any* $k \leq \bar{K} - 1$*,* $H_0^{\epsilon_k}$ *is false, and for* $k \geq \bar{K}$*,* $H_0^{\epsilon_k}$ *is true.*

The definition of the FWER in Definition A.5.1 is motivated by the nested structure of $H_0^{\epsilon_k}$, where the null hypothesis is true for any $k \geq \bar{K}$. The critical issue is determining how to reject all the $H_0^{\epsilon_k}$ hypotheses when identifying the boundary of interference, while still ensuring control over the FWER.

A.5.1 A Valid Procedure to Determine the Neighborhood of Interference

A major challenge in testing the extent of interference with respect to distance lies in addressing the issue of multiple hypothesis testing when conducting a series of tests to identify the neighborhood of interference. To manage the increased error rate arising from multiple tests, and drawing inspiration from Meinshausen (2008) and Section 15.4.4 of Lehmann & Romano (2005), I propose Algorithm A.5.1.

Algorithm A.5.1 is designed to control the FWER while leveraging the nested

Algorithm A.5.1 Sequential Testing Procedure

```
Inputs : Test statistic $T = T(Y(d), d)$, observed assignment $D^{obs}$, observed outcome
         $Y^{obs}$, and treatment assignment mechanism $P$.
Set    : $\hat{K} = 0$.
for $k = 0$ to $K - 1$ do
    Test $H_0^{\epsilon_k}$ using the PIRT procedure and collect $pval^k$.
    If $pval^k \leq \alpha$, set $\hat{K} = k + 1$ and reject $H_0^{\epsilon_k}$.
    If $pval^k > \alpha$, break.
end
Output: Significant spillover within distance $\epsilon_{\hat{K}}$.
```

structure of sequential hypothesis testing. Unlike traditional multiple hypothesis testing procedures, such as the Bonferroni-Holm method, which require rejecting at a smaller level than α, this algorithm maintains the significance level without adjustment, potentially increasing power compared to conventional methods (Meinshausen, 2008). Moreover, if the unadjusted p-values increase as k increases, indicating that interference diminishes with distance, there is no loss of power compared to not adjusting for multiple hypothesis testing, as we would naturally stop rejecting beyond a certain distance. When using the PIRT for each k, rejecting at the $\alpha/2$ level ensures size control. For the partially sharp null hypothesis $H_0^{\epsilon_k}$, a natural choice for ϵ_c is ϵ_{k+1}. Theorem A.5.1 guarantees the FWER control of Algorithm A.5.1.

Theorem A.5.1. *The sequential testing procedure constructed by Algorithm A.5.1 controls the FWER at α.*

Proof of Theorem A.5.1. Without loss of generality, consider the minimization-based PIRT below. The same proof holds when using the PIRT with a rejection level of $\alpha/2$.

Suppose for any $k < \bar{K}$, $H_0^{\epsilon_k}$s are false and $H_0^{\epsilon_{\bar{K}}}$ is true. Then, by Algorithm

A.5.1, if there exist $k \geq \bar{K}$ such that $H_0^{\epsilon_k}$ is rejected, it must be the case that $H_0^{\epsilon_{\bar{K}}}$ is rejected. Thus, by Definition A.5.1,

$$FWER = P(pval^1 \leq \alpha, pval^2 \leq \alpha, \ldots, pval^{\bar{K}} \leq \alpha) \leq P(pval^{\bar{K}} \leq \alpha) \leq \alpha$$

because $H_0^{\epsilon_{\bar{K}}}$ is true. □

For example, suppose $K = 2$ with $(\epsilon_0, \epsilon_1, \epsilon_2) = (0, 1, 2)$. Algorithm A.5.1 can be implemented in two steps. First, collect $pval^0$ for H_0^0 and reject H_0^0 if $pval^0 \leq \alpha$. If H_0^0 is not rejected, report that no significant interference was found. If H_0^0 is rejected, proceed to the second step, collect $pval^1$ for H_0^1, and reject H_0^1 if $pval^1 \leq \alpha$. If H_0^1 is rejected, report significant interference within distance 2; if H_0^2 is not rejected, report significant interference within distance 1.

A.5.2 Rationale for Using the FWER

Controlling the family-wise error rate (FWER) is not the only option in multiple testing. As Anderson (2008) suggests, a false discovery rate (FDR) control may be more suitable for exploratory analyses by allowing a small number of type I errors in return for greater power. An FDR-based approach could be explored in future work. However, when policymakers intend to apply a policy in distant regions under a positive interference effect, the more restrictive FWER control prevents overly optimistic conclusions about the interference boundary. In such settings, FWER provides a conservative distance threshold and better accounts for interference in expected welfare calculations.

This procedure also aids in identifying a pure control group by defining a "safe distance" ϵ_c, as mentioned in Section 1.3.1. One natural choice is ϵ_K, the greatest distance at which non-trivial testing power remains. However, researchers might reduce this distance to include more control units and boost power. Algorithm A.5.1 offers a principled method for selecting ϵ_c, but the resulting value may be smaller than the true interference boundary due to the algorithm's conservative nature. Weighing these trade-offs is crucial when deciding whether to incorporate a pre-testing step.

A.6 INCORPORATING COVARIATE ADJUSTMENT

In practice, we often have access to covariates X, and incorporating this information is crucial for enhancing the power of tests, particularly when these covariates are predictive of potential outcomes (Wu & Ding, 2021). Since the choice of test statistic does not affect the validity of the testing procedure for the partially sharp null hypothesis of interest, I propose three approaches for incorporating covariates in the analysis.

The first approach is PIRT with regressions. As illustrated in the main text, this method involves conducting the PIRT using regression coefficients from a simple OLS model as the test statistic. This OLS model includes a binary variable indicating whether a unit receives spillovers at a certain distance and known covariates, such as information about the neighborhood and social center points. A similar approach is discussed in Puelz et al. (2021).

The second approach is PIRT with residual outcomes. The key idea here is to use the residuals from a model-based approach, such as regression with covariates of interest, rather than the raw outcome variables. I first obtain predicted values $\hat{Y}_i$ for the sample outcomes and then use the residuals, defined as the difference between observed outcomes and predicted values $\hat{e}_i = Y_i^{obs} - \hat{Y}_i$, for the PIRT procedures as the Y defined in the main text. A similar approach for FRTs is proposed by Rosenbaum (2020), with detailed discussion in Sections 7 and 9.2 of Basse & Feller (2018).

The third approach is PIRT using pairwise residuals. In this method, for each pair of treatment assignments (D^{obs}, D), I conduct a regression with covariates within the imputable units set to transform the outcomes into residuals before testing and constructing the *p*-values accordingly. This approach can be viewed as combining the first and second methods.

A.6.1 Investigation on the Power of Incorporating Covariates

In this investigation, we extend the potential outcomes described in Table A.7.1 by incorporating two covariates, X_1 and X_2. The new control potential outcomes, $Y_i^C(\text{new})$, are simulated based on the original control outcomes Y_i^C from Table A.7.1 as follows:

$$Y_i^C(\text{new}) = 2 + 0.5 \times X_1 + 0.3 \times X_2 + Y_i^C,$$

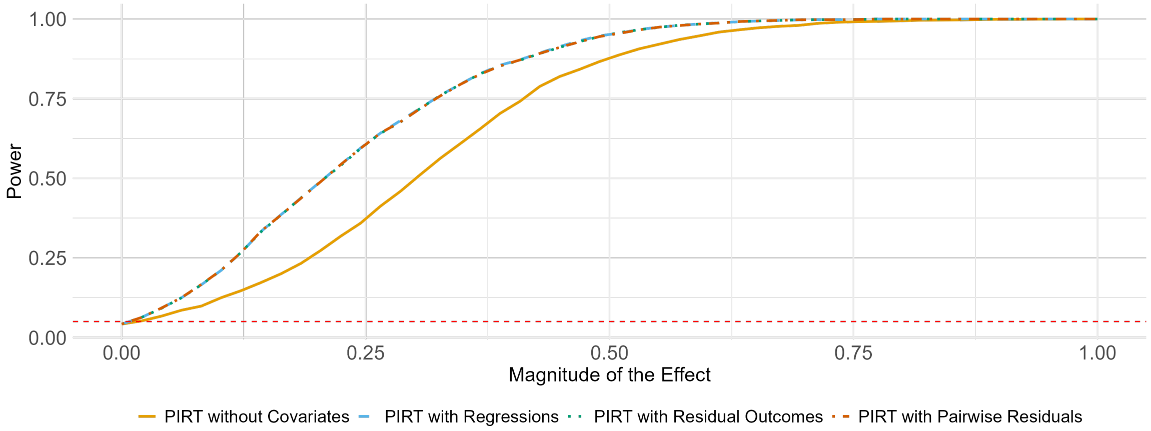
where X_1 is a binary covariate drawn from a Bernoulli distribution with parameter 0.5, and X_2 is a continuous covariate drawn from a standard normal distribution:

$$X_1 \sim \text{Bernoulli}(0.5), \quad X_2 \sim \mathcal{N}(0, 1).$$

It is important to note that only the control potential outcomes, Y_i^C, are modified by these covariates. The remaining potential outcomes for treated units follow the same functional relationships as described in Table A.7.1. By introducing X_1 and X_2, the control potential outcomes become more variable, reflecting the added noise from the covariates.

Next, I apply the three methods introduced earlier—PIRT with regressions, PIRT with residual outcomes, and PIRT using pairwise residuals—to construct the power curves. The simulation procedure remains consistent with that described in the main text, focusing on displacement effects and one-sided tests using non-absolute coefficients.

Figure A.6.1: Power Comparison of Testing Methods for Different Covariate Adjustments



Notes: The red line indicates the size level $\alpha = 0.05$. Power is based on the PIRT with rejection at level α. I consider 50 equally spaced values of τ between 0 and 1, conducting 2,000 simulations for each τ to compute the average rejection rate for each method.

Figure A.6.1 illustrates the power gains achieved by incorporating covariate information. While all methods involving covariate adjustments demonstrate similar power performance, leveraging covariates consistently results in higher power. When $\tau = 0$, the rejection rates for all methods align with the nominal size of the test. As τ increases, the power also increases. For instance, when $\tau \approx 0.25$, the power of the test with covariate adjustment reaches approximately 0.65, compared to less than 0.4 for the test without covariate adjustments. Therefore, in practice, researchers should select the method that best suits their specific context and data.

A.6.2 Robustness of Results to Adjustment Methods

The application of the above methods yields the results presented in Table A.6.1. The regression models closely follow the framework outlined in Blattman et al. (2021), with slight modifications.

First, the regression includes the same covariates used in Blattman et al. (2021), such as police station fixed effects, but excludes those related to the municipal services intervention.[1] In the original study, randomization testing was conducted jointly for both the policing and municipal services interventions, complicating interpretation when interaction effects are present. In this analysis, I hold the municipal services intervention fixed to isolate the effect of intensive policing.

Second, Blattman et al. (2021) employs inverse propensity weighting in the weighted regression, using weights that account for both the hotspot policing and municipal services interventions.[2] In this replication, I use weights that only consider the hotspot policing intervention to focus specifically on the impact of intensive policing.

As shown in Table A.6.1, the p-values are very similar across the different methods, allowing researchers to choose the most practical implementation. Additionally, as discussed in Section C.3 of Basse et al. (2024), one can stratify potential assignments based on covariates to balance the focal units. This is done by stratifying both the permutations and the test statistic by an additional discrete covariate. However, we could not implement and compare p-values from this method due to

---

[1] The covariates include the following: number of crimes (20122015); average patrol time per day; square meters built (100 meters around) per meter of longitude; distance to the nearest shopping center, educational center, religious/cultural center, health center, and additional services office (e.g., justice); transport infrastructure (e.g., bus/BRT station); indicators for industry/commerce zones and service sector zones; income level; eligibility for municipal services; and interactions with the crime hotspot indicator.

[2] Although this method does not fully eliminate bias, as discussed in Aronow et al. (2020), it helps address imbalance in the spillover group.

Table A.6.1: p-Values: PIRT with Different Specifications

	Unadjusted p-values		
	$(0m, \infty)$	$(125m, \infty)$	$(250m, \infty)$
Violent crime			
Reg (WLS)	0.105	0.719	0.158
Reg (OLS)	0.156	0.767	0.110
Pair residuals	0.119	0.726	0.142
Residuals outcome	0.114	0.757	0.166
Property crime			
Reg (WLS)	0.508	0.232	0.619
Reg (OLS)	0.494	0.462	0.560
Pair residuals	0.481	0.252	0.565
Residuals outcome	0.455	0.250	0.578

Notes: The table shows p-values of PIRT across different methods, using the number of crimes as the outcome variable. Reg (WLS) is PIRT with regression, using the coefficient from the covariates-included regression with inverse propensity weighting as the test statistic. Reg (OLS) is the PIRT with regression, using the coefficient from the covariates-included regression without weighting as the test statistic. Pair residuals are PIRT with pairwise residuals, where residuals are constructed from the pairwise subset regression in the first step. The coefficient from the no-covariates regression with inverse propensity weighting is then used as the test statistic. Residuals outcome is PIRT with the residuals outcome, where residuals are constructed for all units in the first step, followed by using the coefficient from the no-covariates regression with inverse propensity weighting as the test statistic.

limitations in the original dataset.

In line with Puelz et al. (2021), including covariates raises p-values, suggesting that distance alone may not capture all heterogeneity in spillover effects.[3] Covariates can reveal that the partially sharp null hypothesis does not fully account for unit-level heterogeneity. In an extreme scenario, if spillovers are perfectly correlated with these covariates, the partially sharp null would be rejected; however, regression adjustment could then eliminate the spillover signal, raising p-values

---

[3] Most of this p-value increase stems from demographic covariates such as income levels and building density.

under the same null. Future research may refine distance measures by incorporating additional factors (e.g., socioeconomic disparities) to better capture spillover intensity (Puelz et al., 2021).

Researchers should interpret these results cautiously and decide on the null hypothesis of interest beforehand. If a researcher is interested in testing for no spillover effects after controlling for covariates, PIRTs can be extended to accommodate the work by Ding et al. (2016). One can refer to Owusu (2023) for investigating heterogeneous effects in network settings. Alternatively, if interested in the weak null of the average effect being equal to zero (see Zhao & Ding (2020); Basse et al. (2024)), one should note that the construction of p-values in PIRTs differs from those in CRTs and FRTs, making classical approaches for weak nulls potentially inapplicable. Further investigation into these differences would be of interest to future research.

A.7 ALGORITHM FOR SIMULATION EXERCISE

I generate $N = 1,000$ points from a bivariate Gaussian distribution with non-diagonal covariance to simulate the network on a $[0,1] \times [0,1]$ space. Figure A.7.1 shows the unit distribution within this space.

I focus on two distance thresholds, with $(\epsilon_0, \epsilon_1, \epsilon_2) = (0, 0.1, 0.2)$. Across different treatment assignments, the distance interval $(0, 0.1]$ comprises approximately 420 units, $(0.1, 0.2]$ around 250 units, and the pure control group $(0.2, \infty)$ around 320 units.

The algorithm for the simulation exercise in Section 1.4.1 is outlined in Algorithm A.7.1.

Figure A.7.1: Unit Distribution

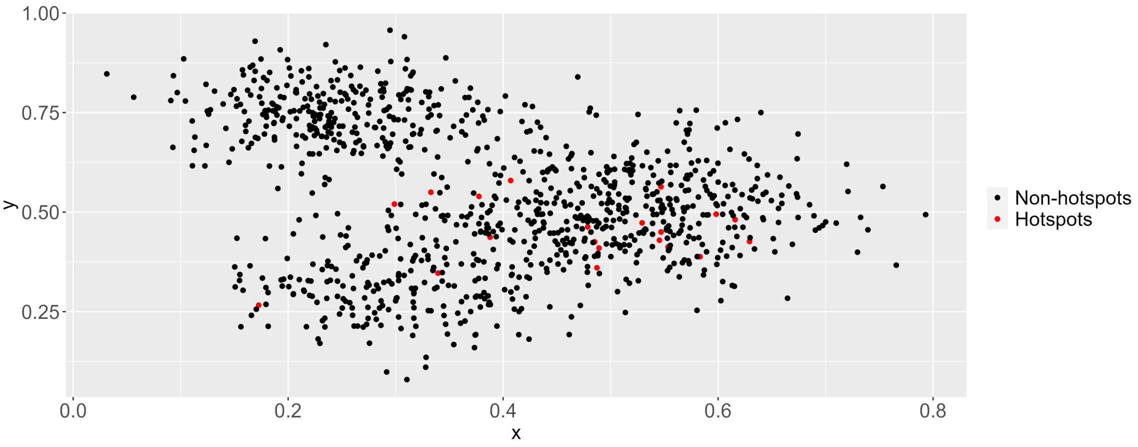


Table A.7.1: Potential Outcome Schedule in the Simulation

Pure control for non-hotspots:	$Y_i^C \sim Gamma(0.086, 3.081)$
Pure control for hotspots:	$Y_i^C \sim Gamma(0.737, 1.778)$
Treated unit:	$Y_i^T = \max(Y_i^C - 1, 0)$
Short-range spillover:	$Y_i(d) = Y_i^C + \tau \quad \forall d \in \mathcal{D}_i(0)/\mathcal{D}_i(0.1)$
Long-range spillover:	$Y_i(d) = Y_i^C + 0.5\tau \quad \forall d \in \mathcal{D}_i(0.1)/\mathcal{D}_i(0.2)$

Notes: The outcome schedule is calibrated to the observed dataset. For $Gamma(k, \theta)$, k is the shape parameter and θ is the scale parameter. Y_i^C represents the pure control potential outcome for unit i, and Y_i^T represents the potential outcome for unit i when treated.

Algorithm A.7.1 Simulation Study Procedure

```
Inputs : 5,000 randomly chosen assignments as the potential assignments set, $\mathbb{D}_S$.
         The biclique decomposition of $\mathbb{D}_S$ from Puelz et al. (2021).
Set    : Spillover effect $\tau$ and corresponding schedule of potential outcomes.
for $s = 1 : S$ do
    Sample $D_s^{obs}$ from $\mathbb{D}_S$, and generate $Y_s^{obs}$.
    Implement the algorithms and collect corresponding $pval(D_s^{obs})$ using $R = 1,000$.
end
Output: Average the number of rejections to obtain the power for that fixed $\tau$.
```

APPENDIX B

Supplementary Materials for Chapter Two

B.1 PROOFS FOR THE MAIN RESULTS

The proofs will make repeated use of the following mean-value identity.

Lemma B.11 (Outer Mean-Value Lemma)**.** *For any $g(\cdot)$ continuous differentiable on Θ with Jacobian $G(\cdot)$, let $\overline{G}(\theta_1, \theta_2) = \int_0^1 G(\omega\theta_1 + (1-\omega)\theta_2)d\omega$. For any $\theta_1, \theta_2 \in \Theta$:*

$$g(\theta_1) - g(\theta_2) = \overline{G}(\theta_1, \theta_2)(\theta_1 - \theta_2).$$

Proof of Lemma B.11: Let $h : [0,1] \to \mathbb{R}^{d_g}$ be defined as $h(\omega) = g(\omega\theta_1 + (1-\omega)\theta_2)$, so that $g(\theta_1) - g(\theta_2) = h(1) - h(0) = \int_0^1 \partial_\omega h(\omega)d\omega$. By composition and the chain rule: $\partial_\omega h(\omega) = \partial_\theta g(\omega\theta_1 + (1-\omega)\theta_2)(\theta_1 - \theta_2) = G(\omega\theta_1 + (1-\omega)\theta_2)(\theta_1 - \theta_2)$. Plug this into the integral to find: $g(\theta_1) - g(\theta_2) = \overline{G}(\theta_1, \theta_2)(\theta_1 - \theta_2)$, as desired. □

B.1.1 Implications of Assumption 1

In the following we will use the notation: $\overline{g}_n(\theta) = 1/n \sum_{i=1}^n g(\theta; x_i)$, $g(\theta) = \mathbb{E}[\overline{g}_n(\theta)]$, $G(\theta; x_i) = \partial_\theta g(\theta; x_i)$, $G_n(\theta) = 1/n \sum_{i=1}^n G(\theta; x_i)$, $G(\theta) = \mathbb{E}[G_n(\theta)]$, $Q_n(\theta) = \overline{g}_n(\theta)' W_n \overline{g}_n(\theta)$, and $Q(\theta) = g(\theta)' W g(\theta)$. W_n and W are symmetric. With probability approaching 1 will be abbreviated as wpa1. $\mathcal{B}_R(\theta^\dagger)$ is a closed ball of radius R, centered around $\theta^\dagger$.

Assumption B.15. *With probability approaching 1: i. Q_n has a unique minimum $\hat{\theta}_n \in$ interior(Θ), ii. $\overline{g}_n$ is twice continuously differentiable, iii. G_n is Lipschitz continuous with constant $L \geq 0$, and for some $R_G > 0$ such that, $\sigma_{\min}[G_n(\theta)] \geq \underline{\sigma} > 0$ for all $\|\theta - \hat{\theta}_n\| \leq R_G$, iv. The parameters space Θ is convex and compact, v. W_n is such that $0 < \underline{\lambda}_W \leq \lambda_{\min}(W_n) \leq \lambda_{\max}(W_n) \leq \overline{\lambda}_W < \infty$.*

Remarks. The condition that x_i are iid can also be weakened to allow for non-identically distributed dependent observations by appropriately adjusting the moment conditions in 1i, iii which are used to derive uniform laws of large numbers for $\overline{g}_n$ and G_n.

Lemma B.12. *Assumption 1 implies Assumption B.15.*

Lemma B.13. *Suppose Assumption 1 holds. Then, for some $r > 0$, Assumption 2 (a) holds for all $\theta \in \mathcal{B}_r(\theta^\dagger)$ with the same choice of ρ, $\underline{\sigma}$.*

The following results are stated in terms of $\overline{G}_n(\theta) = \int_0^1 \{G_n(\omega\theta + (1-\omega)\hat{\theta}_n)\}d\omega$.

Assumption B.16. *With probability approaching 1, for all $\theta \in \Theta$: (a) $\sigma_{\min}[G_n(\theta)'W_n\overline{G}_n(\theta)] \geq \rho\underline{\sigma}$, (b) $\|G_n(\theta)'W_n\overline{G}_n(\theta)(\theta - \theta^\dagger)\| \geq \rho\underline{\sigma}\|\theta - \theta^\dagger\|$.*

Lemma B.14. *Suppose Assumptions 1 holds. 1) If Assumption 2 (a) holds, Assumption B.16 (a) holds. 2) If Assumption 2 (b), Assumption B.16 (b) holds.*

Proof of Lemma B.12. In the following, all the strict inequalities are replaced by weak inequalities with some slackness $\delta > 0$, e.g. $\sigma_{\min}(G(\theta)) \geq (1+\delta)\underline{\sigma} > 0$ instead of $\sigma_{\min}(G(\theta)) > \underline{\sigma} > 0$, and $\lambda_{\max}(W) \leq (1-\delta)\overline{\lambda}_W < \infty$ instead of $\lambda_{\max}(W) < \overline{\lambda}_W < \infty$. Assumption B.15ii, iv follow from 1ii, iv. Use Weyl's perturbation inequality for singular values (Bhatia, 2013, Problem III.6.5) to find $\lambda_{\min}(W_n) \geq \lambda_{\min}(W) - \sigma_{\max}(W_n - W) \geq (1+\delta)\underline{\lambda}_W - o_p(1) \geq \underline{\lambda}_W$, wpa 1. Likewise, $\lambda_{\max}(W_n) \leq \overline{\lambda}_W$, wpa1. This yields Assumption B.15v.

Assumption 1iii and compactness imply uniform convergence of the sample Jacobian $\sup_{\theta\in\Theta} \|G_n(\theta) - G(\theta)\| = o_p(1)$, see Jennrich (1969). We also have uniform convergence for the same moments. Condition ii implies $\overline{g}_n(\theta) - g(\theta) = o_p(1)$, for all θ. Notice that $\|[\overline{g}_n(\theta_1) - g(\theta_1)] - [\overline{g}_n(\theta_2) - g(\theta_2)]\| = \|[\overline{G}_n(\theta_1, \theta_2) -$

$\overline{G}(\theta_1, \theta_2)](\theta_1 - \theta_2)\| \leq [\sup_{\theta \in \Theta} \|G_n(\theta) - G(\theta)\|]\|\theta_1 - \theta_2\|$, where the $\sup$ is a $o_p(1)$ by uniform convergence of G_n, and $\overline{G}(\theta_1, \theta_2) = \int_0^1 G(\omega\theta_1 + (1-\omega)\theta_2)d\omega$. Using a finite cover and arguments similar to Jennrich (1969), this implies uniform convergence: $\sup_{\theta \in \Theta} \|\overline{g}_n(\theta) - g(\theta)\| = o_p(1)$.

Then, uniform convergence of $\overline{g}_n$ and $W_n \xrightarrow{p} W$ imply uniform converge of Q_n to Q. Continuity and the global identification condition 1i. imply $\hat{\theta}_n \xrightarrow{p} \theta^\dagger$ (Newey & McFadden, 1994, Th2.1). This implies that $\|\theta - \hat{\theta}_n\| \leq R_G \Rightarrow \|\theta - \theta^\dagger\| \leq R_G + o_p(1) \leq (1+\delta)R_G$, wpa 1, i.e. $\mathcal{B}_{R_G}(\hat{\theta}_n) \subseteq \mathcal{B}_{(1+\delta)R_G}(\theta^\dagger) \subseteq \Theta$. This implies $\hat{\theta}_n \in$ interior(Θ), wpa1. Then, for the same θ, $\sigma_{\min}[G(\theta)] \geq (1+\delta)\underline{\sigma}$, wpa1. Apply Weyl's inequality for singular values to find that, uniformly in θ: $\sigma_{\min}[G_n(\theta)] \geq \sigma_{\min}[G_n(\theta)] - \sigma_{\max}[G(\theta) - G_n(\theta)] \geq (1+\delta)\underline{\sigma} - o_p(1) \geq \underline{\sigma} > 0$, wpa 1. Take any two θ_1, θ_2 in Θ, $\|G_n(\theta_1) - G_n(\theta_2)\| \leq 1/n \sum_{i=1}^n \bar{L}(x_i)\|\theta_1 - \theta_2\| \leq [(1-\delta)L + o_p(1)]\|\theta_1 - \theta_2\| \leq L\|\theta_1 - \theta_2\|$, wpa1, using a law of large numbers for $\bar{L}(x_i)$. This yields all the conditions in Assumption B.15iii.

□

Proof of Lemma B.13: Under Assumption 1, $\sigma_{\min}[G(\theta)] \geq (1+\delta)\underline{\sigma}$ for all $\theta \in \mathcal{B}_{R_G}(\theta^\dagger)$ and some $\delta > 0$. Also, G is Lipschitz continuous with constant L since $\|G(\theta_1) - G(\theta_2)\| \leq \mathbb{E}[\|G(\theta_1; x_i) - G(\theta_2; x_i)\|] \leq L\|\theta_1 - \theta_2\|$. As a result, $\|\overline{G}(\theta) - G(\theta^\dagger)\| \leq L\|\theta - \theta^\dagger\|$. Then,

$$\|G(\theta)'W\overline{G}(\theta) - G(\theta^\dagger)'WG(\theta^\dagger)\| \leq 2\overline{\sigma}\overline{\lambda}_W L\|\theta - \theta^\dagger\|.$$

Apply Weyl's inequality to find:

$$\sigma_{\min}[G(\theta)'W\overline{G}(\theta)] \geq \left\{(1+\delta)[\underline{\lambda}_W\underline{\sigma}] - 2\frac{\overline{\sigma}\overline{\lambda}_W L}{\underline{\sigma}}\|\theta - \theta^\dagger\|\right\}\underline{\sigma}.$$

Pick $\|\theta - \theta^\dagger\| \leq r$ with r such that $\delta > 2\overline{\sigma} L \overline{\lambda}_W / [\underline{\lambda}_W \underline{\sigma}^2] r$ to find: $\sigma_{\min}[G(\theta)' W \overline{G}(\theta)] > [\underline{\lambda}_W \underline{\sigma}]\underline{\sigma}$, given that $0 < \rho \leq \underline{\lambda}_W \underline{\sigma}$ in Assumption 2 (a), this yields the result. $\square$

Proof of Lemma B.14. Lemma B.12 applies so that Assumption B.15 holds. Hence, G_n is uniformly convergent and Lipschitz continuous, $\hat{\theta}_n$ is consistent. 1) This implies that:

$$\begin{aligned} \|\overline{G}_n(\theta) - \overline{G}(\theta)\| &= \| \int_0^1 \{G_n(\omega\theta + (1-\omega)\hat{\theta}_n) - G(\omega\theta + (1-\omega)\theta^\dagger)\} d\omega \| \\ &\leq L\|\hat{\theta}_n - \theta^\dagger\| + \sup_{\theta \in \Theta} \|G_n(\theta) - G(\theta)\| = o_p(1). \end{aligned}$$

Then apply Weyl's inequality to find that, uniformly in θ and wpa1: $\sigma_{\min}[G_n(\theta)] \geq \sigma_{\min}[G(\theta)] - o_p(1)$, $\sigma_{\min}[\overline{G}_n(\theta)] \geq \sigma_{\min}[\overline{G}(\theta)] - o_p(1)$, and $\sigma_{\min}[G_n(\theta)' W_n \overline{G}_n(\theta)] \geq \sigma_{\min}[G(\theta)' W \overline{G}(\theta)] - o_p(1)$, which yields the result.

2) Lemma B.13 implies Assumption 2 (a) holds locally, i.e. for $\|\theta - \theta^\dagger\| \leq r$, with $r > 0$. With the derivations above, this implies that Assumption B.16 (a) holds locally as well, i.e. for $\|\theta - \hat{\theta}_n\| \leq r/2$, wpa1. Recall that Assumption B.16 (a) implies Assumption B.16 (b).

Take $\|\theta - \hat{\theta}_n\| \geq r/2$. By uniform consistency and boundedness of G_n and $\overline{G}_n$, we have: $G_n(\theta)' W_n \overline{G}_n(\theta) = G(\theta)' W \overline{G}(\theta) + o_p(1)$, uniformly in θ using $\sigma_{\max}[G_n(\theta)] \leq \overline{\sigma}$ wpa1. Since $\hat{\theta}_n$ is consistent, we have uniformly in $\|\theta - \hat{\theta}_n\| \geq r/2$:

$$\begin{aligned} \|G_n(\theta)' W_n \overline{G}_n(\theta)(\theta - \hat{\theta}_n)\| &\geq \|G(\theta)' W \overline{G}(\theta)(\theta - \hat{\theta}_n)\| - o_p(1)\|\theta - \hat{\theta}_n\| \\ &\geq \|G(\theta)' W \overline{G}(\theta)(\theta - \theta^\dagger)\| - o_p(1)\|\theta - \hat{\theta}_n\| - \overline{\sigma}^2 \overline{\lambda}_W o_p(1) \\ &\geq (1+\delta)\rho\underline{\sigma}\|\theta - \theta^\dagger\| - o_p(1)\|\theta - \hat{\theta}_n\| - \overline{\sigma}^2 \overline{\lambda}_W o_p(1) \end{aligned}$$

$$\geq [(1+\delta)\rho\underline{\sigma} - o_p(1)]\|\theta - \hat{\theta}_n\| - [\overline{\sigma}^2\overline{\lambda}_W + (1+\delta)\rho\underline{\sigma}]o_p(1)$$
$$\geq \left[(1+\delta)\rho\underline{\sigma} - o_p(1) - o_p(1)2\frac{\overline{\sigma}^2\overline{\lambda}_W + (1+\delta)\rho\underline{\sigma}}{r}\right]\|\theta - \hat{\theta}_n\|,$$

using $\|\theta - \hat{\theta}_n\|/(r/2) \geq 1$ for the last inequality. The leading term is greater or equal than $\rho\underline{\sigma}$ wpa1 which yields the result. □

B.1.2 Proofs for Section 2.2.1

Proof of Proposition 2 (Gauss-Newton). Take $\theta_k \in \Theta$, the update (2.1) can be re-written as:

$$\begin{aligned}\theta_{k+1} - \hat{\theta}_n =& \Big(I_d - \gamma P_k G_n(\theta_k)' W_n G_n(\theta_k)\Big)(\theta_k - \hat{\theta}_n) \\ &- \gamma P_k G_n(\theta_k)' W_n[\overline{g}_n(\theta_k) - G_n(\theta_k)(\theta_k - \hat{\theta}_n)].\end{aligned} \tag{B.1.1}$$

For GN, $P_k G_n(\theta_k)' W_n G_n(\theta_k) = I_d$ so that we have:

$$\begin{aligned}\theta_{k+1} - \hat{\theta}_n =& (1-\gamma)(\theta_k - \hat{\theta}_n) \\ &- \gamma P_k G_n(\theta_k)' W_n[\overline{g}_n(\theta_k) - \overline{g}_n(\hat{\theta}_n) - G_n(\theta_k)(\theta_k - \hat{\theta}_n)] \\ &- \gamma P_k[G_n(\theta_k) - G_n(\hat{\theta}_n)]' W_n \overline{g}_n(\hat{\theta}_n),\end{aligned} \tag{B.1.1'}$$

using the first-order condition $G_n(\hat{\theta}_n)' W_n \overline{g}_n(\hat{\theta}_n) = 0$. From Assumption B.15, there exists $R_G > 0$ such that: $\underline{\sigma} \leq \sigma_{\min}[G_n(\theta_k)]$ for any $\|\theta_k - \hat{\theta}_n\| \leq R_G$, which implies that P_k is well defined and bounded. Since G_n is Lipschitz continuous with constant $L \geq 0$:

$$\|P_k G_n(\theta_k)' W_n[\overline{g}_n(\theta_k) - \overline{g}_n(\hat{\theta}_n) - G_n(\theta_k)(\theta_k - \hat{\theta}_n)]\| \leq \underline{\sigma}^{-1}\sqrt{\overline{\lambda}_W/\underline{\lambda}_W} L \|\theta_k - \hat{\theta}_n\|^2,$$

We also have:

$$\|P_k[G_n(\theta_k) - G_n(\hat{\theta}_n)]' W_n \overline{g}_n(\hat{\theta}_n)\| \leq \underline{\sigma}^{-2}(\sqrt{\overline{\lambda}_W}/\underline{\lambda}_W) L \|\overline{g}_n(\hat{\theta}_n)\|_{W_n} \|\theta_k - \hat{\theta}_n\|.$$

Combine these two inequalities into (B.1.1′) to find:

$$\begin{aligned}&\|\theta_{k+1} - \hat{\theta}_n\| \leq \\ &\left(1 - \gamma + \gamma \left[\underline{\sigma}^{-1}\sqrt{\overline{\lambda}_W/\underline{\lambda}_W} L \|\theta_k - \hat{\theta}_n\| + \underline{\sigma}^{-2}(\sqrt{\overline{\lambda}_W}/\underline{\lambda}_W) L \|\overline{g}_n(\hat{\theta}_n)\|_{W_n}\right]\right) \|\theta_k - \hat{\theta}_n\|.\end{aligned} \tag{B.1.1″}$$

Now take any $\tilde{\gamma} \in (0, \gamma)$, let:

$$\tilde{R}_n = \frac{\gamma - \tilde{\gamma}}{\gamma} \left[L^{-1} \underline{\sigma} \sqrt{\underline{\lambda}_W/\overline{\lambda}_W}\right] - (\underline{\sigma}^{-1}/\sqrt{\underline{\lambda}_W}) \|\overline{g}_n(\hat{\theta}_n)\|_{W_n}.$$

Let $R_n = \min(\tilde{R}_n, R_G)$, for any $\|\theta_k - \hat{\theta}_n\| \leq R_n$, we have $\|\theta_{k+1} - \hat{\theta}_n\| \leq (1 - \tilde{\gamma})\|\theta_k - \hat{\theta}_n\| \leq R_n$. By recursion, we then have for any $\|\theta_0 - \hat{\theta}_n\| \leq R_n$:

$$\|\theta_{k+1} - \hat{\theta}_n\| \leq (1 - \tilde{\gamma})\|\theta_k - \hat{\theta}_n\| \leq \cdots \leq (1 - \tilde{\gamma})^{k+1}\|\theta_0 - \hat{\theta}_n\|,$$

as stated in (2.2). □

Proof of Proposition 2 (General Case). Take $\theta_k \in \Theta$, the update (2.1) can be re-written as:

$$\begin{aligned}\theta_{k+1} - \hat{\theta}_n =& \Big(I_d - \gamma P_k G_n(\theta_k)' W_n G_n(\theta_k)\Big)(\theta_k - \hat{\theta}_n) \\ &- \gamma P_k G_n(\theta_k)' W_n [\overline{g}_n(\theta_k) - G_n(\theta_k)(\theta_k - \hat{\theta}_n)].\end{aligned} \tag{B.1.1}$$

Taking norms on both sides this identity yields:

$$\begin{aligned}\|\theta_{b+1} - \hat{\theta}_n\| \leq & \sigma_{\max}\Big[I_d - \gamma P_k G_n(\theta_k)' W_n G_n(\theta_k)\Big] \|\theta_b - \hat{\theta}_n\| \\ & + \gamma \|P_k G_n(\theta_k)' W_n [\overline{g}_n(\theta_k) - G_n(\theta_k)(\theta_k - \hat{\theta}_n)]\|,\end{aligned} \tag{B.1.1$'$}$$

where $\sigma_{\max}$ returns the largest singular value. We will now bound each of these two terms. First, note that $\sigma_{\max}[I_d - \gamma P_k G_n(\theta_k)' W_n G_n(\theta_k)] = \sigma_{\max}[I_d - \gamma P_k^{1/2} G_n(\theta_k)' W_n G_n(\theta_k) P_k^{1/2}] = \max_{j=1,\dots,d} |\lambda_j[I_d - \gamma P_k^{1/2} G_n(\theta_k)' W_n G_n(\theta_k) P_k^{1/2}]|$, where λ_j are the eigenvalues. Because this is a difference of Hermitian matrices, Weyl's perturbation inequality (Bhatia, 2013, Corollary III.2.2) implies the following bounds:

$$\begin{aligned}1 - \gamma \lambda_{\max}[P_k^{1/2} G_n(\theta_k)' W_n G_n(\theta_k) P_k^{1/2}] &\leq \lambda_{\min}[I_d - \gamma P_k^{1/2} G_n(\theta_k)' W_n G_n(\theta_k) P_k^{1/2}] \\ &\leq \lambda_{\max}[I_d - \gamma P_k^{1/2} G_n(\theta_k)' W_n G_n(\theta_k) P_k^{1/2}] \\ &\leq 1 - \gamma \lambda_{\min}[P_k^{1/2} G_n(\theta_k)' W_n G_n(\theta_k) P_k^{1/2}].\end{aligned}$$

Let $\overline{\sigma} = \max_{\theta \in \Theta} \sigma_{\max}[G_n(\theta)]$, suppose $0 < \gamma < [\overline{\lambda}_P \overline{\lambda}_W \overline{\sigma}^2]^{-1}$, we then have:

$$0 \leq 1 - \gamma \lambda_{\max}[P_k^{1/2} G_n(\theta_k)' W_n G_n(\theta_k) P_k^{1/2}] \leq 1 - \gamma \lambda_{\min}[P_k^{1/2} G_n(\theta_k)' W_n G_n(\theta_k) P_k^{1/2}],$$

so that we are only concerned with the upper bound. From Assumption B.15, $\|\theta - \hat{\theta}_n\| \leq R_G \Rightarrow \sigma_{\min}[G_n(\theta)] \geq \underline{\sigma}$. Combine with the bound for γ to find:

$$0 \leq \sigma_{\max}[I_d - \gamma P_k G_n(\theta_k)' W_n G_n(\theta_k)] \leq 1 - \gamma \underline{\lambda}_P \underline{\lambda}_W \underline{\sigma}^2 < 1,$$

for any choice of $\gamma \in (0, [\overline{\lambda}_P \overline{\lambda}_W \overline{\sigma}^2]^{-1})$. For the second term in (B.1.1), using the identity $G_n(\hat{\theta}_n)' W_n \overline{g}_n(\hat{\theta}_n) = 0$ and Lemma B.11:

$$\begin{aligned} P_k G_n(\theta_k)' W_n [\overline{g}_n(\theta_k) - G_n(\theta_k)(\theta_k - \hat{\theta}_n)] =& P_k G_n(\theta_k)' W_n [\overline{G}_n(\theta_k) - G_n(\theta_k)](\theta_k - \hat{\theta}_n) \\ &+ P_k [G_n(\theta_k) - G_n(\hat{\theta}_n)]' W_n \overline{g}_n(\hat{\theta}_n), \end{aligned}$$

where $\overline{G}_n(\theta_k) = \int_0^1 \{G_n(\omega\theta_k + (1-\omega)\hat{\theta}_n)\} d\omega$. Since G_n is Lipschitz continuous with constant $L \geq 0$:

$$\begin{aligned} \|(B.1.1')\| \leq& (1 - \gamma \underline{\lambda}_P \underline{\lambda}_W \underline{\sigma}^2) \|\theta_b - \hat{\theta}_n\| + \gamma \overline{\lambda}_P \overline{\lambda}_W \overline{\sigma} L \|\theta_b - \hat{\theta}_n\|^2 \\ &+ \gamma \overline{\lambda}_P \overline{\lambda}_W^{1/2} L \|\overline{g}_n(\hat{\theta}_n)\|_{W_n} \|\theta_b - \hat{\theta}_n\| \\ &= \Big(1 - \gamma \underline{\lambda}_P \underline{\lambda}_W \underline{\sigma}^2 + \gamma \Big[\overline{\lambda}_P \overline{\lambda}_W \overline{\sigma} L \|\theta_b - \hat{\theta}_n\| \\ &+ \overline{\lambda}_P \overline{\lambda}_W^{1/2} L \|\overline{g}_n(\hat{\theta}_n)\|_{W_n} \Big] \Big) \|\theta_b - \hat{\theta}_n\|. \end{aligned}$$

Let $c_1 = \overline{\lambda}_P \overline{\lambda}_W \overline{\sigma} L$, $c_2 = \overline{\lambda}_P \overline{\lambda}_W^{1/2} L$, pick $\tilde{\gamma} \in (0, \gamma \underline{\lambda}_P \underline{\lambda}_W \underline{\sigma}^2)$, and assume:

$$\|\theta_k - \hat{\theta}_n\| \leq \frac{\gamma \underline{\lambda}_P \underline{\lambda}_W \underline{\sigma}^2 - \tilde{\gamma}}{\gamma c_1} - \frac{c_2}{c_1} \|\overline{g}_n(\hat{\theta}_n)\|_{W_n} := \tilde{R}_n. \tag{B.1.2}$$

Take $R_n = \min(R_G, \tilde{R}_n)$, $\|\theta_k - \hat{\theta}_n\| \leq R_n$ implies that, by construction:

$$\|\theta_{k+1} - \hat{\theta}_n\| \leq (1 - \tilde{\gamma}) \|\theta_k - \hat{\theta}_n\| \leq \cdots \leq (1 - \overline{\gamma})^{k+1} \|\theta_0 - \hat{\theta}_n\|,$$

by recursion, if $\|\theta_0 - \hat{\theta}_n\| \leq R_n$. □

Proof of Theorem 2 In the just-identified case, we will repeatedly use the identities $\overline{g}_n(\hat{\theta}_n) = 0$ and $Q_n(\hat{\theta}_n) = \frac{1}{2} \overline{g}_n(\hat{\theta}_n)' W_n \overline{g}_n(\hat{\theta}_n) = 0$. Take any $\theta \in \Theta$, by Lemma

B.11, $\overline{g}_n(\theta) = \overline{G}_n(\theta)(\theta - \hat{\theta}_n)$ which implies:

$$\left(\frac{1}{2}\min_{\theta\in\theta}\lambda_{\min}[\overline{G}_n(\theta)'W_n\overline{G}_n(\theta)]\right)\|\theta - \hat{\theta}_n\|^2 \leq Q_n(\theta)$$
$$\leq \left(\frac{1}{2}\max_{\theta\in\theta}\lambda_{\max}[\overline{G}_n(\theta)'W_n\overline{G}_n(\theta)]\right)\|\theta - \hat{\theta}_n\|^2,$$

and, using the full rank assumption, let $0 < \underline{\lambda} \leq \overline{\lambda} < \infty$, denote respectively the $\min$ and the $\max$. This yields an equivalence between the two distances $\|\theta - \hat{\theta}_n\|$ and $Q_n(\theta)$. Apply the Mean Value Theorem to Q_n, for some $\tilde{\theta}_k$ between θ_k and θ_{k+1} in (2.1):

$$Q_n(\theta_{k+1}) = Q_n(\theta_k) + \partial_\theta Q_n(\theta_k)'(\theta_{k+1} - \theta_k) + [\partial_\theta Q_n(\tilde{\theta}_k) - \partial_\theta Q_n(\theta_k)]'(\theta_{k+1} - \theta_k), \tag{B.1.3}$$

where $\partial_\theta Q_n(\theta_k) = G_n(\theta_k)'W_n\overline{g}_n(\theta_k)$ and $\theta_{k+1} - \theta_k = -\gamma P_k G_n(\theta_k)'W_n\overline{g}_n(\theta_k)$. This yields a first inequality:

$$\partial_\theta Q_n(\theta_k)'(\theta_{k+1} - \theta_k) = -\gamma\overline{g}_n(\theta_k)'W_n^{1/2}\left(W_n^{1/2}G_n(\theta_k)P_kG_n(\theta_k)'W_n^{1/2}\right)W_n^{1/2}\overline{g}_n(\theta_k)$$
$$\leq -\gamma c_1\|\overline{g}_n(\theta_k)\|^2_{W_n},$$

where $c_1 := \min_{\theta\in\Theta}\lambda_{\min}[W_n^{1/2}G_n(\theta)P_kG_n(\theta)'W_n^{1/2}] > 0$, using the full rank assumption, and $\|\overline{g}_n(\theta_k)\|^2_{W_n} = Q_n(\theta_k)$. For the second inequality, since $\overline{g}_n$ is twice continuously differentiable and Θ is compact, $\partial_\theta Q_n$ is Lipschitz continuous with constant $L_Q \geq 0$. This implies:

$$\|[\partial_\theta Q_n(\tilde{\theta}_k) - \partial_\theta Q_n(\theta_k)]'(\theta_{k+1} - \theta_k)\| \leq L_Q\|\tilde{\theta}_k - \theta_k\| \times \|\theta_{k+1} - \theta_k\| \leq L_Q\|\theta_{k+1} - \theta_k\|^2$$
$$\leq \gamma^2 c_2\|\overline{g}_n(\theta_k)\|^2_{W_n},$$

since $\|\theta_{k+1} - \theta_k\| = \gamma\|P_k G_n(\theta_k)' W_n \overline{g}_n(\theta_k)\|$ and $\|\tilde{\theta}_k - \theta_k\| \leq \|\theta_{k+1} - \theta_k\|$, setting $c_2 := L_Q \max_{\theta\in\Theta} \sigma^2_{\max}[P_k G_n(\theta)' W_n^{1/2}] < +\infty$. Combine the two inequalities into (B.1.3) to find:

$$Q_n(\theta_{k+1}) \leq (1 - \gamma c_1 + \gamma^2 c_2) Q_n(\theta_k),$$

for any $\theta_k \in \Theta$. The polynomial $P(\gamma) = 1 - \gamma c_1 + \gamma^2 c_2$ is such that $P(0) = 1, d_\gamma P(0) < 0$ which implies $P(\gamma) \in (0,1)$ strictly for any $\gamma > 0$ sufficiently small. Take any such γ and let $(1-\overline{\gamma})^2 = P(\gamma) \in (0,1)$ for some $\overline{\gamma} \in (0,1)$, by construction. Take any $\theta_0 \in \Theta$, by recursion:

$$Q_n(\theta_{k+1}) \leq (1-\overline{\gamma})^2 Q_n(\theta_k) \leq \cdots \leq (1-\overline{\gamma})^{2(k+1)} Q_n(\theta_0).$$

Now apply the distance equivalence derived earlier to get the desired result:

$$\|\theta_{k+1} - \hat{\theta}_n\| \leq (1-\overline{\gamma})^{k+1} \sqrt{\overline{\lambda}/\underline{\lambda}} \|\theta_0 - \hat{\theta}_n\|.$$

□

Proof of Theorem 3. The general layout of the proof is similar to the just-identified case. Differences arise because $Q_n(\hat{\theta}_n) \neq 0$ in general and $G_n(\theta)$ only has rank d_θ which is less than the dimension of $\overline{g}_n$ so that several parts of the proof do not apply anymore. First:

$$Q_n(\theta) - Q_n(\hat{\theta}_n) = \frac{1}{2}\left[\overline{g}_n(\theta) - \overline{g}_n(\hat{\theta}_n)\right]' W_n \left[\overline{g}_n(\theta) - \overline{g}_n(\hat{\theta}_n)\right] + \left[\overline{g}_n(\theta) - \overline{g}_n(\hat{\theta}_n)\right]' W_n \overline{g}_n(\hat{\theta}_n)$$

The leading term equals $\frac{1}{2}(\theta - \hat{\theta}_n)' \overline{G}_n(\theta)' W_n \overline{G}_n(\theta)(\theta - \hat{\theta}_n)$ which can be bounded above and below using the same approach as before. For the last term, use the first-order condition $G_n(\hat{\theta}_n)'W_n\overline{g}_n(\hat{\theta}_n) = 0$ to get, using $\overline{g}_n(\theta) - \overline{g}_n(\hat{\theta}_n) = [\overline{G}_n(\theta) - G_n(\hat{\theta}_n) + G_n(\hat{\theta}_n)](\theta - \hat{\theta}_n)$:

$$\begin{aligned} \|[\overline{g}_n(\theta) - \overline{g}_n(\hat{\theta}_n)]'W_n\overline{g}_n(\hat{\theta}_n)\| & \\ &\leq \sqrt{\lambda_{\max}(W_n)}\|\theta - \hat{\theta}_n\| \times \|G_n(\hat{\theta}_n) - \overline{G}_n(\theta)\| \times \|\overline{g}_n(\hat{\theta}_n)\|_{W_n} \\ &\leq \sqrt{\lambda_{\max}(W_n)}L\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}\|\theta - \hat{\theta}_n\|^2, \end{aligned}$$

where $L \geq 0$ is the Lipschitz constant of G_n. Let $0 < \underline{\lambda} = \min_{\theta\in\Theta}\lambda_{\min}[\overline{G}_n(\theta)'W_n\overline{G}_n(\theta)] \leq \overline{\lambda} = \max_{\theta\in\Theta}\lambda_{\max}[\overline{G}_n(\theta)'W_n\overline{G}_n(\theta)] < \infty$, apply the triangular inequality and its reverse to find the relation:

$$\left(\underline{\lambda} - C\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}\right)\|\theta - \hat{\theta}_n\|^2 \leq 2[Q_n(\theta) - Q_n(\hat{\theta}_n)] \leq \left(\overline{\lambda} + C\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}\right)\|\theta - \hat{\theta}_n\|^2, \tag{B.1.4}$$

where $C := 2\sqrt{\lambda_{\max}(W_n)}L \geq 0$ is finite. As in the just-identified case, we can write:

$$Q_n(\theta_{k+1}) = Q_n(\theta_k) + \partial_\theta Q_n(\theta_k)'(\theta_{k+1} - \theta_k) + [\partial_\theta Q_n(\tilde{\theta}_k) - \partial_\theta Q_n(\theta_k)]'(\theta_{k+1} - \theta_k),$$

and bound each of the last two terms. As before, we have:

$$\partial_\theta Q_n(\theta_k)'(\theta_{k+1} - \theta_k) = -\gamma\overline{g}_n(\theta_k)'W_nG_n(\theta_k)P_kG_n(\theta_k)'W_n\overline{g}_n(\theta_k),$$

However, $\dim(W_n^{1/2}\overline{g}_n(\theta_k)) > \text{rank}[W_n^{1/2}G_n(\theta_k)P_kG_n(\theta_k)'W_n^{1/2}]$, the model being over-identified. Hence, we only have $\partial_\theta Q_n(\theta_k)'(\theta_{k+1} - \theta_k) \leq 0$ which, unlike the just-identified case, does not imply a strict contraction. Nevertheless, we have

$\overline{g}_n(\theta_k) = \overline{G}_n(\theta_k)(\theta_k - \hat{\theta}_n) + \overline{g}_n(\hat{\theta}_n)$ where $\dim(\theta_k - \hat{\theta}_n)$ equals the rank of the matrix above. Let $A_k = W_n^{1/2} G_n(\theta_k) P_k G_n(\theta_k)' W_n^{1/2}$:

$$\begin{aligned}
&-\gamma \overline{g}_n(\theta_k)' W_n G_n(\theta_k) P_k G_n(\theta_k)' W_n \overline{g}_n(\theta_k) \\
&\qquad = -\gamma \left[\overline{g}_n(\theta_k) - \overline{g}_n(\hat{\theta}_n)\right]' W_n^{1/2} A_k W_n^{1/2} \left[\overline{g}_n(\theta_k) - \overline{g}_n(\hat{\theta}_n)\right] && \text{(B.1.5)} \\
&\qquad\quad - \gamma \overline{g}_n(\hat{\theta}_n)' W_n^{1/2} A_k W_n^{1/2} \overline{g}_n(\hat{\theta}_n) && \text{(B.1.6)} \\
&\qquad\quad - 2\gamma \overline{g}_n(\hat{\theta}_n)' W_n^{1/2} A_k W_n^{1/2} \left[\overline{g}_n(\theta_k) - \overline{g}_n(\hat{\theta}_n)\right]. && \text{(B.1.7)}
\end{aligned}$$

Now, bound these terms one at a time:

$$\begin{aligned}
(B.1.5) &= -\gamma(\theta_k - \hat{\theta}_n)' \overline{G}_n(\theta_k)' W_n^{1/2} A_k W_n^{1/2} \overline{G}_n(\theta_k)(\theta_k - \hat{\theta}_n) \\
&\leq -\gamma c_{1n} [Q_n(\theta_k) - Q_n(\hat{\theta}_n)],
\end{aligned}$$

where the last inequality comes from (B.1.4) above, and

$$c_{1n} := \min_{\theta_k \in \Theta} \lambda_{\min}[\overline{G}_n(\theta_k)' W_n^{1/2} A_k W_n^{1/2} \overline{G}_n(\theta_k)] (\overline{\lambda}/2 + C/2 \|\overline{g}_n(\hat{\theta}_n)\|_{W_n})^{-1} > 0,$$

is bounded below by a strictly positive value with probability approaching 1, using Assumption 3 and Lemma B.14.[1] To see why Lemma B.14 (ii) is critical, notice that $\overline{G}_n(\theta_k)' W_n^{1/2} A_k W_n^{1/2} \overline{G}_n(\theta_k) = \overline{G}_n(\theta_k)' W_n G_n(\theta_k) P_k G_n(\theta_k)' W_n \overline{G}_n(\theta_k)$ is symmetric and has full rank if both $G_n(\theta_k)' W_n \overline{G}_n(\theta_k)$ and P_k have full rank. Both Lemma B.14 (ii) and Assumption 3 need to hold for c_{1n} to be non-zero. Then $(B.1.6) \leq -\gamma Q_n(\hat{\theta}_n) \lambda_{\min}(A_k) \leq 0$. For the remaining term, apply the Cauchy-

---

[1] An explicit lower bound is given in the proof of Theorem 4.

Schwarz inequality, Lemma B.11, and (B.1.4) to find the last bound:

$$\|(B.1.7)\| \leq 2\gamma\sqrt{Q_n(\hat{\theta}_n)}\frac{\max_{\theta\in\Theta}\sigma_{\max}[A_kW_k^{1/2}\overline{G}_n(\theta)]}{[\underline{\lambda}/2 - C/2\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}]^{1/2}}[Q_n(\theta_k) - Q_n(\hat{\theta}_n)]^{1/2}$$

Let $c_{3n} := 2\max_{\theta\in\Theta}\sigma_{\max}[A_kW_k^{1/2}\overline{G}_n(\theta)][\underline{\lambda}/2 - C/2\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}]^{-1/2}$. As in the just-identified case, $\partial_\theta Q_n$ is Lipschitz continuous with constant $L_Q \geq 0$, which yields the same inequality as in the proof of Theorem 2:

$$\|[\partial_\theta Q_n(\tilde{\theta}_k) - \partial_\theta Q_n(\theta_k)]'(\theta_{k+1} - \theta_k)\| \leq \gamma^2 L_Q \max_{\theta\in\Theta}\sigma^2_{\max}[P_kG_n(\theta)'W_n^{1/2}]Q_n(\theta_k).$$

Let $c_2 := L_Q\max_{\theta\in\Theta}\sigma^2_{\max}[P_kG_n(\theta)'W_n^{1/2}] \geq 0$. Combine all the inequalities above to get:

$$\begin{aligned} Q_n(\theta_{k+1}) - Q_n(\hat{\theta}_n) &\leq (1 - \gamma c_{1n} + \gamma^2 c_2)[Q_n(\theta_k) - Q_n(\hat{\theta}_n)] \\ &\quad + \gamma^2 c_2 Q_n(\hat{\theta}_n) + \gamma c_{3n}\sqrt{Q_n(\hat{\theta}_n)}\left[Q_n(\theta_k) - Q_n(\hat{\theta}_n)\right]^{1/2}. \end{aligned}$$

Because of the square root on $Q_n(\theta_k) - Q_n(\hat{\theta}_n)$, this is a non-linear recursion. To derive explicit convergence results, we will bound it by a linear recursion using:

i. If $[Q_n(\theta_k) - Q_n(\hat{\theta}_n)]^{1/2} \geq 2c_{3n}/c_{1n}\sqrt{Q_n(\hat{\theta}_n)}$, then:

$$Q_n(\theta_{k+1}) - Q_n(\hat{\theta}_n) \leq (1 - \gamma\frac{c_{1n}}{2} + \gamma^2 c_2)[Q_n(\theta_k) - Q_n(\hat{\theta}_n)] + \gamma^2 c_2 Q_n(\hat{\theta}_n).$$

ii. Otherwise:

$$Q_n(\theta_{k+1}) - Q_n(\hat{\theta}_n) \leq (1 - \gamma c_{1n} + \gamma^2 c_2)[Q_n(\theta_k) - Q_n(\hat{\theta}_n)] + \left(\gamma^2 c_2 + 2\gamma\frac{c_{3n}^2}{c_{1n}}\right)Q_n(\hat{\theta}_n).$$

A majorization of these two bounds implies:

$$Q_n(\theta_{k+1}) - Q_n(\hat{\theta}_n) \tag{B.1.8}$$

$$\leq \left(1 - \gamma\frac{c_{1n}}{2} + \gamma^2 c_2\right)\left[Q_n(\theta_k) - Q_n(\hat{\theta}_n)\right] + \left(\gamma^2 c_2 + 2\gamma\frac{c_{3n}^2}{c_{1n}}\right) Q_n(\hat{\theta}_n). \tag{B.1.9}$$

Let $P_n(\gamma) = (1 - \gamma\frac{c_{1n}}{2} + \gamma^2 c_2)$. Then, using the same arguments as in the just-identified case, for $\gamma > 0$ sufficiently small, we have $P_n(\gamma) = (1 - \overline{\gamma})^2 \in (0, 1)$, i.e. (2.1) is a strict contraction globally. Iterate on the recursion to find:

$$Q_n(\theta_{k+1}) - Q_n(\hat{\theta}_n) \leq (1 - \overline{\gamma})^{2(k+1)}[Q_n(\theta_0) - Q_n(\hat{\theta}_n)] + \frac{\gamma^2 c_2 + 2\gamma\frac{c_{3n}^2}{c_{1n}}}{1 - (1 - \overline{\gamma})^2} Q_n(\hat{\theta}_n).$$

Apply the distance equivalence to find:

$$\begin{aligned}\|\theta_{k+1} - \hat{\theta}_n\|^2 \leq & (1 - \overline{\gamma})^{2(k+1)} \left(\underline{\lambda}/2 - C/2\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}\right)^{-1} [Q_n(\theta_0) - Q_n(\hat{\theta}_n)] \\ & + \left(\underline{\lambda}/2 - C/2\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}\right)^{-1} \frac{\gamma^2 c_2 + 2\gamma\frac{c_{3n}^2}{c_{1n}}}{1 - (1 - \overline{\gamma})^2} Q_n(\hat{\theta}_n).\end{aligned}$$

For the choice of $\gamma > 0$ which yields the result, there exists a $R_n > 0$ for which Proposition 2 holds. Then in large samples we have:

$$\left(\underline{\lambda}/2 - C/2\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}\right)^{-1} \frac{\gamma^2 c_2 + 2\gamma\frac{c_{3n}^2}{c_{1n}}}{1 - (1 - \overline{\gamma})^2} Q_n(\hat{\theta}_n) \leq R_n^2/2,$$

with increasing probability. For k large enough:

$$(1 - \overline{\gamma})^{2(k)} \left(\underline{\lambda}/2 - C/2\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}\right)^{-1} [Q_n(\theta_0) - Q_n(\hat{\theta}_n)] \leq R_n^2/2,$$

as well.[2] Then, with increasing probability, for this choice of k we have:

$$\|\theta_k - \hat{\theta}_n\| \leq R_n,$$

apply Proposition 2 for another $j \geq 0$ iterations to find:

$$\|\theta_{k+j} - \hat{\theta}_n\| \leq (1 - \tilde{\gamma})^j R_n,$$

where $\tilde{\gamma}$ is the convergence rate in Proposition 2 which need not the same as the global rate derived above. This concludes the proof. $\square$

B.1.3 Proofs for Section 2.2.2

Proof of Proposition 3 (Gauss-Newton): Following the proof of Proposition 2:

$$\begin{aligned} &\|\theta_{k+1} - \hat{\theta}_n\| / \|\theta_k - \hat{\theta}_n\| \\ &\leq 1 - \gamma + \gamma \left[\underline{\sigma}^{-1} \sqrt{\overline{\lambda}_W / \underline{\lambda}_W} L \|\theta_k - \hat{\theta}_n\| + \underline{\sigma}^{-2} (\sqrt{\overline{\lambda}_W} / \underline{\lambda}_W) L \|\overline{g}_n(\hat{\theta}_n)\|_{W_n} \right] \end{aligned} \tag{B.1.1''}$$

Take $\tilde{\gamma} \in (0, \gamma)$, and $\tilde{R}_n$ such that:

$$\tilde{R}_n = \frac{\gamma - \tilde{\gamma}}{\gamma} \left[L^{-1} \underline{\sigma} \sqrt{\underline{\lambda}_W / \overline{\lambda}_W} \right] - (\underline{\sigma}^{-1} / \sqrt{\underline{\lambda}_W}) \|\overline{g}_n(\hat{\theta}_n)\|_{W_n}.$$

We have

$$\begin{aligned} \text{plim}_{n \to \infty} \tilde{R}_n = \tilde{R} = \frac{\gamma - \tilde{\gamma}}{\gamma} [L^{-1} \underline{\sigma} \sqrt{\underline{\lambda}_W / \overline{\lambda}_W}] - (\underline{\sigma}^{-1} / \sqrt{\underline{\lambda}_W}) \varphi > 0 \\ \Leftrightarrow \varphi < [1 - \tilde{\gamma}/\gamma] \frac{\underline{\sigma}^2 \underline{\lambda}_W}{L \sqrt{\overline{\lambda}_W}} \end{aligned}$$

---

[2] Let $d_{0n} = [\underline{\lambda}/2 - C/2 \|\overline{g}_n(\hat{\theta}_n)\|_{W_n}]^{-1} [Q_n(\theta_0) - Q_n(\hat{\theta}_n)]$, pick $k \geq \frac{2 \log R_n - \log 2 - \log d_{0n}}{2 \log(1 - \overline{\gamma})}$.

Under the stated Assumptions, for any $\varphi \geq 0$ such that $\varphi < \frac{\underline{\sigma}^2\underline{\lambda}_W}{L\sqrt{\overline{\lambda}_W}}$, there exists $\tilde{\gamma} \in (0,\gamma)$, sufficiently small such that the above strict inequality holds. Then, $\tilde{R}_n \geq (1-\varepsilon)\tilde{R} > 0$ with probability approaching 1 for any $\varepsilon \in (0,1)$. Let $R_n = \min(\tilde{R}_n, R_G)$, take $\|\theta_0 - \hat{\theta}_n\| \leq R_n$, by recursion:

$$\|\theta_{k+1} - \hat{\theta}_n\| \leq (1-\tilde{\gamma})\|\theta_k - \hat{\theta}_n\| \leq \cdots \leq (1-\tilde{\gamma})^{k+1}\|\theta_0 - \hat{\theta}_n\|,$$

with probability approaching 1, for all $k \geq 0$. This is the desired result. □

Proof of Theorem 4: The layout of the proof closely follows that of Theorem 3. Recall inequality (B.1.4):

$$\left(\underline{\lambda} - C\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}\right)\|\theta - \hat{\theta}_n\|^2 \leq 2[Q_n(\theta) - Q_n(\hat{\theta}_n)] \leq \left(\overline{\lambda} + C\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}\right)\|\theta - \hat{\theta}_n\|^2,$$

where $0 < \underline{\lambda} = \min_{\theta\in\Theta}\lambda_{\min}[\overline{G}_n(\theta)'W_n\overline{G}_n(\theta)] \leq \overline{\lambda} = \max_{\theta\in\Theta}\lambda_{\max}[\overline{G}_n(\theta)'W_n\overline{G}_n(\theta)] < \infty$, and $C := 2\sqrt{\lambda_{\max}(W_n)}L \geq 0$ are finite. Condition (2.5′) implies $0 < \underline{\lambda} - C\varphi$ since $\underline{\lambda} \geq \underline{\sigma}^2\underline{\lambda}_W$ when Assumption 2 (b) holds. Then, for any $\delta \in (0,1)$, we have $(\underline{\lambda} - C\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}) \geq (1-\delta)[\underline{\lambda} - C\varphi] > 0$, with probability approaching 1 (wpa1). This implies that the norm equivalence holds and is informative, with high probability, in large samples. Now recall inequality (B.1.9):

$$\begin{aligned} Q_n(\theta_{k+1}) - & Q_n(\hat{\theta}_n) \\ & \leq \left(1 - \gamma\frac{c_{1n}}{2} + \gamma^2 c_2\right)\left[Q_n(\theta_k) - Q_n(\hat{\theta}_n)\right] + \left(\gamma^2 c_2 + 2\gamma\frac{c_{3n}^2}{c_{1n}}\right)Q_n(\hat{\theta}_n), \end{aligned}$$

where:

$$c_{1n} = \min_{\theta_k\in\Theta}\lambda_{\min}[\overline{G}_n(\theta_k)'W_n^{1/2}A_kW_n^{1/2}\overline{G}_n(\theta_k)](\overline{\lambda}/2 + C/2\|\overline{g}_n(\hat{\theta}_n)\|_{W_n})^{-1},$$

$$
\begin{aligned}
c_2 &= L_Q \max_{\theta\in\Theta} \sigma^2_{\max}[P_k G_n(\theta)' W_n^{1/2}],\\
c_{3n} &= 2\max_{\theta\in\Theta} \sigma_{\max}[A_k W_n^{1/2} G_n(\theta)][\underline{\lambda}/2 - C/2\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}]^{-1/2},
\end{aligned}
$$

L_Q is the Lipschitz constant of $\partial_\theta Q_n$ and $A_k = W_n^{1/2} G_n(\theta_k) P_k G_n(\theta_k)' W_n^{1/2}$ is an idempotent matrix for GN. Together, Assumption 2 (b) and (2.5′) imply the following upper and lower bounds holds wpa1:

$$
0 < \underline{c}_1 := \frac{2}{3}\rho^2([\underline{\sigma}/\overline{\sigma}]^2 \kappa_W^{-1})^2 \leq c_{1n} \leq 2[\overline{\sigma}/\underline{\sigma}]^2 \kappa_W := \overline{c}_1 < \infty,
$$

where $\overline{\sigma} \geq \underline{\sigma}$ is such that $\sigma_{\max}[G_n(\theta)] \leq \overline{\sigma}$ for all $\theta \in \Theta$ and $\kappa_W = \overline{\lambda}_W/\underline{\lambda}_W$. The upper bound relies on $\sigma_{\max}(A_k) = 1$ so the numerator is less than $\overline{\sigma}^2\overline{\lambda}_W$, while for the denominator $\|\overline{g}_n(\hat{\theta}_n)\|_{W_n} \geq 0$ and $\overline{\lambda} \geq \underline{\sigma}^2\underline{\lambda}_W$. For the lower bound, Assumption 2 (b) implies the numerator is greater than $\inf_{\theta_k\in\Theta} \sigma_{\min}[\overline{G}_n(\theta)' W_n G_n(\theta_k)]^2 \lambda_{\min}(P_k) \geq [\underline{\sigma}^2\underline{\lambda}_W]^2[\overline{\sigma}^2\overline{\lambda}_W]^{-1}$. For the denominator of the lower bound, notice that $0 \leq \varphi < \underline{\lambda}/C$ implies $C\|\overline{g}_n(\hat{\theta}_n)\|_{W_n} \leq 2C\varphi \leq 2\underline{\lambda} \leq 2\overline{\lambda}$, wpa 1, which – with a bound on $\overline{\lambda}$ – yields the resulting bound $\underline{c}_1$. Also, wpa1:

$$
0 \leq c_{3n} \leq \overline{c}_3[\underline{\lambda}/2 - C/2\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}]^{-1/2}.
$$

where $\overline{c}_3 = 2\overline{\sigma}\overline{\lambda}_W^{-1/2}$ since $\sigma_{\max}(A_k) = 1$ for GN. Combine these bounds to find that, wpa1 and uniformly in θ_k we have:[3]

$$
\begin{aligned}
Q_n(\theta_{k+1}) - Q_n(\hat{\theta}_n)& \\
&\leq \left(1 - \gamma\frac{c_1}{2} + \gamma^2 c_2\right)\left[Q_n(\theta_k) - Q_n(\hat{\theta}_n)\right] + \left(\gamma^2 c_2 + 2\gamma\frac{c_{3n}^2}{c_{1n}}\right) Q_n(\hat{\theta}_n),
\end{aligned}
$$

---

[3] The inequality is uniform in θ_k because the bound involves the same event on $\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}$ for all θ_k.

which does not depend on φ. For $\gamma \in (0,1)$ small enough, pick $\overline{\gamma} \in (0,1)$ such that $(1 - \gamma\frac{\underline{c}_1}{2} + \gamma^2 c_2) = (1-\overline{\gamma})^2$ and

$$C_n = \frac{\gamma^2 c_2 + 2\gamma c_{3n}^2/c_{1n}}{[1-(1-\overline{\gamma})^2][\underline{\lambda}/2 - C/2\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}]} = O_p(1),$$

as in Theorem 3. Then we have:

$$\begin{aligned} Q_n(\theta_{k+1}) - & Q_n(\hat{\theta}_n) \\ & \leq (1-\overline{\gamma})^{2(k+1)} \left[Q_n(\theta_0) - Q_n(\hat{\theta}_n) \right] + C_n[\underline{\lambda}/2 - C/2\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}] Q_n(\hat{\theta}_n), \end{aligned}$$

iterate on this inequality and apply the norm equivalence to find that (2.4′) holds wpa1.

As in Theorem 3, we further need to invoke the local convergence results to show that $\theta_k \to \hat{\theta}_n$ as k increases. For that, we need to show that for some $\delta \in (0,1)$, sufficient small, we have $C_n Q_n(\hat{\theta}_n) \leq (1-\delta)R_n^2$, defined in Proposition 3. Note that Assumption 2 implies, without loss of generality, that $R_G > \tilde{R}_n = R_n$ in Proposition 3.

If $\varphi = 0$, we have $Q_n(\hat{\theta}_n) = o_p(1)$ and $C_n = O_p(1)$ which together yield $C_n Q_n(\hat{\theta}_n) = o_p(1) \leq (1-\delta)\tilde{R}_n^2$ wpa1, as in Theorem 3. Now suppose $\varphi > 0$, then we have, for any $\delta \in (0,1)$, that $Q_n(\hat{\theta}_n) \leq [1-\delta]^{-1}\varphi^2$ wpa1. We also have: $\{\underline{\lambda}/2 - C/2\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}\}^{-1} \leq \{(1-\delta)\Delta\}^{-1}$ wpa1, where $\Delta = 1/2[\underline{\sigma}^2 \underline{\lambda}_W - 2L\overline{\lambda}_W^{1/2}\varphi]$. Combine these with the bounds for c_{1n}, c_{3n} above to find that wpa1:

$$C_n Q_n(\hat{\theta}_n) \leq \frac{\gamma^2 c_2 + 2\gamma\overline{c}_3^2/[(1-\delta)\Delta\underline{c}_1]}{[\gamma\underline{c}_1 - \gamma^2 c_2](1-\delta)^2\Delta}\varphi^2 \leq \frac{\Delta\gamma^2 c_2 + 2\gamma\overline{c}_3^2/\underline{c}_1}{[\gamma\underline{c}_1 - \gamma^2 c_2](1-\delta)^3\Delta^2}\varphi^2,$$

using $(1-\overline{\gamma})^2 = (1-\gamma\underline{c}_1/2+\gamma^2 c_2)$. If inequality (2.6) holds strictly, then for $\delta \in (0,1)$ small enough we also have:

$$\frac{\Delta\gamma^2 c_2 + 2\gamma\overline{c}_3^2/\underline{c}_1}{[\gamma\underline{c}_1/2 - \gamma^2 c_2]\Delta^2}\varphi^2 \leq (1-\delta)^5 \left((1-\varepsilon)\frac{\underline{\sigma}}{L\sqrt{\kappa_W}} - \frac{\varphi}{\underline{\sigma}\sqrt{\underline{\lambda}_W}} \right)^2. \tag{B.1.10}$$

Next, note that wpa1: $(1-\delta)^2[(1-\varepsilon)\frac{\underline{\sigma}}{L\sqrt{\kappa_W}} - \frac{\varphi}{\underline{\sigma}\sqrt{\underline{\lambda}_W}}]^2 \leq (1-\delta)[(1-\varepsilon)\frac{\underline{\sigma}}{L\sqrt{\kappa_W}} - \frac{\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}}{\underline{\sigma}\sqrt{\underline{\lambda}_W}}]^2 = (1-\delta)\tilde{R}_n^2$ from Proposition 3. Set $\tilde{\gamma}$ such that $\varepsilon = \tilde{\gamma}/\gamma$ (or smaller). Putting these inequalities together implies that wpa1:

$$C_n\|\overline{g}_n(\hat{\theta}_n)\|_{W_n}^2 \leq (1-\delta)\tilde{R}_n^2,$$

for the same small enough $\delta \in (0,1)$. Now take $k \geq k_n$ given in the Theorem, wpa1 $\|\theta_k - \hat{\theta}_n\| \leq \tilde{R}_n$ when $k \geq k_n$ because k_n was chosen such that the leading term is less than $\delta\tilde{R}_n^2$ to be added to $(1-\delta)\tilde{R}_n^2$ above. Since the conditions for Proposition 3 hold, we have for $k = k_n + j$, $j \geq 0$: $\|\theta_k - \hat{\theta}_n\| \leq (1-\tilde{\gamma})^j\tilde{R}_n$, as desired. □

Theorem B.1.1. *Suppose Assumptions 1, 2, 3 hold and* $Q_n(\hat{\theta}_n) = 0$. *Then, for* γ *small enough, there exists* $\overline{\gamma} \in (0,1)$ *and* $0 < \underline{\lambda} \leq \overline{\lambda} < +\infty$ *such that:* $\|\theta_{k+1} - \hat{\theta}_n\| \leq (1-\overline{\gamma})^{k+1}\sqrt{\overline{\lambda}/\underline{\lambda}}\|\theta_0 - \hat{\theta}_n\|$, *for any starting value* $\theta_0 \in \Theta$, *with probability approaching* 1.

Proof of Theorem B.1.1: Since Assumptions B.15 and B.16 hold (using Lemmas B.12, B.14), Proposition 4 (1)-(2) hold with probability approaching with the same choice of strictly positive constants C_1, C_2, C_3. Denote by L_Q the Lipschitz constant of $\partial_\theta Q_n$. The mean value value theorem implies that for some $\tilde{\theta}_k$ between θ_k and θ_{k+1}:

$$Q_n(\theta_{k+1}) = Q_n(\theta_k) - \gamma\partial_\theta Q_n(\theta_k)P_k\partial_\theta Q_n(\theta_k) - \gamma\{\partial_\theta Q_n(\tilde{\theta}_k) - \partial_\theta Q_n(\theta_k)\}P_k\partial_\theta Q_n(\theta_k)$$

$$\begin{aligned} &\leq Q_n(\theta_k) - \gamma\underline{\lambda}_P\|\partial_\theta Q_n(\theta_k)\|^2 + \gamma^2 L_Q \overline{\lambda}_P^2 \|\partial_\theta Q_n(\theta_k)\|^2 \\ &\leq Q_n(\theta_k) + C_1\{-\gamma\underline{\lambda}_P + \gamma^2 \overline{\lambda}_P^2 L_Q\}(Q_n(\theta_k) - Q_n(\hat{\theta}_n)). \end{aligned}$$

Substract $Q_n(\hat{\theta}_n)$ on both sides to find:

$$Q_n(\theta_{k+1}) - Q_n(\hat{\theta}_n) = \{1 - \gamma C_1\underline{\lambda}_P + \gamma^2 C_1 \overline{\lambda}_P^2 L_Q\}(Q_n(\theta_k) - Q_n(\hat{\theta}_n)),$$

where $0 < 1 - \gamma C_1\underline{\lambda}_P + \gamma^2 C_1\overline{\lambda}_P^2 L_Q < 1$ if $0 < \gamma < \underline{\lambda}_P/[L\overline{\lambda}_P^2]$. Set $(1-\overline{\gamma})^2 = 1 - \gamma C_1\underline{\lambda}_P + \gamma^2 C_1 \overline{\lambda}_P^2 L_Q$ and iterate over $k = 0, \dots$ to find:

$$\|\theta_{k+1} - \hat{\theta}_n\| \leq (1-\overline{\gamma})^{k+1} C_2/C_3 \|\theta_0 - \hat{\theta}_n\|,$$

which is the desired result. □

B.2 PROOFS AND ADDITIONAL RESULTS FOR SECTION 2.3

B.2.1 Additional Results for Over-Identified Models

Proposition B.216. *(Sufficient Conditions: Over-Identified) Consider the following three conditions: (a)* $\sigma_{\min}[G(\theta)'W\overline{G}(\theta_1,\theta_2)] > \underline{\sigma} > 0$*, for all* $\theta, \theta_1, \theta_2 \in \Theta$*, (b) for all* $\theta \in \Theta$*,* $G(\theta) = US(\theta)V$ *for* U, V *full rank,* $S(\theta)$ *symmetric with* $0 < \underline{\lambda}_S < \lambda_{\min}[S(\theta)] < \lambda_{\max}[S(\theta)] < \overline{\lambda}_S < \infty$*, and* $U'WU$ *invertible.*

The following holds: (1) (b) $\Rightarrow$ *(a)* $\Rightarrow$ *Assumption 2 (a), (2) (a) implies* $G(\theta_1)'Wg(\cdot)$ *is one-to-one, for any* $\theta_1 \in \Theta$*.*

Proposition B.217. *(Reparameterization: Over-Identified) Take* h *as in Proposition 10. If Assumption 2 (a) holds for* g *and* $\underline{\sigma} > \overline{\lambda}_W[C_1\overline{\sigma}_h\overline{\sigma}^2 + C_2 L\overline{\sigma}_h^2\overline{\sigma}]/\underline{\sigma}_h^2$*, then Assumption 2 (a) holds for* $g \circ h$*. In particular, if* $h = Au + b$ *is affine with* A *invertible then* $C_1 = C_2 = 0$ *and Assumption 2 (a) holds for* $g \circ h$*.*

B.2.2 Proofs

Proof of Proposition 4: We first prove (2). For any $\theta \in \Theta$, $g(\theta) = g(\theta) - g(\theta^\dagger) = \overline{G}(\theta)(\theta - \theta^\dagger)$, for correctly specified models. This implies that $Q(\theta) = 1/2(\theta - \theta^\dagger)'\overline{G}(\theta)'W\overline{G}(\theta)(\theta - \theta^\dagger)$. Assumption 1 (iii) implies $\sigma_{\max}[\overline{G}(\theta)] \leq \max_{\theta\in\Theta}\sigma_{\max}[G(\theta)] \leq \overline{\sigma} < +\infty$. Assumption 2 (b) implies $\overline{\sigma}\overline{\lambda}_W^{-1/2}\|W^{1/2}\overline{G}(\theta)(\theta-\theta^\dagger)\| \geq \rho\underline{\sigma}\|\theta - \theta^\dagger\|$ and $\|W^{1/2}\overline{G}(\theta)(\theta - \theta^\dagger)\| = \sqrt{2[Q(\theta) - Q(\theta^\dagger)]}$. Putting these together yields:

$$1/2\frac{\rho^2\underline{\sigma}^2}{\overline{\sigma}^2\overline{\lambda}_W}\|\theta - \theta^\dagger\|^2 \leq Q(\theta) - Q(\theta^\dagger) \leq 1/2\overline{\sigma}^2\overline{\lambda}_W\|\theta - \theta^\dagger\|^2.$$

Now, we prove (1). We have $\partial_\theta Q(\theta) = G(\theta)'Wg(\theta) = G(\theta)'W\overline{G}(\theta)(\theta-\theta^\dagger)$. Assumption 2 (b) implies:

$$\|\partial_\theta Q(\theta)\|^2 \geq \rho^2\underline{\sigma}^2\|\theta - \theta^\dagger\|^2 \geq \frac{\rho^2\underline{\sigma}^2}{1/2\overline{\sigma}^2\overline{\lambda}_W}[Q(\theta) - Q(\theta^\dagger)],$$

using (2). This is the desired result. □

Proof of Proposition 5: For correctly specified models, $\partial_{\theta'}Q(\theta) = G(\theta)'W\overline{G}(\theta)(\theta - \theta^\dagger)$. 1) If the PL inequality holds, the quadratic lower bound implies $\|G(\theta)'W\overline{G}(\theta)(\theta - \theta^\dagger)\|^2 \geq \mu C_2\|\theta - \theta^\dagger\|^2$, i.e. Assumption 2 (b) holds.

2) By definition, Q is quasar-convex if, and only if, there are $\lambda \geq 1$ and $\mu \geq 0$ such that:

$$\partial_\theta Q(\theta)(\theta - \theta^\dagger) \geq \frac{1}{\lambda}\{Q(\theta) - Q(\theta^\dagger)\} + \frac{\mu}{2\lambda}\|\theta - \theta^\dagger\|^2,$$

where $\partial_\theta Q(\theta)(\theta - \theta^\dagger) = (\theta - \theta^\dagger)'G(\theta)'W\overline{G}(\theta)(\theta - \theta^\dagger)$. Since $Q(\theta) - Q(\theta^\dagger) \geq 0$ we have:

$$(\theta - \theta^\dagger)'G(\theta)'W\overline{G}(\theta)(\theta - \theta^\dagger) \geq \frac{\mu}{2\lambda}\|\theta - \theta^\dagger\|^2.$$

Now apply the Cauchy-Schwarz inequality to find:

$$\|\theta - \theta^\dagger\| \|G(\theta)' W \overline{G}(\theta)(\theta - \theta^\dagger)\| \geq (\theta - \theta^\dagger)' G(\theta)' W \overline{G}(\theta)(\theta - \theta^\dagger) \geq \frac{\mu}{2\lambda} \|\theta - \theta^\dagger\|^2,$$

which implies Assumption 2 (b). □

Proof of Proposition 6: 1) Strong monotonicity of Ag implies $(\theta_1 - \theta_2)' A\overline{G}(\theta_1, \theta_2)(\theta_1 - \theta_2) \geq \mu\|\theta_1 - \theta_2\|^2$ since $g(\theta_1) - g(\theta_2) = \overline{G}(\theta_1, \theta_2)(\theta_1 - \theta_2)$. For any unit vector v, take $\theta_2 = \theta_1 + \varepsilon v$ and let $\varepsilon \to 0$ to find $v' AG(\theta_1) v = \frac{1}{2} v'[AG(\theta_1) + G(\theta_1)' A'] v \geq \mu$ so that $G(\theta_1)$ has full rank and $AG(\theta_1) + G(\theta_1)' A'$ is positive definite. We have $\sigma_{\min}[G(\theta)] \geq \mu \sigma_{\min}(A)^{-1} := \underline{\sigma} > 0$. Pick $\theta_2 = \theta^\dagger$, use the Cauchy-Schwarz inequality to find $\|A'(\theta - \theta^\dagger)\| \|\overline{G}(\theta, \theta^\dagger)(\theta - \theta^\dagger)\| \geq (\theta - \theta^\dagger)' A\overline{G}(\theta, \theta^\dagger)(\theta - \theta^\dagger) \geq \mu\|\theta - \theta^\dagger\|^2$. Because $G(\theta)' W$ is invertible, so we can write $\|G(\theta)' W \overline{G}(\theta, \theta^\dagger)(\theta - \theta^\dagger)\| = \|G(\theta)' W A^{-1} A\overline{G}(\theta, \theta^\dagger)(\theta - \theta^\dagger)\| \geq \underline{\sigma} \underline{\lambda}_W \|A\overline{G}(\theta, \theta^\dagger)(\theta - \theta^\dagger)\| \geq \mu \underline{\lambda}_W \underline{\sigma} \|\theta - \theta^\dagger\|^2$ Assumption 2 (b) also holds for an appropriate $0 < \rho \leq \mu \underline{\lambda}_W$.
2) Strong injectivity of g implies $\|\overline{G}(\theta_1, \theta_2)(\theta_1 - \theta_2)\| \geq \mu\|\theta_1 - \theta_2\|$, for any pair θ_1, θ_2. Using the same arguments as above: $G(\theta)$ has full rank for all θ and $\|G(\theta)' W \overline{G}(\theta, \theta^\dagger)(\theta - \theta^\dagger)\| \geq \underline{\sigma} \underline{\lambda}_W \mu \|\theta - \theta^\dagger\|$. □

Proof of Proposition 7. Assumption 1 (ii)-(vi) implies Assumption 2 (a) holds locally (Lemma B.13). Hence, for $\|\theta - \theta^\dagger\| \leq r$, we have $\|G(\theta)' W \overline{G}(\theta - \theta^\dagger)\| \geq \rho \underline{\sigma} \|\theta - \theta^\dagger\|$. Condition (N) implies that for $R \geq \|\theta - \theta^\dagger\| \geq r$ we have:

$$\inf_{\theta, R \geq \|\theta - \theta^\dagger\| \geq r} \|\partial_\theta Q(\theta)\| \geq \delta(r, R) \geq \frac{\delta(r, R)}{R} \|\theta - \theta^\dagger\|,$$

by continuity, compactness and the Weierstrass Theorem. We can pick $\rho < \frac{\delta(r,R)}{R\underline{\sigma}}$.

□

Proof of Proposition 8: For any $\theta \in \Theta$, we have:

$$\begin{aligned} Q(\theta) - Q(\theta^\dagger) &= \frac{1}{2}\left(g(\theta)'Wg(\theta) - g(\theta^\dagger)'Wg(\theta^\dagger)\right) \\ &= \frac{1}{2}\left(g(\theta) + g(\theta^\dagger)\right)' W \left(g(\theta) - g(\theta^\dagger)\right) \\ &= \frac{1}{2}\left(g(\theta) + g(\theta^\dagger)\right)' W\overline{G}(\theta)(\theta - \theta^\dagger) \\ &= \frac{1}{2}(\theta - \theta^\dagger)'\overline{G}(\theta)'W\overline{G}(\theta)(\theta - \theta^\dagger) - g(\theta^\dagger)'W\overline{G}(\theta)(\theta - \theta^\dagger), \end{aligned}$$

the first term in the last display matches the one in the proof of Proposition 4. Note that $g(\theta^\dagger)'WG(\theta^\dagger) = 0$ and $\|G(\theta^\dagger) - \overline{G}(\theta)\| \leq L\|\theta - \theta^\dagger\|$, together these allow to bound the second term:

$$\|g(\theta^\dagger)'W\overline{G}(\theta)(\theta - \theta^\dagger)\| = \|g(\theta^\dagger)'W[\overline{G}(\theta) - G(\theta^\dagger)](\theta - \theta^\dagger)\| \leq \overline{\lambda}_W^{1/2} L\sqrt{\varphi}\|\theta - \theta^\dagger\|^2.$$

Let $C_2 = 1/2\frac{\rho^2\underline{\sigma}^2}{\overline{\sigma}^2\overline{\lambda}_W}$ and $C_3 = 1/2\overline{\sigma}^2\overline{\lambda}_W$, as in the proof of Proposition 4. Take $C_4 = \overline{\lambda}_W^{1/2} L$, this yields (2):

$$(C_2 - C_4\sqrt{\varphi})\|\theta - \theta^\dagger\|^2 \leq Q(\theta) - Q(\theta^\dagger) \leq (C_3 + C_4\sqrt{\varphi})\|\theta - \theta^\dagger\|^2.$$

For (1), we have $\partial_\theta Q(\theta) = G(\theta)'Wg(\theta)$ and $G(\theta^\dagger)'Wg(\theta^\dagger) = 0$, so that:

$$\partial_\theta Q(\theta) = G(\theta)'W\overline{G}(\theta)(\theta - \theta^\dagger) + \{G(\theta) - G(\theta^\dagger)\}'Wg(\theta^\dagger).$$

Apply the reverse triangular inequality to find:

$$\begin{aligned} \|\partial_\theta Q(\theta)\| &\geq \rho\underline{\sigma}\|\theta - \theta^\dagger\| - \sqrt{\varphi\overline{\lambda}_W}L\|\theta - \theta^\dagger\| \\ &= \left(\rho\underline{\sigma} - \sqrt{\varphi\overline{\lambda}_W}L\right)\|\theta - \theta^\dagger\|, \end{aligned}$$

where L is the Lipschitz constant of G. Finally, (1′) can be derived from (1) and (2) assuming $(\rho\underline{\sigma} - \sqrt{\varphi\overline{\lambda}_W}L) > 0$. □

Proof of Proposition 9: We first prove (1). (a) ⇒ Assumption 2 (a) is immediate. Under (c), $G(\theta) = \partial^2_{\theta,\theta'}F(\theta)$ is symmetric and strictly positive definite so (b) holds. Suppose (b) holds, then $\overline{G}(\theta_1,\theta_2) = U\{\int_0^1 S(\omega\theta_1 + (1-\omega)\theta_2)d\omega\}V$ where $\int_0^1 S(\omega\theta_1 + (1-\omega)\theta_2)d\omega$ is symmetric. Concavity of the smallest positive eigenvalue on the set of positive definite matrices, and Jensen's inequality imply: $\lambda_{\min}[\int_0^1 S(\omega\theta_1 + (1-\omega)\theta_2)d\omega] \geq \int_0^1 \lambda_{\min}[S(\omega\theta_1 + (1-\omega)\theta_2)]d\omega \geq \min_{\theta\in\Theta}\lambda_{\min}[S(\theta)] > 0$, by positive definiteness and continuity of $S(\cdot)$. Finally,

$$\sigma_{\min}[\overline{G}(\theta_1,\theta_2)] \geq \sigma_{\min}(U)\sigma_{\min}(V)\min_{\theta\in\Theta}\lambda_{\min}[S(\theta)] > \underline{\lambda}_S\underline{\sigma}_U\underline{\sigma}_V > 0,$$

taking $\underline{\sigma}_U\underline{\sigma}_V$ to be smallest singular values of U, V. Hence (a) holds.

For (2), note that $g(\theta_1) - g(\theta_2) = \overline{G}(\theta_1,\theta_2)(\theta_1 - \theta_2)$, using Lemma B.11. With condition (a), we have $g(\theta_1) - g(\theta_2) = 0 \Leftrightarrow \theta_1 = \theta_2$, i.e. $g(\cdot)$ is one-to-one.

For (3), $g(\cdot)$ is one-to-one, take $\phi(\cdot) = g^{-1}(\cdot)$, one-to-one, and $\psi = I_d - \theta^\dagger$, we get $h(\theta) = \theta - \theta^\dagger$ linear for which strong convexity is immediate. □

Proof of Proposition 10: Under Assumption 2 (a), G has full rank for all $\theta \in \Theta$. Take $u \in \mathcal{U}$, let $\theta = h(u)$, the chain rule implies that $\partial_u g \circ h(u) = \partial_\theta g \circ h(u)\partial_u h(u)$ has full rank for all $u \in \mathcal{U}$. Then, we have:

$$\int_0^1 G \circ h(\omega u + (1-\omega)u)\partial_u h(\omega u + (1-\omega)u^\dagger)d\omega$$

$$= \int_0^1 G(\omega\theta + (1-\omega)\theta^\dagger)d\omega\partial_u h(u^\dagger)$$

$$+\int_0^1 G(\omega\theta+(1-\omega)\theta^\dagger)[\partial_u h(\omega u+(1-\omega)u^\dagger)-\partial_u h(u^\dagger)]d\omega$$
$$+\int_0^1 [G\circ h(\omega u+(1-\omega)u)-G(\omega\theta+(1-\omega)\theta^\dagger)]\partial_u h(\omega u+(1-\omega)u^\dagger)d\omega,$$

using Weyl's inequality and a minoration of the singular value for a matrix product, we get:

$$\sigma_{\min}[\int_0^1 \partial_u h(\omega u+(1-\omega)u^\dagger)G\circ h(\omega u+(1-\omega)u)d\omega]\geq \underline{\sigma}_h\underline{\sigma}-C_1\overline{\sigma}-C_2L\overline{\sigma}_h,$$

which is strictly positive under the stated condition. After the change of variable, the Assumption 2 (a) holds if:

$$\partial_u h(u)'G(g(u))'W\Big\{\int_0^1 \partial_u h(\omega u+(1-\omega)u^\dagger)G\circ h(\omega u+(1-\omega)u)d\omega\Big\},$$

has singular values bounded below by a strictly positive term which holds for C_1, C_2 bounded as in the Proposition statement. In particular, when h is affine, $C_1 = C_2 = 0$ and $0 < \underline{\sigma}_h = \sigma_{\min}[A] \leq \sigma_{\max}[A] \leq \overline{\sigma}_h < \infty$, so that the condition is automatically satisfied. □

Proof of Proposition B.216: First, we prove (1). (a) ⇒ Assumption 2 (a) is immediate. Suppose (b) holds, take any $\theta, \theta_1, \theta_2 \in \Theta$, then $G(\theta)'W\overline{G}(\theta_1,\theta_2) = V'S(\theta)U'WU\int_0^1\{S(\omega\theta_1+(1-\omega)\theta_2)\}d\omega V$. By assumption, $V'S(\theta)$ and $U'WU$ have full rank. As in the proof of Proposition 9, $\int_0^1\{S(\omega\theta_1+(1-\omega)\theta_2)\}d\omega$ has full rank for any θ_1,θ_2, and V is invertible. Hence, $S(\theta)U'WU\int_0^1\{S(\omega\theta_1+(1-\omega)\theta_2)\}d\omega V$ is invertible, U has full rank so that $G(\theta)'W\overline{G}(\theta_1,\theta_2)$ has full rank for all $\theta, \theta_1, \theta_2$.

For part (2), take any $\theta_1, \theta_2, \theta_3$. Suppose $G(\theta_1)'Wg(\theta_2) = G(\theta_1)'Wg(\theta_3)$, apply Lemma B.11 to find $G(\theta_1)'W\overline{G}(\theta_2, \theta_3)(\theta_2 - \theta_3) = 0 \Rightarrow \theta_2 = \theta_3$ under condition (a). □

Proof of Proposition B.217: We'll proceed similarly to the proof of Proposition 10:

$$\begin{aligned}
&\int_0^1 \partial_u' h(\omega u + (1-\omega)u^\dagger) G \circ h(\omega u + (1-\omega)u^\dagger)' d\omega W G \circ h(u) \partial_u h(u) \\
&= \partial_u' h(u^\dagger) \int_0^1 G(\omega\theta + (1-\omega)\theta^\dagger)' d\omega W G(\theta) \partial_u h(u) \\
&+ \int_0^1 [\partial_u h(\omega u + (1-\omega)u^\dagger) - \partial_u h(u^\dagger)]' G(\omega\theta + (1-\omega)\theta^\dagger)' d\omega W G(\theta) \partial_u h(u) \\
&+ \int_0^1 \partial_u' h(\omega u + (1-\omega)u^\dagger) [G \circ h(\omega u + (1-\omega)u^\dagger) \\
&- G(\omega\theta + (1-\omega)\theta^\dagger)]' d\omega W G(\theta) \partial_u h(u).
\end{aligned}$$

As before, we get: $\sigma_{\min}[\int_0^1 \partial_u' h(\omega u + (1-\omega)u^\dagger) G \circ h(\omega u + (1-\omega)u^\dagger)' d\omega W G \circ h(u) \partial_u h(u)] \geq \underline{\sigma\sigma}_h^2 - C_1 \overline{\sigma}_h \overline{\sigma}^2 \overline{\lambda}_W - C_2 L \overline{\sigma}_h^2 \overline{\sigma} \overline{\lambda}_W$ which is positive under the stated condition. As before, for h affine we have $C_1 = C_2 = 0$ so that the condition holds for A finite and invertible. □

B.3 COMMON METHODS AND THEIR PROPERTIES

B.3.1 A survey of empirical practice

Survey methodology: The survey covers empirical papers published in the American Economic Review (AER) between 2016 and 2018. The focus on this specific outlet is driven by the mandatory data and code policy enacted in 2005. Indeed, since a number of papers provide little or no detail in the paper on the methodology used to compute estimates numerically, it is important to read the replication codes to determine what was implemented. The search function in JSTOR was used to find the papers matching the survey criteria. The database did not include more recent publications at the time of the survey.[4] Table B.3.1 was constructed by reading through the main text, supplemental material, and all available replication codes of the selected papers.

Survey results: Table B.3.1 provides an overview of the quantitative results of the survey. Additional details on the algorithms in the table are given below. There are 23 papers in total, a little over 7 papers per year. Excluding the estimation with 147 parameters, the average estimation has around 10 coefficients, and the median is 6. 3 papers used more than one starting value, and the remaining 20 papers either used the solver default or typed in a specific value in the replication code. There is generally no information provided on the origin of these specific starting values. Of the papers using multiple starting values, one did not provide replication codes, and the other two used 12 and 50 starting points. Some of the estimations are very

---

[4] The search function in JSTOR allows to search for keywords within the title, abstract, main text, and supplemental material of a paper. Further screening ensures that each paper in the search results actually implements at least one of the estimations considered. The search criteria include keywords: “Method of Moments," “Indirect Inference," “Method of Simulated Moments," “Minimum Distance," and “MM."

Table B.3.1: American Economic Review 2016-2018: GMM and related empirical estimations

Method	# Papers	# Parameters (p)	Data available
Nelder-Mead - one starting value	7	2,6 (×2),11,13 (×2),147	3
Simulated Annealing + Nelder-Mead	2	4,13	1
Nelder-Mead - multiple starting values	2	?,6	1‡
Pattern Search	2	6,147	1†
Genetic Algorithm	2	9,14	1
Simulated Annealing	2	4,13	2†
MCMC	1	15	1
Grid Search	1	5	1
No description	3	-	-
Stata/Mata default	4	3,6 (×2),38	3⋆

Legend: # Parameters correspond to the size of the largest specification. Data avail. reports if the dataset is included with the replication files. Estimations surveyed include: Generalized Method of Moments (GMM), Minimum Distance (MD), Simulated Method of Moments (SMM), and Indirect Inference. ?: information not available due to the lack of replication codes. ⋆: one of the 3 papers reported to include data requires to download the PSID dataset separately. †: two papers in total also rely separately on Nelder-Mead, so they are also reported under Nelder-Mead. ‡: one paper provides data without codes.

time-consuming. For instance, Lise & Robin (2017) use MCMC for estimation (but not inference) and report that each evaluation of the moments takes 45s. In total, their estimation takes more than a week to run in a 96-core cluster environment.

As mentioned in the introduction, although convex optimizers such as (stochastic) gradient-descent and quasi-Newton methods are commonly used to solve large scale convex minimization problems, they are virtually absent from the survey. Overall, 11 papers rely on the Nelder-Mead algorithm, alone or in combination with another method, making it the most popular optimizer in this survey. Pattern search, used in 2 papers, belongs to the same family of algorithms as Nelder-Mead. The following provides a brief overview of the properties of the main Algorithms found used in Table B.3.1.

B.3.2 A brief summary of the Algorithms' properties

The following briefly discussed the properties of four algorithms from Table B.3.1: Nelder-Mead, Grid Search, Multi-Start, and Simulated Annealing. Further discussion, descriptions, and references can be found in Appendix B.6.

Nelder-Mead (NM) is the most popular method in the survey, it can be used even if Q_n is discontinuous. Its convergence properties, which measure its ability to find valid estimates, are somewhat limited however. For some smooth convex problems, it can be shown to converge to values that are neither locally nor globally optimal. The grid-search converges to the solution under weak conditions, unlike NM. It is very slow, however, and often not practical when estimating three or more coefficients. Simulated annealing (SA) is not deterministic. Still it converges, in probability, under weak conditions to the solution. Albeit, the convergence is predicted to be slower than grid search. A common approach to improve the convergence of a given algorithm is to combine it with multiple starting values. The required number of starting values depends on Q_n and the choice of algorithm. Andrews (1997) provides an asymptotically valid stopping rule for correctly specified GMM models.

When Q_n is strongly convex, several gradient-based methods discussed below are rapidly, globally convergent and do not suffer from a curse of dimensionality. This implies that it is possible to estimate a large number of parameters in a reasonable amount of time. Similar convergence properties are derived in this paper, under rank conditions instead of convexity.

B.3.3 Revisting some empirical results

In the survey, 11 papers provide replication files with the codes and data necessary to replicate the results. Excluding those requiring Stata, Fortran, Eviews, and C/C++ to run, 3 papers remain: Gill & Prowse (2012), Sieg & Yoon (2017) and Kelly et al. (2016). The first two involve a Simulated Method of Moments (SMM) estimation with discrete outcomes, which are non-smooth moments and fall outside the framework of this paper. Kelly et al. (2016) use Simulated Annealing for estimation. The rank condition used in this paper is not satisfied for their estimation. As shown below, when the rank condition holds: any local optimum is a valid estimator and the model is globally identified. Here, there are values $\tilde{\theta}_n = (9.67, -0.19, 0.47)$ with a lower objective $Q_n(\tilde{\theta}_n) = 1.0419$ than in the original results $\hat{\theta}_n = (3.20, -0.34, 0.54)$ and $Q_n(\hat{\theta}_n) = 1.0469$.[5] The parameters $\theta = (\omega, \vartheta, \delta)$ are coefficients in a Merton Jump model (see Kelly et al., 2016, SecB).[6] Figure B.5.5, Appendix B.5.4, illustrates for a one-dimensional subproblem how the estimation is non-convex and the rank conditions fail. The figure also illustrates that their estimates are not globally optimal.

B.4 R CODE FOR THE MA(1) EXAMPLE

```
library(stats) # fit an AR(p) model
library(pracma) # compute jacobian


n = 200     # sample size n
theta = -1/2 # MA(1) coefficient
```

---

[5] Another value, within their specified bounds, is $\tilde{\theta}_n = (50.31, -0.10, 0.48)$ with a lower objective value $Q_n(\tilde{\theta}_n) = 1.0367$.

[6] The authors do not report standard errors, but it seems that the objective values are close enough and the parameters far enough that identification of the Merton Jump parameters could be a concern.

```
set.seed(123) # set the seed for random numbers
e = rnorm(n+1)       # draw innovations
y = e[2:(n+1)] - theta*e[1:n] # generate MA(1) data
p = 12      # number of lags for the AR(p) models

beta <- function(theta) {
   # computes the p-limit of the OLS estimates
   # V = covariance matrix of (y_{t-1},...,y_{t-p})
   V = diag(p+1)*(1+theta^2) # variances on the diagonal
   diag(V[,-1]) = -theta # autocovariance
   V = t(V)              # transpose
   diag(V[,-1]) = -theta # autocovariance
   return(
      solve( V[2:(p+1),2:(p+1)], V[1,2:(p+1)] )
      # p-limit = inv(V)*( vector of autocovariances )
   )
}

# Fit the AR(p) auxiliary model:
ols_p = c(ar.ols( y, aic = FALSE, order.max = p, demean = FALSE,
   intercept = FALSE )$ar)

moments <- function(theta) {
   # computes the sample moments gn
   return( ols_p - beta(theta) ) # gn = psi_n - psi(theta)
}

objective <- function(theta,disp = FALSE) {
   # compute the sample objective Qn
   if (disp == TRUE) {
      print(round(theta,3)) # print to tack R’s optimization paths
   }
   mm = moments(theta) # compute sample moments gn
   return( t(mm)%*%mm ) # compute Qn = gn’*gn (W = Id)
}

dQ <- function(theta,disp=FALSE) {
   # compute the derivative of Qn
   # gradient of Qn = -2*d psi(theta)/ d theta’ * gn(theta)
   return(-2*t(jacobian(beta,theta))%*%moments(theta))
}

# L-BFGS-B: with bound constraints
o1 = optim(0.95,objective,gr=dQ,method="L-BFGS-B",lower=c(-1),upper=c(1)
    ,disp=TRUE)
# BFGS: without bound constraints
o2 = optim(0.95,objective,gr=dQ,method="BFGS",disp=TRUE)

# *********************************
#        Gauss-Newton
# *********************************
gamma = 0.1 # learning rate
coefsGN = rep(0,150) # 150 iterations in total
coefsGN[1] = 0.95 # starting value: theta = 0.95

for (b in 2:150) { # main loop for Gauss-Newton
   Gn = -jacobian(beta,coefsGN[b-1]) # 1. compute Jacobian
   mom = moments(coefsGN[b-1]) # 2. compute moments
   coefsGN[b] = coefsGN[b-1] - gamma*solve(t(Gn)%*%Gn,t(Gn)%*%mom) # 3.
      update
}  # repeat for each b
```

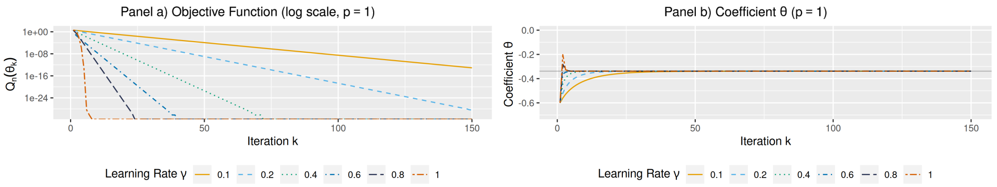
```
# Put the results into a table:
results = matrix(NA,2,3)
colnames(results) = c('L-BFGS-B','BFGS','GN')
results[1,] = c(o1$par,o2$par,coefsGN[150])
results[2,] = sapply(results[1,],objective)
rownames(results) = c('theta','Qn(theta)')

print(results,digits=3)
# Output should look like this:
#         L-BFGS-B BFGS   GN
# theta      -1.0 -6.979 -0.626
# Qn(theta)  1.7 0.397 0.101
```

B.5 ADDITIONAL SIMULATION, EMPIRICAL RESULTS

B.5.1 Estimating an MA(1) model

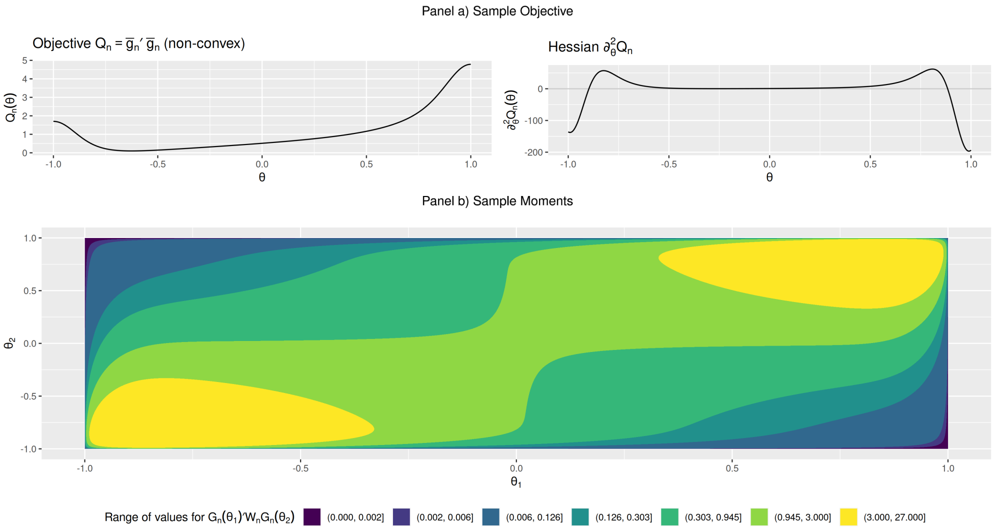
The following reports GN results with $\gamma \in \{0.1, 0.2, 0.4, 0.6, 0.8, 1\}$, $p = 1$ and $p = 12$, equal and optimal weighting ($p = 12$).

Figure B.5.1: GN iterations: different learning rates



Legend: simulated sample of size $n = 200$, $\theta^\dagger = -1/2$, $\overline{g}_n(\theta) = \hat{\beta}_n - \beta(\theta)$. Grey horizontal line: $\hat{\theta}_n = -0.339$. Just Identified ($p = 1$).

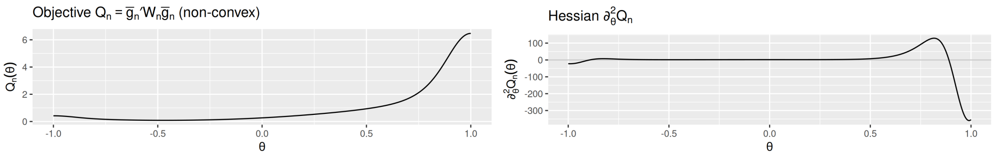
Figure B.5.2: Non-convexity and the rank condition ($p = 12$, equal weighting $W_n = I_d$)



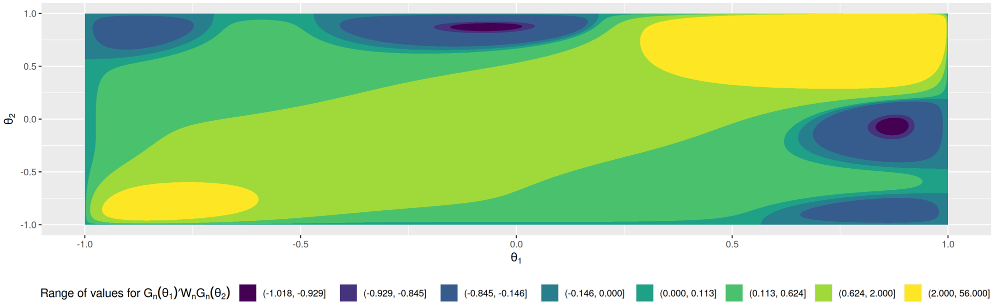
Legend: simulated sample of size $n = 200$, $\theta^\dagger = -1/2$, $\overline{g}_n(\theta) = \hat{\beta}_n - \beta(\theta)$, $W_n = I_d$. The GMM objective (panel a) is non-convex but the sample moments (panel b) satisfy the rank condition: $G_n(\theta_1)' G_n(\theta_2)$ is full rank (non-zero) for all $(\theta_1, \theta_2) \in (-1, 1) \times (-1, 1)$.

Figure B.5.3: Non-convexity and the rank condition ($p = 12$, optimal weighting $W_n = V_n^{-1}$)

Panel a) Sample Objective

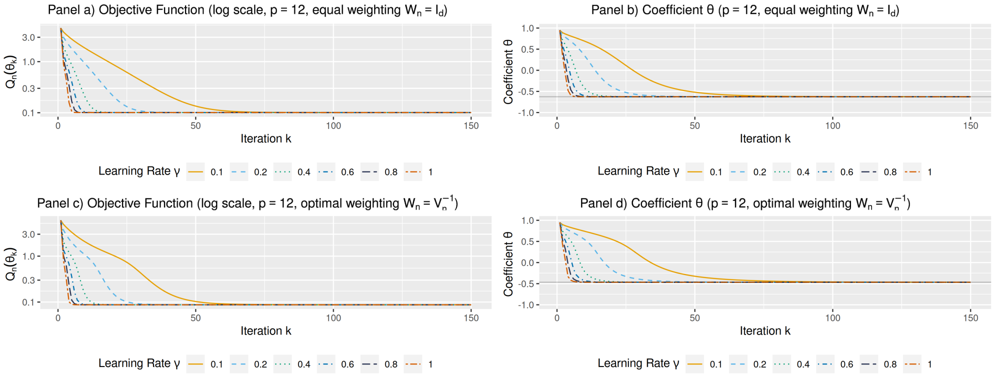


Panel b) Sample Moments



Legend: simulated sample of size $n = 200$, $\theta^\dagger = -1/2$, $\overline{g}_n(\theta) = \hat{\beta}_n - \beta(\theta)$, $W_n = V_n^{-1}$ with $V_n = n\hat{\text{var}}(\hat{\beta}_n)$. The GMM objective (panel a) is non-convex but the sample moments (panel b) does not satisfy the rank condition: $G_n(\theta_1)' V_n^{-1} G_n(\theta_2)$ is not full rank (non-zero) for all $(\theta_1, \theta_2) \in (-1, 1) \times (-1, 1)$.

Figure B.5.4: GN iterations: equal and optimal weighting, different learning rates



Legend: simulated sample of size $n = 200$, $\theta^{\dagger} = -1/2$, $\overline{g}_n(\theta) = \hat{\beta}_n - \beta(\theta)$, $W_n = I_d$ (equal weighting, top panel), or $W_n = V_n^{-1}$ (optimal weighting, bottom panel), with $V_n = n\hat{\text{var}}(\hat{\beta}_n)$. Grey horizontal line: $\hat{\theta}_n = -0.626$ (equal weighting), $\hat{\theta}_n = -0.466$ (optimal weighting).

B.5.2 Demand for Cereal

B.5.3 Impulse Response Matching

The following tables report results for GN using a range of tuning parameters γ. Since the rank condition does not hold towards the lower bound for η, ν, GN alone can crash and/or fail to converge. Following Forneron (2023), we can introduce a global step:

$$\theta_{k+1} = \theta_k - \gamma P_k G_n(\theta_k)' W_n \overline{g}_n(\theta_k) \tag{2.1}$$

$$\text{if } \|\overline{g}_n(\theta^{k+1})\|_{W_n} < \|\overline{g}_n(\theta_{k+1})\|_{W_n}, \text{ set } \theta_{k+1} = \theta^{k+1}$$

Table B.5.1: Demand for Cereal: GN with different learning rates

		STDEV				INCOME				objs	# of crashes
		const.	price	sugar	mushy	const.	price	sugar	mushy		
TRUE	est	0.28	2.03	-0.01	-0.08	3.58	0.47	-0.17	0.69	33.84	-
	se	0.11	0.76	0.01	0.15	0.56	3.06	0.02	0.26	-	
$\gamma = 0.1$	avg	0.28	2.03	-0.01	-0.08	3.58	0.47	-0.17	0.69	33.84	0
	std	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.2$	avg	0.28	2.03	-0.01	-0.08	3.58	0.47	-0.17	0.69	33.84	0
	std	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.4$	avg	0.28	2.03	-0.01	-0.08	3.58	0.47	-0.17	0.69	33.84	0
	std	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.6$	avg	0.28	2.03	-0.01	-0.08	3.58	0.47	-0.17	0.69	33.84	0
	std	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.8$	avg	0.28	2.03	-0.01	-0.08	3.58	0.47	-0.17	0.69	33.84	0
	std	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
$\gamma = 1$	avg	0.28	2.03	-0.01	-0.08	3.58	0.47	-0.17	0.69	33.84	0
	std	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	

Legend: Comparison for 50 starting values in $[-10, 10] \times \cdots \times [-10, 10]$. Avg, Std: sample average and standard deviation of optimizer outputs. TRUE: full sample estimate (est) and standard errors (se). Objs: avg and std of minimized objective value. # of crashes: optimization terminated because the objective function returned an error. GN run with $\gamma \in \{0.1, 0.2, 0.4, 0.6, 0.8, 1\}$ for $k = 150$ iterations for all starting values.

where the sequence $(\theta^k)_{k \geq 0}$ is predetermined and dense in Θ. The results rely on the Sobol sequence, independently randomized for each of the 50 starting values.[7] Results are reported with and without the global step. Also, the former implements error-handling (try-catch).

---

[7] We take $(s_k)_{k \geq 0}$ in $[0, 1]^p$, $p \geq 1$ is the number of parameters, draw one vector $(u_1, \ldots, u_p) \sim \mathcal{U}_{[0,1]^p}$, for each starting value, and compute $\tilde{s}_k = (s_k + u)$ modulo 1, then map $\tilde{s}_k$ to the bounds for $\theta = (\theta_1, \ldots, \theta_p)$. The randomization is used to create independent variation in the global step between starting values to emphasize that convergence does not rely on a specific value in the sequence $(\theta^k)_{k \geq 0}$; this is called a random shift (see Lemieux, 2009, Ch6.2.1).

Table B.5.2: Without reparameterization : GN with different learning rates

		η	ν	ρ_s	σ_s	objs	# of crashes
TRUE	est	0.30	0.29	0.39	0.17	4.65	-
GN WITHOUT GLOBAL STEP							
$\gamma = 0.1$	avg	0.30	0.29	0.39	0.17	4.65	2
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.2$	avg	0.30	0.29	0.39	0.17	4.65	2
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.4$	avg	0.30	0.29	0.39	0.17	4.65	4
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.6$	avg	0.30	0.29	0.39	0.17	4.65	8
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.8$	avg	0.30	0.29	0.39	0.17	4.65	12
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 1$	avg	0.30	0.29	0.39	0.17	4.65	28
	std	0.00	0.00	0.00	0.00	0.00	
GN WITH GLOBAL STEP							
$\gamma = 0.1$	avg	0.30	0.29	0.39	0.17	4.65	0
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.2$	avg	0.30	0.29	0.39	0.17	4.65	0
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.4$	avg	0.30	0.29	0.39	0.17	4.65	0
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.6$	avg	0.30	0.29	0.39	0.17	4.65	0
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.8$	avg	0.30	0.29	0.39	0.17	4.65	0
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 1$	avg	0.30	0.29	0.39	0.17	4.65	0
	std	0.00	0.00	0.00	0.00	0.00	
lower bound		0.05	0.01	-0.95	0.01	-	-
upper bound		0.99	0.90	0.95	12	-	-

Legend: Comparison for 50 starting values. TRUE: full sample estimate (est). GN WITH GLOBAL STEP: Gauss-Netwon augmented with a global sequence. Both are run for $k = 150$ iterations in total, for all starting values. Objs: avg and std of minimized objective value. # of crashes: optimization terminated because objective returned error. Lower/upper bound used for the reparameterization.

Table B.5.3: With reparameterization : GN with different learning rates

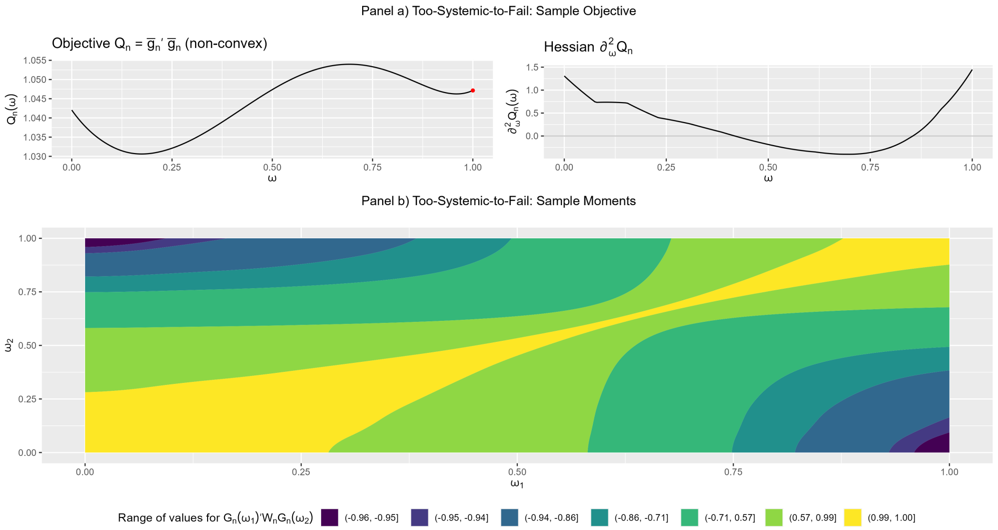
		η	ν	ρ_s	σ_s	objs	# of crashes
TRUE	est	0.30	0.29	0.39	0.17	4.65	-
GN WITHOUT GLOBAL STEP							
$\gamma = 0.1$	avg	0.30	0.29	0.39	0.17	4.65	5
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.2$	avg	0.30	0.29	0.39	0.17	4.65	10
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.4$	avg	0.30	0.29	0.39	0.17	4.65	20
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.6$	avg	0.30	0.29	0.39	0.17	4.65	22
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.8$	avg	0.30	0.29	0.39	0.17	4.65	25
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 1$	avg	0.30	0.29	0.39	0.17	4.65	29
	std	0.00	0.00	0.00	0.00	0.00	
GN WITH GLOBAL STEP							
$\gamma = 0.1$	avg	0.30	0.29	0.39	0.17	4.65	0
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.2$	avg	0.30	0.29	0.39	0.17	4.65	0
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.4$	avg	0.30	0.29	0.39	0.17	4.65	0
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.6$	avg	0.30	0.29	0.39	0.17	4.65	0
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 0.8$	avg	0.30	0.29	0.39	0.17	4.65	0
	std	0.00	0.00	0.00	0.00	0.00	
$\gamma = 1$	avg	0.30	0.29	0.39	0.17	4.65	0
	std	0.00	0.00	0.00	0.00	0.00	
lower bound		0.05	0.01	-0.95	0.01	-	-
upper bound		0.99	0.90	0.95	12	-	-

Legend: Comparison for 50 starting values. TRUE: full sample estimate (est). GN WITH GLOBAL STEP: Gauss-Netwon augmented with a global sequence. Both are run for $k = 150$ iterations in total, for all starting values. Objs: avg and std of minimized objective value. # of crashes: optimization terminated because objective returned error. Lower/upper bound used for the reparameterization.

B.5.4 Comparison of Rank Condition and Convexity

Non-convexity and failure of rank conditions in Kelly et al. (2016). The moments are computationally intensive to evaluate in Kelly et al. (2016), for each value θ they use a numerical solver to compute option prices and then evaluate the moments. The illustrate how convexity and the rank conditions fail in this example, consider a one-dimensional sub-problem $\theta(\omega) = \omega\tilde{\theta}_n + (1-\omega)\hat{\theta}_n$ where $\hat{\theta}_n = (3.20, -0.34, 0.54)$ are the estimates reported in Kelly et al. (2016) and $\tilde{\theta}_n = (9.67, -0.19, 0.47)$ is a value for which Q_n is strictly smaller. Figure B.5.5 is similar to figure B.5.2, the moments were evaluated on a coarse grid and interpolated to reduce the computational burden. The top left panel shows the objective function which is non-convex. The red dot indicates the estimates reported in Kelly et al. (2016). The Hessian can be positive, negative or zero depending on the value ω (top right panel). However, here the rank condition also fails since the scalar $\partial_\omega \overline{g}_n(\theta(\omega_1))'\partial_\omega \overline{g}_n(\theta(\omega_2))$ can change sign. This implies that there are values for which $\sigma_{\min}[\partial_\omega \overline{g}_n(\theta(\omega_1))'\partial_\omega \overline{g}_n(\theta(\omega_2))] = |\partial_\omega \overline{g}_n(\theta(\omega_1))'\partial_\omega \overline{g}_n(\theta(\omega_2))| = 0$, a violation of the rank condition. In fact, Q_n has multiple local minima: an indication that the rank conditions do not hold.

Figure B.5.5: Non-convexity and the rank condition



Legend: One-dimensional plots over $\theta(\omega) = \omega\tilde{\theta}_n + (1-\omega)\hat{\theta}_n$, where $\omega \in [0,1]$, $\hat{\theta}_n = (3.20, -0.34, 0.54)$, $\tilde{\theta}_n = (9.67, -0.19, 0.47)$. The GMM objective (panel a) is non-convex, the sample moments (panel b) do not satisfy the rank condition.

B.6 ADDITIONAL MATERIAL FOR ALGORITHMS

B.6.1 General overview of Algorithms properties

The following describes three of the algorithms in Table B.3.1: Nelder-Mead, Grid Search, Multi-Start, and Simulated Annealing. The goal is to give a brief overview of their known convergence properties; further description for each method is given in Appendix B.6.

Notation: Q_n is a continuous objective function to be minimized over Θ, a convex and compact subset of $\mathbb{R}^p$, $p \geq 1$, $\hat{\theta}_n$ denotes the solution to this minimization problem.

Nelder-Mead. Also called the simplex algorithm, the Nelder & Mead (1965, NM) algorithm comes out as a standard choice for empirical work in our survey. Notably, it was used in Berry et al. (1995, Sec6.5) to estimate the BLP model for the automobile industry. Its main feature is that it can be used even if Q_n is not continuous. It is often referred to as a *local derivative-free* optimizer. It belongs to the direct search family, which includes pattern search seen in Table B.3.1 above.

Despite being widely used, formal convergence results for the simplex algorithm are few. Notably, Lagarias et al. (1998) proved convergence for strictly convex continuous functions for $p = 1$, and a smaller class of functions for $p = 2$ parameters. McKinnon (1998) gave counter-examples for $p = 2$ of smooth, strictly convex functions for which the algorithm converges to a point that is neither a local nor a global optimum, i.e. does not satisfy a first-order condition.[8] Using the algorithm once may not produce consistent estimates in well-behaved problems so

---

[8] Powell (1973) gives additional counter-examples for the class of direct search algorithms which includes NM and Pattern Search.

it is sometimes combined with a multiple starting value strategy, described below. The TIKTAK Algorithm of Arnoud et al. (2019) builds on NM with multiple starting values. Despite these potential limitations, NM remains popular in empirical work.

Grid-Search. As the name suggests, a grid-search returns the minimizer of Q_n over a finite grid of points. In Economics, it is sometimes used to estimate models where the number of parameters p is not too large. One notable example is Donaldson (2018), who estimates $p = 3$ non-linear coefficients in a gravity model.

Contrary to NM above, grid-search has global convergence guarantees. However, convergence is very slow. Suppose we want the minimizer $\tilde{\theta}_k$ over a grid of k points to satisfy: $Q_n(\tilde{\theta}_k) - Q_n(\hat{\theta}_n) \leq \varepsilon$. Then the search requires at least $k \geq C\varepsilon^{-p}$ grid points where C depends on Q_n and the bounds used for the grid. Suppose $C = 1$, $p = 3$, $\varepsilon = 10^{-2}$, at least $k \geq 10^6$ grid points are needed, which is quite large. If each moment evaluation requires 45s, as in Lise & Robin (2017), this translates into 1.5 years of computation time.

Simulated Annealing. Unlike the methods above, Simulated Annealing (SA) is not a deterministic but a Monte Carlo based optimization method. Along with NM, SA stands out as the standard choice in empirical work. Like the grid-search, SA is guaranteed to converge, with high probability, as the number of iterations increases for an appropriate choice of tuning parameters. The main issue is that tuning parameters for which convergence results have been established result in very slow convergence: $\|\theta_k - \hat{\theta}_n\| \leq O_p(1/\sqrt{\log[k]})$, after k iterations. As a result, SA could - in theory - converge more slowly than a grid-search. Chernozhukov & Hong (2003) consider the frequentist properties of a GMM-based quasi-Bayesian posterior distribution. Draws can be sampled using the random-walk Metropolis-

Hastings algorithm, which is closely related to SA.

Multiple Starting Values. To accommodate some of the limitations of optimizers, especially the lack of global convergence guarantees, it is common to run a given algorithm with multiple starting values. Setting the starting values is similar to choosing a grid for a grid-search. Andrews (1997) provides a stopping rule which can be used to determine if sufficiently many starting values were used or not. The required number of starting values depends on the objective function Q_n, the choice of the optimizer, and the properties of the sequence used to generate starting values.

B.6.2 Implementation of the algorithms

The Nelder-Mead algorithm. The following description of the algorithm is based on Nash (1990, Ch14) which R implements in the optimizer *optim*. The first step is to build a simplex for the p-dimensional parameters, i.e. $p+1$ distinct points $\theta_1, \dots, \theta_{p+1}$ ordered s.t. $Q_n(\theta_1) \leq \cdots \leq Q_n(\theta_{p+1})$. The simplex is then transformed at each iteration using four operations called *reflection*, *expansion*, *reduction*, and *contraction*. The algorithm also repeatedly computes the centroid θ_c of the best p points, to do so: take the best p guesses $\theta_1, \dots, \theta_p$ and compute their average: $\theta_c = 1/p \sum_{\ell=1}^{p} \theta_\ell$. Once this is done, go to step **R** below.

Nelder-Mead Algorithm:

Inputs: Initial simplex $\theta_1, \ldots, \theta_{p+1}$, parameters $\alpha, \gamma, \beta, \beta'$. NM suggest to use $\alpha = 1, \gamma = 2, \beta = \beta' = 1/2$.

Re-order the points so that $Q_n(\theta_1) \leq \cdots \leq Q_n(\theta_{p+1})$, compute the centroid $\theta_c = 1/p \sum_{\ell=1}^{p} \theta_\ell$ (average of the best p points)

Start at **R** and run until convergence:

R: The *reflection* step computes $\theta_r = \theta_c + \alpha(\theta_c - \theta_{p+1}) = 2\theta_c - \theta_{p+1}$ for $\alpha = 1$. There are now several possibilities:

- If $Q_n(\theta_r) < Q_n(\theta_1)$ got to step **E**.
- If $Q_n(\theta_1) \leq Q_n(\theta_r) \leq Q_n(\theta_p)$, replace θ_{p+1} with θ_r, re-order the points, compute the new θ_c, and do **R** again.
- By elimination: $Q_n(\theta_r) > Q_n(\theta_p)$. If $Q_n(\theta_r) < Q_n(\theta_{p+1})$, replace θ_{p+1} with θ_r. Either way, go to step **R′**.

E: The *expansion* step computes $\theta_e = \theta_r + (\gamma - 1)(\theta_r - \theta_c) = 2\theta_r - \theta_c$ for $\gamma = 2$. If $Q_n(\theta_e) < Q_n(\theta_r)$, then θ_e replaces θ_{p+1}. Otherwise, θ_r replaces θ_{p+1}. Once θ_{p+1} is replaced, re-order the points, compute the new θ_c, and go to **R**.

R′: The *reduction* step computes $\theta_s = \theta_c + \beta(\theta_{p+1} - \theta_c) = (\theta_c + \theta_{p+1})/2$ for $\beta = 1/2$. If $Q_n(\theta_s) < Q_n(\theta_{p+1})$, θ_s replaces θ_{p+1}, then re-order the points, compute the new θ_c, and go to **R**. Otherwise, go to **C**.

C: The *contraction* step updates $\theta_2, \ldots, \theta_{p+1}$ using $\theta_\ell = \theta_1 + \beta'(\theta_\ell - \theta_1) = (\theta_\ell + \theta_1)/2$ for $\beta' = 1/2$. Re-order the points, compute the new θ_c, and go to **R**.

Clearly, the choice of initial simplex can affect the convergence of the algorithm.

Typically, one provides a starting value θ_1 and then the software picks the remaining p points of the simplex without user input. NM proposed their algorithm with statistical estimation in mind, so they considered using the standard deviation $\sqrt{\sum_{\ell=1}^{n+1}(Q_n(\theta_\ell) - \bar{Q}_n)^2/n} <$ tol as a convergence criterion, setting tol $= 10^{-8}$ and $\bar{Q}_n$ the average of $Q_n(\theta_\ell)$ in their application. Here convergence occurs when the simplex collapses around a single point.

The Grid-Search algorithm. The procedure is very simple, pick a grid of k points $\theta_1, \dots, \theta_k$, and compute:

$$\tilde{\theta}_k = \text{argmin}_{\ell=1,\dots,k} Q_n(\theta_\ell).$$

The optimization error $\|\tilde{\theta}_k - \hat{\theta}_n\|$ depends on both k and the choice of grid. The following gives an overview of the approximation error and feasible error rates.

For simplicity, suppose that the parameter space is the unit ball in $\mathbb{R}^p$: $\Theta = \mathcal{B}_2^p$, and Q_n is continuous. Under these assumptions, there is an $L \geq 0$ such that $|Q_n(\theta_1) - Q_n(\theta_2)| \leq L\|\theta_1 - \theta_2\|$. $L > 0$, unless Q_n is constant. This implies: $|Q_n(\tilde{\theta}_k) - Q_n(\hat{\theta}_n)| \leq L(\inf_{1 \leq \ell \leq k} \|\theta_\ell - \hat{\theta}_n\|)$. Suppose we want to ensure $|Q_n(\tilde{\theta}_k) - Q_n(\hat{\theta}_n)| \leq \varepsilon$, then we need $\inf_{1 \leq \ell \leq k} \|\theta_\ell - \hat{\theta}_n\| \leq \varepsilon/L$. Packing arguments (e.g. Vershynin, 2018, Proposition 4.2.12) give a lower bound for k over all grids, and all possible $\hat{\theta}_n$: $k \geq \text{vol}(\mathcal{B}_2^p)/\text{vol}([\varepsilon/L]\mathcal{B}_2^p) = [\varepsilon/L]^{-p}$, where vol is the volume.

For the choice of grid, Niederreiter (1983, Theorem 3) shows that low-discrepancy sequences, e.g. the Sobol or Halton points sets, can achieve this rate, up to a logarithmic term.[9] This is indeed a common choice for multi-start and grid search optimization.

---

[9] In comparison, using uniform random draws in a grid search would require $O([\varepsilon/L]^{-2p})$ iterations to achieve the same level of accuracy with high-probability. Fang & Wang (1993, Ch3.1) give a review of these results.

In practice, $Q_n(\tilde{\theta}_k) - Q_n(\hat{\theta}_n)$ is typically not the quantity of interest for empirical estimations, rather we are interested in $\|\tilde{\theta}_k - \hat{\theta}_n\|$. Suppose, in addition, that $\hat{\theta}_n \in \text{int}(\Theta)$, and Q_n is twice continuously differentiable with positive definite Hessian $H_n(\hat{\theta}_n)$, a local identification condition. Then there exists $0 < \underline{\lambda} \leq \overline{\lambda} < \infty$ and $\varepsilon_1 > 0$ s.t. $\|\theta - \hat{\theta}_n\| \leq \varepsilon_1$ implies:

$$\underline{\lambda}\|\theta - \hat{\theta}_n\|^2 \leq Q_n(\theta) - Q_n(\hat{\theta}_n) \leq \overline{\lambda}\|\theta - \hat{\theta}_n\|^2, \tag{B.6.11}$$

i.e. Q_n is locally strictly convex.[10] If $\hat{\theta}_n$ is the unique minimizer of Q_n, there is a $0 < \varepsilon_2 \leq \varepsilon_1$ such that $\inf_{\|\theta - \hat{\theta}_n\| \geq \varepsilon_1} Q_n(\theta) > Q_n(\hat{\theta}_n) + \overline{\lambda}\varepsilon_2^2$, using a global identification condition. Now, by local identification: $\|\theta - \hat{\theta}_n\| \leq \varepsilon_2 \Rightarrow Q_n(\theta) \leq Q_n(\hat{\theta}_n) + \overline{\lambda}\varepsilon_2^2 < \inf_{\|\theta - \hat{\theta}_n\| \geq \varepsilon_1} Q_n(\theta)$. As soon as $k \geq k_0$ where $\inf_{1 \leq \ell \leq k_0} \|\theta_\ell - \hat{\theta}_n\| \leq \varepsilon_2$, we have $\|\tilde{\theta}_k - \hat{\theta}_n\| \leq \varepsilon_1$. Then, for any $k \geq k_0$: $\underline{\lambda}\|\tilde{\theta}_k - \hat{\theta}_n\|^2 \leq Q_n(\tilde{\theta}_k) - Q_n(\hat{\theta}_n) \leq \overline{\lambda}(\inf_{1 \leq \ell \leq k} \|\theta_\ell - \hat{\theta}_n\|^2)$ and $\|\tilde{\theta}_k - \hat{\theta}_n\| \leq [\overline{\lambda}/\underline{\lambda}]^{1/2}(\inf_{1 \leq \ell \leq k} \|\theta_\ell - \hat{\theta}_n\|)$.

This reveals the interplay between the identification conditions and the optimization error. The best value $\tilde{\theta}_k$ is only guaranteed to be near $\hat{\theta}_n$ when $k \geq \varepsilon_2^{-p}$ iterations (using packing arguments for the unit ball), where ε_2 depends on the global identification condition. Local convergence depends on the ratio $\overline{\lambda}/\underline{\lambda} \geq 1$ which is infinite when $H_n(\hat{\theta}_n)$ is singular. The main drawback of a grid search is its slow convergence. To illustrate, Colacito et al. (2018, pp3443-3445) estimate $p = 5$ parameters using a grid search with $k = 1551$ points. For simplicity, suppose $\overline{\lambda}/\underline{\lambda} = 1$, $k_0 < k$, and $\Theta = \mathcal{B}_2^p$, the unit ball, then the worst-case optimization error is $\sup_{\hat{\theta}_n \in \Theta}(\inf_{1 \leq \ell \leq k} \|\theta_\ell - \hat{\theta}_n\|) \geq k^{-1/p} \simeq 0.23$. This is ten times larger than all but one of the standard errors reported in the chapter.

---

[10] The three $\varepsilon_1, \underline{\lambda}, \overline{\lambda}$ only depend on $H_n(\cdot)$.

Simulated Annealing. Implementations can vary across software, the following will focus on the implementation used in R's *optim* function.

```
Simulated Annealing Algorithm:

    Inputs: Starting value $\theta_1 \in \Theta$, temperature schedule $\infty > T_2 \geq T_3 \geq \cdots > 0$,
    a sequence $\infty > \eta_2 \geq \eta_3 \geq \cdots > 0$, and maximum number of iterations $k$.
    Common choice: $T_\ell = T_1/\log(\ell)$ for $\ell \geq 2$ and $\eta_\ell$ proportional to $T_\ell$.

    For $\ell \in \{2, \dots, k\}$, repeat:

        1. Draw $\theta^\star \sim \mathcal{N}(\theta_{\ell-1}, \eta_\ell I_d)$, and $u_\ell \sim \mathcal{U}_{[0,1]}$
        2. Set $\theta_\ell = \theta^\star$ if $u_\ell \leq \exp(-[Q_n(\theta^\star) - Q_n(\theta_{\ell-1})]/T_\ell)$, otherwise set $\theta_\ell = \theta_{\ell-1}$

    Output: Return $\tilde{\theta}_k = \text{argmin}_{1 \leq \ell \leq k} Q_n(\theta_\ell)$
```

The implementation described above relies on the random-walk Metropolis update. Notice that if $Q_n(\theta^\star) \leq Q_n(\theta_{\ell-1})$, the exponential term in step 2 is greater than 1 and $\theta^\star$ is always accepted as the next θ_ℓ, regardless of u_ℓ. Bélisle (1992) gave sufficient condition for $\tilde{\theta}_k \overset{a.s.}{\to} \hat{\theta}_n$ when $k \to \infty$ and Q_n is continuous. In practice, the performance of the Algorithm can be measured by its convergence rate. To get some intuition, we give some simplified derivations below which highlight the role of T_k and several quantities which appeared in our discussion of the grid search.

First, notice that for each k, steps 1-2 implement the Metropolis algorithm also used for Bayesian inference using random-walk Metropolis-Hastings. The invariant distribution of these two steps is:

$$f_k(\theta) = \frac{\exp(-[Q_n(\theta) - Q_n(\hat{\theta}_n)]/T_k)}{\int_\Theta \exp(-[Q_n(\theta) - Q_n(\hat{\theta}_n)]/T_k)d\theta},$$

this is called the Gibbs-Boltzmann distribution. When $T_\infty = +\infty$, f_∞ puts all the probability mass on the unique minimum $\hat{\theta}_n$. To build intuition, suppose that $k \geq 1$: $\theta_k \sim f_k$. Because SA is a stochastic algorithm, the approximation error $\|\theta_k - \hat{\theta}_n\|$ is random, but can be quantified using $\mathbb{P}(\|\theta_k - \hat{\theta}_n\| \geq \varepsilon)$. In the following we will assume the temperature schedule to be $T_k = T_1/\log(k)$, as implemented in R.

The following relies on the same setting, notation and assumptions as the grid search above. First, we can bound the probability that θ_k is outside the ε_1-local neighborhood of $\hat{\theta}_n$ where Q_n is approximately quadratic: $\mathbb{P}(\|\theta_k - \hat{\theta}_n\| \geq \varepsilon_1)$. Using the global identification condition:

$$\exp(-[Q_n(\theta) - Q_n(\hat{\theta}_n)]/T_k) \leq \exp(-\overline{\lambda}\varepsilon_2^2/T_k) = k^{-\overline{\lambda}\varepsilon_2^2/T_1}, \text{ if } \|\theta - \hat{\theta}_n\| \geq \varepsilon_1,$$

where ε_1, ε_2 were defined in the grid search section above. This gives an upper bound for the numerator in $f_k(\theta_k)$. A lower bound is also required for the denominator. Using (B.6.11) and the change of variable $\theta = \hat{\theta}_n + \sqrt{T_k}h$, we have:

$$\begin{aligned}\exp(-\overline{\lambda}\|h\|^2) &\leq \exp(-[Q_n(\hat{\theta}_n + \sqrt{T_k}h) - Q_n(\hat{\theta}_n)]/T_k) \\ &\leq \exp(-\underline{\lambda}\|h\|^2), \text{ if } \|\sqrt{T_k}h\| \leq \varepsilon_1.\end{aligned}$$

Suppose $T_k \leq \varepsilon_1^2$, the two inequalities give us the bound:

$$\mathbb{P}(\|\theta_k - \hat{\theta}_n\| \geq \varepsilon_1) \leq \frac{k^{-\overline{\lambda}\varepsilon_2^2/T_1}\text{vol}(\Theta)}{|T_k|^{p/2}\int_{\|h\|\leq 1}\exp(-\overline{\lambda}\|h\|^2)dh} = C[\log(k)]^{d/2}k^{-\overline{\lambda}\varepsilon_2^2/T_1}.$$

This upper bound declines more slowly than for the grid search when $\overline{\lambda}\varepsilon_2^2/T_1 < 1/p$, which can be the case if T_1 large and/or ε_2 is small. For the lower bound, pick any $\varepsilon \in (0, \varepsilon_1/\sqrt{T_k})$:

$$\mathbb{P}(\|\theta_k - \hat{\theta}_n\| \leq \sqrt{T_k}\varepsilon) \geq \frac{\int_{\|h\|\leq\varepsilon}\exp(-\overline{\lambda}\|h\|^2)dh}{\int_{\|h\|\in\mathbb{R}}\exp(-\underline{\lambda}\|h\|^2)dh + |T_k|^{-p/2}\text{vol}(\Theta)k^{-\overline{\lambda}\varepsilon_2^2/T_1}},$$

which has a strictly positive limit. This implies that $\sqrt{\log(k)}\|\theta_k - \hat{\theta}_n\| \geq O_p(1)$, since $T_k = T_1/\log(k)$. This $\sqrt{\log(k)}$ rate is slower than the grid search. To get faster convergence, some authors have suggested using $T_k = T_1/k$ and, by default, Matlab sets $T_k = T_1 \cdot 0.95^k$. However, theoretical guarantees to have $\theta_k \xrightarrow{p} \hat{\theta}_n$, as $k \to \infty$ are only available when $T_k = T_1/\log(k)$.[11]

---

[11] See Spall (2005, Ch8.4-8.6) for additional details and references.

APPENDIX C

Supplementary Materials for Chapter Three

C.1 PROOFS

Lemma 1: (Inverse Mills ratio). If X is a normally distributed random variable with mean μ and variance σ^2, then

$$E(X \mid X > \alpha) = \mu + \sigma \frac{\phi(\frac{\alpha-\mu}{\sigma})}{1 - \Phi(\frac{\alpha-\mu}{\sigma})}$$

where ϕ and Φ are the p.d.f. and c.d.f. of the Normal, respectively.

Proof of Proposition 2:

Part 1: mean and variance of box-office revenue conditional on production

(i) Given two normal distributions $\pi \mid t$ and $y|\pi, t$, $f(\pi|y,t) \propto f(y \mid \pi, t) f(\pi \mid t)$. Hence

$$\pi|y,t \sim N(E(\pi|y,t), Var(\pi|y,t)), \quad y|t \sim N(\mu_t, \sigma^2_{\pi t} + \sigma^2_{yt})$$

where:

$$E(\pi|y,t) = \frac{\sigma^2_{\pi t}}{\sigma^2_{\pi t} + \sigma^2_{yt}} y + \frac{\sigma^2_{yt}}{\sigma^2_{\pi t} + \sigma^2_{yt}} \mu_t \sim N(\mu_t, \frac{\sigma^4_{\pi t}}{\sigma^2_{\pi t} + \sigma^2_{yt}})$$

$$Var(\pi|y,t) = \frac{\sigma^2_{\pi t}\sigma^2_{yt}}{\sigma^2_{\pi t} + \sigma^2_{yt}}$$

With an abuse of notation, denote π_t as $\pi \mid t$, and y_t as $y \mid t$, then by Lemma 1 and the law of total expectation:

$$\begin{aligned} E(\pi_t|y_t > \bar{y}_t) &= E(\pi_t \mid E(\pi_t \mid y_t) > \pi_0) \\ &= E(E(\pi_t \mid y_t) \mid E(\pi_t \mid y_t) > \pi_0) \\ &= \mu_t + \sigma \frac{\phi(\frac{\pi_0 - \mu_t}{\sigma})}{1 - \Phi(\frac{\pi_0 - \mu_t}{\sigma})} (3) \end{aligned}$$

where $\sigma^2 = \frac{\sigma^4_{\pi_t}}{\sigma^2_{\pi_t} + \sigma^2_{y_t}}$

END.

(ii) Now for variance:

$$\begin{aligned} Var(\pi_t|y_t > \bar{y}_t) &= Var(\pi_t|E(\pi_t|y_t) > \pi_0) \\ &= E(\pi_t^2|E(\pi_t|y_t) > \pi_0) - E^2(\pi_t|E(\pi_t|y_t) > \pi_0) \\ &= E(E(\pi_t^2|y_t)|E(\pi_t|y_t) > \pi_0) - E^2(\pi_t|E(\pi_t|y_t) > \pi_0) \\ &= E([Var(\pi_t|y_t) + E^2(\pi_t|y_t)]|E(\pi_t|y_t) > \pi_0) \\ &- E^2(E(\pi_t|y_t)|E(\pi_t|y_t) > \pi_0) \\ &= \frac{\sigma^2_{\pi_t}\sigma^2_{y_t}}{\sigma^2_{\pi_t} + \sigma^2_{y_t}} + E\Big(E^2(\pi_t|y_t)|E(\pi_t|y_t) > \pi_0\Big) \\ &- E^2\Big(E(\pi_t|y_t)|E(\pi_t|y_t) > \pi_0\Big)(4) \end{aligned}$$

For a standard normal distribution, $z \sim N(0,1)$.

$$\begin{aligned}
E(z^2|z>c) &= \frac{1}{1-\Phi(c)}\int\limits_c^{\infty}\frac{z^2}{\sqrt{2\pi}}exp\Big(-\frac{z^2}{2}\Big)dz \\
&= \frac{1}{1-\Phi(c)}\int\limits_c^{\infty}\Bigg(\frac{1}{\sqrt{2\pi}}exp\Big(-\frac{z^2}{2}\Big)-\Big(\frac{z}{\sqrt{2\pi}}exp\Big(-\frac{z^2}{2}\Big)\Big)'\Bigg)dz \\
&= \frac{1}{1-\Phi(c)}\int\limits_c^{\infty}\Bigg(\frac{1}{\sqrt{2\pi}}exp\Big(-\frac{z^2}{2}\Big)-\Big(\frac{z}{\sqrt{2\pi}}exp\Big(-\frac{z^2}{2}\Big)\Big)'\Bigg)dz \\
&= \frac{1}{1-\Phi(c)}\int\limits_c^{\infty}\Bigg(\frac{1}{\sqrt{2\pi}}exp\Big(-\frac{z^2}{2}\Big)-\Big(\frac{z}{\sqrt{2\pi}}exp\Big(-\frac{z^2}{2}\Big)\Big)'\Bigg)dz \\
&= 1+\frac{c\phi(c)}{1-\Phi(c)}
\end{aligned}$$

So, for $x \sim N(\mu, \sigma^2)$

$$\begin{aligned}
1+\frac{\frac{c-\mu}{\sigma}\phi(\frac{c-\mu}{\sigma})}{1-\Phi(\frac{c-\mu}{\sigma})} &= E\Bigg(\Big(\frac{x-\mu}{\sigma}\Big)^2|\frac{x-\mu}{\sigma}>\frac{c-\mu}{\sigma}\Bigg) \\
&= \frac{1}{\sigma^2}\Big(E(x^2|x>c)-2\mu E(x|x>c)+\mu^2\Big)
\end{aligned}$$

Combining with

$$E(x|x>c)=\mu+\sigma\frac{\phi(\frac{c-\mu}{\sigma})}{1-\Phi(\frac{c-\mu}{\sigma})}$$

we obtain

$$E(x^2|x>c)=\sigma^2+\sigma^2\frac{\frac{c-\mu}{\sigma}\phi(\frac{c-\mu}{\sigma})}{1-\Phi(\frac{c-\mu}{\sigma})}+\mu^2+2\mu\sigma\frac{\phi(\frac{c-\mu}{\sigma})}{1-\Phi(\frac{c-\mu}{\sigma})}$$

Plugging in (4) yields

$$\begin{aligned}
Var(\pi_t|E(\pi_t|y_t)>\pi_0) &= \frac{\sigma_{\pi_t}^2\sigma_{y_t}^2}{\sigma_{\pi_t}^2+\sigma_{y_t}^2}+\sigma^2+\sigma^2\frac{\frac{\pi_0-\mu_t}{\sigma}\phi(\frac{\pi_0-\mu_t}{\sigma})}{1-\Phi(\frac{\pi_0-\mu_t}{\sigma})} \\
&\quad -\sigma^2\Bigg(\frac{\phi(\frac{\pi_0-\mu_t}{\sigma})}{1-\Phi(\frac{\pi_0-\mu_t}{\sigma})}\Bigg)^2 \qquad (I)
\end{aligned}$$

Then, $(I) = \sigma^2_{\pi_t} + \sigma^2(x\lambda(x) - \lambda^2(x))$, by $\sigma^2 = \frac{\sigma^4_{\pi_t}}{\sigma^2_{\pi_t}+\sigma^2_{y_t}}, x = \frac{\pi_0-\mu_t}{\sigma}, \lambda(x) = \frac{\phi(x)}{1-\Phi(x)}$

END.

Part 2: comparative statics

Building blocks

Lemma 2: For $\lambda(x) = \frac{\phi(x)}{1-\Phi(x)}$, $\frac{3x+\sqrt{x^2+8}}{4} < \lambda(x) < \frac{x+\sqrt{x^2+4}}{2}$ for $x \in R$.

Proof: Normally, a computer can confirm this lemma. However, when $x > 7$, both the numerator and the denominator of λ are so close to 0 that the value for λ is heavily biased. Hence, this proof will only target the case where $x > 7$.

First, taking the first derivative of $\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$ yields $\phi^{'}(x) = -x\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} = -x\phi(x)$. It follows that

$$
\begin{aligned}
1-\Phi(x) &= \int\limits_x^\infty \phi(u)du \\
&= -\int\limits_x^\infty \frac{\phi'(u)}{u}du \\
&= \frac{\phi(x)}{x} - \frac{\phi(x)}{x^3} + \frac{3\phi(x)}{x^5} - \frac{15\phi(x)}{x^7} + \int\limits_x^\infty \frac{105\phi(u)}{u^8}du \\
&= \frac{\phi(x)}{x} - \frac{\phi(x)}{x^3} + \frac{3\phi(x)}{x^5} - \frac{15\phi(x)}{x^7} + \frac{105\phi(x)}{x^9} - \int\limits_x^\infty \frac{945\phi(u)}{u^{10}}du
\end{aligned}
$$

Then

$$\frac{1}{\frac{1}{x} - \frac{1}{x^3} + \frac{3}{x^5} - \frac{15}{x^7} + \frac{105}{x^9}} < \frac{\phi(x)}{1 - \Phi(x)} < \frac{1}{\frac{1}{x} - \frac{1}{x^3} + \frac{3}{x^5} - \frac{15}{x^7}} \qquad when \quad x > 7$$

Let the left (right) term of the inequality be denoted as LHS (RHS). We first prove that when $x > 7$, $LHS > \frac{3x+\sqrt{x^2+8}}{4}$. Assume that this is true. Then

$$LHS > \frac{3x + \sqrt{x^2+8}}{4}$$

$$\iff (x^9 + 3x^7 - 9x^5 + 45x^3 - 315x)^2 > (x^2+8)(x^8 - x^6 + 3x^4 - 15x^2 + 105)^2$$

$$\iff x^{18} + 6x^{16} - 9x^{14} + 36x^{12} - 279x^{10} - 27000x^8 + 7695x^6 - 28350x^4 + 99225x^2 >$$

$$x^{18} + 6x^{16} - 9x^{14} + 20x^{12} - 39x^{10} + 1692x^8 - 1545x^6 + 3690x^4 - 14175x^2 + 88200$$

$$\iff 16x^{12} - 240x^{10} - 4392x^8 + 9240x^6 - 32040x^4 + 113400x^2 - 88200 > 0$$

Then for $x > 7$, $16x^{12} > 16 * 7^2 x^{10}$, i.e. $16x^{12} > 784x^{10}$, which is true.

We now prove that when $x > 7$, $RHS < \frac{x+\sqrt{x^2+4}}{2}$

$\iff (x^7 + x^5 - 3x^3 + 15x)^2 < (x^2+4)(x^6 - x^4 + 3x^2 - 15)^2$

$\iff x^{14} + 2x^{12} - 5x^{10} + 24x^8 + 39x^6 - 90x^4 + 225x^2 <$

$x^{14} + 2x^{12} - x^{10} - 8x^8 - 105x^6 + 66x^4 - 135x^2 + 900$

$\iff x^{10} - 8x^8 - 36x^6 + 39x^4 - 90x^2 + 225 > 0$

Then for $x > 7$, $x^{10} > 7^2 x^8$, i.e. $x^{10} > 49x^8$, which is also true.

A computer can easily confirm that the lemma holds also for $x < 7$, which completes the proof. In addition, for $x > 0$, we can show that $x < \frac{3x+\sqrt{x^2+8}}{4} < \lambda(x) < \frac{x+\sqrt{x^2+4}}{2} < x + \frac{1}{x}$.

END.

Comparative statics 2(a):

(i) Let $x = \frac{\pi_0 - \mu}{\sigma}$. Then

$$\frac{\mathrm{d}(3)}{\mathrm{d}\mu} = 1 + \sigma\lambda'(x) = 1 + \sigma(-x\lambda(x) + \lambda^2(x))\Big(-\frac{1}{\sigma}\Big)$$
$$= -\left(\lambda(x) - \frac{x+\sqrt{x^2+4}}{2}\right)\left(\lambda(x) - \frac{x-\sqrt{x^2+4}}{2}\right) \qquad (i)$$

By Lemma 2, $\frac{\mathrm{d}(3)}{\mathrm{d}\mu} > 0 \; \forall x \in R$.

END.

(ii) Again let $x = \frac{\pi_0 - \mu}{\sigma}$. Then

$$\frac{\mathrm{d}(I)}{\mathrm{d}\mu} = \sigma^2(\lambda(x) + x\lambda'(x) - 2\lambda(x)\lambda'(x))\Big(-\frac{1}{\sigma}\Big)$$
$$= -\sigma\Big(\lambda(x) - x^2\lambda(x) + 3x\lambda^2(x) - 2\lambda^3(x)\Big)$$
$$= \sigma\lambda(x)\left(\lambda(x) - \frac{3x+\sqrt{x^2+8}}{4}\right)\left(\lambda(x) - \frac{3x-\sqrt{x^2+8}}{4}\right)$$

By Lemma 2, $\frac{\mathrm{d}(I)}{\mathrm{d}\mu} > 0 \; \forall x \in R$.

END.

Comparative statics 2(b):

(i)

$$\frac{\mathrm{d}(3)}{\mathrm{d}\pi_0} = \sigma\lambda'(x) = \sigma(-x\lambda(x) + \lambda^2(x))(\frac{1}{\sigma}) = (-x + \lambda(x))\lambda(x) > 0$$

.

END.

(ii) See the proof for 2(a), (ii).

END.

Comparative statics 2(c):

(i)

$$\frac{\mathrm{d}(3)}{\mathrm{d}\sigma} = \sigma\lambda'(x)+\lambda(x) = \Big(-x\lambda(x)+\lambda^2(x)\Big)\left(-\frac{y_0-\mu_t}{\sigma}\right)+\lambda(x) = \lambda(x)\Big(1+x^2-x\lambda(x)\Big) \tag{4}$$

where, again, $\sigma = \frac{\sigma^2_{\pi t}}{\sqrt{\sigma^2_{\pi t}+\sigma^2_{yt}}}$, $x = \frac{\pi_0-\mu_t}{\sigma}$, $x > 0$, and $\lambda(x) = \frac{\phi(x)}{1-\Phi(x)}$.

When $x > 0$, we can write $(4) = x\lambda(x)(\frac{1}{x} + x - \lambda(x))$. By Lemma 2 and $\frac{x+\sqrt{x^2+4}}{2} < x + \frac{1}{x}$, $(4) > 0$ holds.

When $x < 0$, $(4) > 0$ clearly holds.

Hence, if σ^2_{yt} increases, then σ^2 decreases and (3), i.e. E_t decreases too.

END.

(ii)

$$\begin{aligned}
\frac{\mathrm{d}(I)}{\mathrm{d}\sigma^2_{yt}} &= \sigma^4_{yt}\frac{\Big(x\lambda(x)-\lambda^2(x)\Big)'(\sigma^2_{\pi t}+\sigma^2_{yt}) - \Big(x\lambda(x)-\lambda^2(x)\Big)}{(\sigma^2_{\pi t}+\sigma^2_{yt})^2} \\
&= \sigma^4_{\pi t}\left(\frac{\lambda(x)\Big(1-x^2+3x\lambda(x)-2\lambda^2(x)\Big)}{\sigma^2_{\pi t}+\sigma^2_{yt}}\Big(-\frac{\pi_0-\mu}{\sigma^2}\Big)\Big(-\frac{\sigma^2_{\pi t}}{2(\sigma^2_{\pi t}+\sigma^2_{yt})^{\frac{3}{2}}}\Big)\right. \\
&\quad \left. -\frac{x\lambda(x)-\lambda^2(x)}{(\sigma^2_{\pi t}+\sigma^2_{yt})^2}\right) \\
&= -2x\frac{\lambda(x)\sigma^4_{\pi t}}{2(\sigma^2_{\pi t}+\sigma^2_{yt})^2}\left(\lambda(x)-\frac{x}{2}\right)\left(\lambda(x)-x-\frac{1}{x}\right)
\end{aligned}$$

When $x > 0$, $x < \lambda(x) < x + 1/x$, which implies that $\frac{\mathrm{d}(I)}{\mathrm{d}\sigma_{yt}^2} > 0$.

When $x < 0$, it is easy to see that $\frac{\mathrm{d}(I)}{\mathrm{d}\sigma_{yt}^2} > 0$.

Hence, if σ_{yt}^2 increases, (I), i.e. Var_t increases too.

END.

C.2 ALTERNATIVE DISTRIBUTIONAL ASSUMPTIONS

In this appendix, we explore the robustness of our results to alternative distributional assumptions.

C.2.1 Beta-Binomial distribution

C.2.1.1 Setup

We first consider the case where the producer cares only about a binary criterion, whether the movie will be a "hit" or not.

Assume the object of interest is p, the probability that the movie is a hit. The prior distribution of p is Beta with parameters α and β:

$$p \sim Beta(\alpha, \beta)$$

Therefore:

$$\begin{aligned}
f(p) &\propto p^{\alpha-1}(1-p)^{\beta-1} \\
E(p) &= \frac{\alpha}{\alpha+\beta} \\
V(p) &= \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}
\end{aligned}$$

Producers observe a signal y, which, conditional on the true p, is distributed binomial with parameters n and p. We can think of this as the producer consulting with n critics, and each one independently assessing whether the movie will be a hit or not, with probability p.

$$f(y|p) = \binom{n}{y} p^y (1-p)^{(n-y)}$$

It follows that the posterior density of p given y is

$$f(p|y) \propto p^{\alpha-1}(1-p)^{\beta-1} p^y (1-p)^{n-y}$$
$$\Rightarrow p|y \sim Beta(\alpha + y, \beta + n - y)$$

Therefore,

$$E(p|y) = \frac{\alpha + y}{\alpha + y + \beta + n - y} = \frac{\alpha + y}{\alpha + \beta + n}$$

The producer will produce the movie if $E(p|y) > p_0$, for some predetermined p_0. Therefore, the signal threshold for production $\bar{y}$ is:

$$\bar{y} \equiv p_0(\alpha + \beta + n) - \alpha.$$

It is convenient to use a reparametrization, letting $\kappa = \alpha + \beta$. If $\alpha, \beta > 1$, then κ captures the spread of the distribution: for a given α, a higher value of κ means that the distribution is more concentrated, i.e., the prior is more informative.[1]

---

[1] Under this parameterization, the mean of the prior distribution is $E(p) = \frac{\alpha}{\kappa}$ and the variance is $V(p) = \frac{\alpha(\kappa - \alpha)}{\kappa^2(\kappa - 1)}$. For $\alpha, \beta > 1$, the variance is strictly decreasing in κ.

Let $s = y/n$ be the success rate of the signal. Expressing the signal threshold in terms of s, the movie is produced if and only if

$$s > \bar{s} \equiv p_0 + \frac{\kappa}{n}(p_0 - \frac{\alpha}{\kappa})$$

C.2.1.2 *Comparative statics for production*

(1) **Customer discrimination**. We interpret customer discrimination against non-white movies as $\alpha_b < \alpha_w$. That is, non white movies have a lower prior probability of being a hit. The signal threshold is decreasing in α:

$$\frac{\partial \bar{s}}{\partial \alpha} < 0.$$

Therefore, under customer discrimination, the signal threshold for non-white movies is higher than the signal threshold for white movies.

(2) **Taste-based discrimination**. We interpret taste-based discrimination against non-white movies as $p_{0b} > p_{0w}$. That is, non-white movies are held to a higher standard, and are produced only if the posterior probability of the movie being a hit exceeds a higher threshold. The signal threshold is increasing in p_0:

$$\frac{\partial \bar{s}}{\partial p_0} > 0.$$

Therefore, under taste-based discrimination, the signal threshold for non-white movies is higher than the signal threshold for white movies.

(3) **Statistical discrimination**. Finally, we interpret statistical discrimination as $n_b < n_w$. That is, the signal for non-white movies being is less informative

than that for white movies. The derivative of the signal threshold with respect to n is:

$$\frac{\partial \bar{s}}{\partial n} = -\frac{\kappa}{n^2}(p_0 - \frac{\alpha}{\kappa})$$

The sign of this derivative depends on $p_0 - \alpha/\kappa$. Under this parametrization, α/κ is the prior mean of p. In other words, we have the same qualitative result as in the log-normal model presented in the main text.

(i) If $p_0 > \alpha/\kappa$, (i.e., the producer wants to produce only movies with a very high probability of being a hit),

$$\frac{\partial \bar{s}}{\partial n} < 0;$$

that is, a less precise signal (lower n) raises the signal threshold. The signal threshold for non-white movies is higher.

(ii) If $p_0 < \alpha/\kappa$, (i.e. the producer wants to weed out the very low quality movies),

$$\frac{\partial \bar{s}}{\partial n} > 0;$$

now, a less precise signal (lower n) *lowers* the signal threshold. One can be a bit more tolerant of a bad signal for non-white movies, because it is difficult to say, based on the signal alone, whether the movie is really bad.

It is easy to see that the comparative statics with respect to the precision of the signal mirrors exactly what we had in the normal-normal case.

C.2.1.3 Comparative statics for the observed success rate, conditional on production: simulations

We only observe whether a movie is a hit, conditional on production. Therefore, as in the analysis in the main text, we need to characterize the the posterior distribution of p conditional on $s > \bar{s}$, and derive its comparative statics with respect to p_0, α and n. While it is not possible to derive an analytical solution for the comparative statics, we can proceed by simulation. Specifically, for each set of parameter values, we draw a sample of L movies, apply the production decision rule, and report the mean and standard deviation of the posterior distribution p conditional on production. The results are presented in Table B.1.

Table B.1: Simulation results: Beta-binomial distribution

A: Taste based discrimination: p_0 ↑ for Non-white movies
Fixed $\alpha = 4$, $\kappa = 8$, $n = 5$

p_0	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	Trend
mean	0.500	0.500	0.500	0.500	0.514	0.545	0.545	0.587	0.638	0.638	↑
std	0.167	0.167	0.167	0.167	0.160	0.151	0.151	0.142	0.133	0.133	↓

B: Customer discrimination: α ↓ for Non-white movies
Fixed $p_0 = 0.5$, $\kappa = 8$, $n = 5$

α	6	5.5	5	4.5	4	3.5	3	2.5	2	1.5	
mean	0.752	0.692	0.650	0.597	0.587	0.542	0.555	0.515	0.538	0.497	↓
std	0.142	0.151	0.148	0.151	0.142	0.143	0.136	0.137	0.135	0.135	↓

C1: Statistical discrimination, $p_0 > \alpha/\kappa$: n ↓ for Non-white movies
Fixed $p_0 = 0.6$, $\kappa = 8$, $\alpha = 4$

n	11	10	9	8	7	6	5	4	3	2	
mean	0.673	0.656	0.679	0.659	0.638	0.662	0.638	0.666	0.637	0.600	↓
std	0.115	0.119	0.117	0.122	0.128	0.126	0.133	0.131	0.139	0.148	↑

C2: Statistical discrimination, $p_0 < \alpha/\kappa$: n ↓ for Non-white movies
Fixed $p_0 = 0.4$, $\kappa = 8$, $\alpha = 4$

n	11	10	9	8	7	6	5	4	3	2	
mean	0.549	0.557	0.540	0.548	0.528	0.535	0.545	0.520	0.527	0.500	↓
std	0.145	0.143	0.149	0.148	0.154	0.153	0.151	0.159	0.158	0.167	↑

Legend: simulated data with sample size $L = 10^6$, using R with seed 123. Mean, Std: sample average and standard deviation of the posterior distribution of $p|s, s > \bar{s}$ from the simulation.

Each panel in the table presents a different comparative statics exercise. For example, in Panel A, to examine the role of taste-based discrimination, we fix the values of α, κ and n, and study what happens to the mean and standard deviation of the posterior distribution of p as we increase the value of p_0. The results in Panel A show that as taste-based discrimination increases, the posterior expected value of p increases, and the posterior standard deviation decreases. These results match the predictions in the normal-normal model, derived analytically.

In Panel B we look at the effect of increasing customer discrimination increases.

As in the normal model, both the expected value and the standard deviation increase as the value of α decreases.[2]

In Panels C1 and C2 we study the effect of statistical discrimination, distinguishing between the case in which the producer only wants to produce very high quality movies ($p_0 > \alpha/\kappa$) so that the signal threshold decreases in n (case 3.i in Section C.2.1.2); and the one in which the producer wants to weed out very low quality movies ($p_0 < \alpha/\kappa$) so that the signal threshold increases in n. We see that in both cases the mean of p decreases and the standard deviation increases as the extent of statistical discrimination increases (the signal becomes less prescise, or n decreases). Again, the pattern of comparative statics results mirrors exactly what we obtained in the normal-normal case (Section 3.3 in the main text).

We conclude that all of the main predictions of the theoretical model based on the normal-normal case in the main text remain identical under the beta-binomial model.

C.2.2 Pareto-Normal distribution

C.2.2.1 Setup

We now return to the case considered in the main text, where the producer cares about (log) revenue, but we now depart from the normal-normal model. Specifically, we assume that ex-ante revenue $\tilde{\pi}$ (in dollars) follows a Pareto distribution:

$$\tilde{\pi} \sim Pareto(x_m, a)$$

---

[2] In each panel, as we move in the table from left to right, we *increase* the extent of discrimination. In the case of customer discrimination (Panel B) and statistical discrimination (Panel C), an increase in discrimination implies a decrease in the parameter of interest.

where x_m is the minimum, and a is the shape parameter. The CDF is:

$$F(\tilde{\pi}) = \begin{cases} 1-(\frac{x_m}{\tilde{\pi}})^a, \textit{ if } \tilde{\pi} \geq x_m \\ 0, \textit{ if } \tilde{\pi} < x_m \end{cases}$$

Then, log revenue $\pi \equiv \log(\tilde{\pi})$ has a shifted exponential distribution: $\pi \sim Exp(a) + log(x_m)$, or, equivalently, $\log(\frac{\tilde{\pi}}{x_m}) \sim Exp(a)$.

Therefore, the pdf of π is:

$$f(\pi) = \begin{cases} a * exp(-a(\pi - log(x_m))), \textit{ if } \pi \geq log(x_m) \\ 0, \textit{ if } \pi < log(x_m) \end{cases}$$

Producers observe a signal y, which, conditional on the true π is distributed $N(\pi, \sigma_y^2)$.

$$f(y|\pi) = \frac{1}{\sigma_y\sqrt{2\pi}} exp(-\frac{1}{2}(\frac{y-\pi}{\sigma_y})^2)$$

It follows that the posterior distribution of π given y is

$$f(\pi|y) \propto exp(-a(\pi - \log(x_m)) - \frac{1}{2}(\frac{y-\pi}{\sigma_y})^2), \text{ if } \pi \geq \log(x_m)$$

When $\pi \geq log(x_m)$:

$$\begin{aligned} f(\pi|y) &\propto exp(-a(\pi - log(x_m)) - \frac{1}{2}(\frac{y-\pi}{\sigma_y})^2) \\ &= exp(-\frac{1}{2\sigma_y^2}(\pi^2 - 2y\pi + y^2 + 2\sigma_y^2 a\pi - 2\sigma_y^2 a \log(x_m))) \end{aligned}$$

$$= exp(-\frac{1}{2\sigma_y^2}((\pi - (y - a\sigma_y^2))^2 - a^2\sigma_y^4 + 2ya\sigma_y^2 - 2\sigma_y^2 a\log(x_m)))$$
$$= exp(-\frac{1}{2\sigma_y^2}((\pi - (y - a\sigma_y^2))^2)) \times exp(-a(y - \frac{a\sigma_y^2}{2} - \log(x_m)))$$

Given y, the second term is constant. Therefore, putting everything together, we have that

$$f(\pi|y) \propto exp(-\frac{1}{2\sigma_y^2}((\pi - (y - a\sigma_y^2))^2)), \text{ for } \pi > \log(x_m).$$

This implies that the posterior distribution of π given the signal y is a truncated normal derived from a normal distribution with mean $y - a\sigma_y^2$, variance σ_y^2 and lower truncation point $\log(x_m)$. The posterior mean is therefore

$$E(\pi|y) = (y - a\sigma_y^2) + \sigma_y \frac{\phi(\log(x_m))}{1 - \Phi(\log(x_m))}.$$

The producer will produce the movie if $E(\pi|y) > \pi_0$, for some predetermined π_0. Therefore, the signal threshold for production $\bar{y}$ is:

$$\bar{y} \equiv \pi_0 + a\sigma_y^2 - \sigma_y \frac{\phi(\log(x_m))}{1 - \Phi(\log(x_m))}.$$

C.2.2.2 Comparative statics for production

(1) **Customer discrimination**: The expectation of an exponential distribution with parameter a is $1/a$. Therefore, we interpret customer discrimination

against non-white movies as $a_b > a_w$. The signal threshold is increasing in a:

$$\frac{\partial \bar{y}}{\partial a} > 0$$

As in the normal-normal case, the signal threshold for non-white movies is higher than the signal threshold for white movies.

(2) **Taste-based discrimination**: We interpret taste-based discrimination against non-white movies as $\pi_{0b} > \pi_{0w}$ – non-white movies are held to a higher standard and are produced only if the posterior mean exceeds a threshold that is higher than that set for white movies. The signal threshold increases in π_0:

$$\frac{\partial \bar{y}}{\partial \pi_0} > 0.$$

Therefore, under taste-based discrimination, the signal threshold for non-white movies is higher than that for white movies. This result mirrors that of the normal-normal case.

(3) **Statistical discrimination**: We interpret statistical discrimination as $\sigma_{yb} > \sigma_{yw}$, i.e., the signal for non-white movies is less precise. The derivative of the signal threshold with respect to σ_y is:

$$\frac{\partial \bar{y}}{\partial \sigma_y} = 2a\sigma_y - \frac{\phi(\log(x_m))}{1 - \Phi(\log(x_m))}$$

The sign of this derivative depends on the magnitude of σ_y.

(i) If $\sigma_y > \frac{\phi(\log(x_m))}{2a(1-\Phi(\log(x_m)))}$, (i.e., the movie has a high variance on the potential outcome),

$$\frac{\partial \bar{y}}{\partial \sigma_y} > 0;$$

that is, a less precise signal (lager σ_y) raises the signal threshold. The signal threshold for non-white movies is higher.

(ii) If $\sigma_y < \frac{\phi(\log(x_m))}{2a(1-\Phi(\log(x_m)))}$, (i.e. the signal has good information on the movie revenue),

$$\frac{\partial \bar{y}}{\partial \sigma_y} < 0;$$

now, a less precise signal (larger σ_y) *lowers* the signal threshold. One can be a bit more tolerant of a bad signal for non-white movies, because it is difficult to say, based on the signal alone, whether the movie is really bad.

Although the signal threshold for non-white movies could still have been either higher or lower than that for white movies, the result does not depend on whether producers only want to produce very high quality movies, or they just want to weed out very low quality movies (i.e., it does not depend on whether the revenue threshold π_0 is above or below the prior mean of π,w which is in contrast to the normal-normal model.

C.2.2.3 *Comparative statics for observed revenue, conditional on production: simulations*

As in Section C.2.1.3 we use simulations to characterize the posterior distribution of observed revenue, conditional on production. We choose the parameters of the Pareto distribution to roughly mimic the observed distribution of revenue in our sample. Therefore, in all simulations, we set $x_m = 20$ (roughly equal to the minimum observed revenue in our sample) and set the baseline value of a at 0.2.[3] In this case, $\frac{\phi(\log(x_m))}{2a(1-\Phi(\log(x_m)))} \approx 8.2$, so we choose $\sigma_y = 8$ as the baseline. The prior mean is $1/a + log(x_m) \approx 8$, so we choose $\pi_0 = 8$ as the baseline. The results of the simulations are presented in Table B.2.

---

[3] The maximum likelihood estimate of a in our full sample is 0.09; 0.18 if one excludes the bottom 10% of the distribution; and 0.22 if one excludes the bottom 25%. We chose a slightly higher value of a as the baseline in our simulations because lower values of a will result in an implausibly large fraction of movies with explosive revenues (the mean of a Pareto distribution with $a < 1$ is infinite).

Table B.2: Simulation results: Pareto distribution

A: Taste based discrimination: $\pi_0 \uparrow$ for Non-white movies

Fixed $a = 0.2$, $\sigma_y = 8$, $x_m = 20$

π_0	5	7	9	11	13	15	17	19	21	23	Trend
mean	8.091	8.164	8.252	8.374	8.548	8.767	9.041	9.381	9.800	10.309	$\uparrow$
std	5.041	5.081	5.108	5.173	5.270	5.362	5.521	5.686	5.888	6.137	$\uparrow$

Fixed $a = 0.5$, $\sigma_y = 8$, $x_m = 20$

π_0	5	7	9	11	13	15	17	19	21	23	Trend
mean	5.697	5.858	6.020	6.215	6.474	6.747	7.105	7.373	7.844	8.518	$\uparrow$
std	2.545	2.670	2.781	2.951	3.141	3.327	3.621	3.731	4.094	4.364	$\uparrow$

B: Customer discrimination: $a \uparrow$ for Non-white movies

Fixed $\pi_0 = 8$, $\sigma_y = 8$, $x_m = 20$

a	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	
mean	13.056	9.786	8.198	7.315	6.769	6.428	6.201	6.040	5.944	5.837	$\downarrow$
std	10.012	6.711	5.083	4.168	3.587	3.242	3.008	2.833	2.742	2.628	$\downarrow$

C: Statistical discrimination: $\sigma_y \uparrow$ for Non-white movies

Fixed $a = 0.2$, $\pi_0 = 8$, $x_m = 20$

σ_y	1	2	3	4	5	6	7	8	9	10	
mean	9.736	8.344	8.156	8.112	8.114	8.139	8.156	8.203	8.254	8.324	?
std	5.082	5.072	5.045	5.026	5.027	5.061	5.078	5.087	5.128	5.168	?

Fixed $a = 0.5$, $\pi_0 = 8$, $x_m = 20$

σ_y	1	2	3	4	5	6	7	8	9	10	
mean	6.749	5.469	5.327	5.361	5.447	5.585	5.750	5.930	6.151	6.360	?
std	2.216	2.167	2.152	2.203	2.278	2.412	2.565	2.725	2.933	3.118	?

Legend: simulated data with sample size $L = 10^6$, using R with seed 123. Mean, Std: sample average and standard deviation of the posterior distribution of $\pi|y, y > \bar{y}$ from the simulation.

In Panel A we examine the role of taste-based discrimination. We fix the values of a and σ_y and study what happens to the posterior mean and standard deviation of (log) revenue conditional on production as we increase π_0. The posterior mean increases (as in the normal-normal case), while the standard deviation also increases which is different from the normal-normal case.

In Panel B we look at the effect of increasing customer discrimination by letting a increase. Both the posterior mean and standard deviation decrease as the extent of customer discrimination increases, matching the predictions of the normal-normal model.

Finally, in panel C we vary the extent of statistical discrimination by letting σ_y increase, i.e., making the signal less precise. Here the results stand in contrast with those of the normal-normal model: as the signal becomes less precise, both the posterior mean of log revenue and the posterior standard deviation have a U-shaped pattern, first decreasing and then increasing in the extent of noise in the signal.

The comparative statics in the Pareto model are not identical to those in the normal-normal model presented in the main text. However, the simulations show that both the mean and the variance of log box-office revenue always move in the same direction as we change the discrimination parameter, under all three forms of discrimination. This is in contrast with the observed patterns in the data, where the mean of log revenue is higher for non-white movies, but the variance of log revenue is smaller (see Table 3.6 in the text).

C.3 MACHINE LEARNING ALGORITHM FOR FACIAL CLASSIFICATION

For performers that were not unambiguously classified by the human raters, we applied the facial classification algorithm proposed by Anwar and Islam (2017)[1]. The algorithm is based on a machine learning architecture that combines a convolutional neural network (CNN) and support vector machine (SVM), described below.

---

[1] Link: https://arxiv.org/ftp/arxiv/papers/1709/1709.07429.pdf.

Step 1. We started with a sample of more than 7000 motion pictures released in the United States between 1997 and 2017, taken from Opus Data,[2] a private company that collects data on the industry. For each movie, we took the names of the four top-billed performers. We then scraped and cropped the image appearing on each performer's page on the popular website IMDB.[3]

Step 2. We used the Visual Geometry Group[4] (V.G.G.) technique to locate the actor's face on each picture. The output of this step is a vector of information extracted from each image, or a "feature vector."

Step 3. We repeated step 2 on our training data set, the Chicago Face Database (CFD).[5] This database is intended for use in scientific research. It is useful as it contains images of 597 unique individuals (both male and female) who self-identify as White, Black, Asian, or Latino/a.

Step 4. We used CFD to train our algorithm using the Support Vector Machine (SVM) approach.[6] Intuitively, the purpose of SVM is to find the "best separation line," meaning the hyper-plane that correctly separates white from non-white performers when such performers are located in a multi-dimensional space through their feature vectors.

Step 5. We applied our trained algorithm to the pictures obtained from Steps 1 and 2. We validated our algorithm on a subsample of actors for which we manually coded the racial groups and obtained a success rate of 95%. A few examples of the outcomes of our classification algorithm are presented in Figure C.1.

---

[2] www.opusdata.com

[3] www.imdb.com

[4] See for reference https://www.robots.ox.ac.uk/~vgg/.

[5] The CFD is available at https://www.chicagofaces.org/.

[6] See for reference https://scikit-learn.org/stable/modules/svm.html.

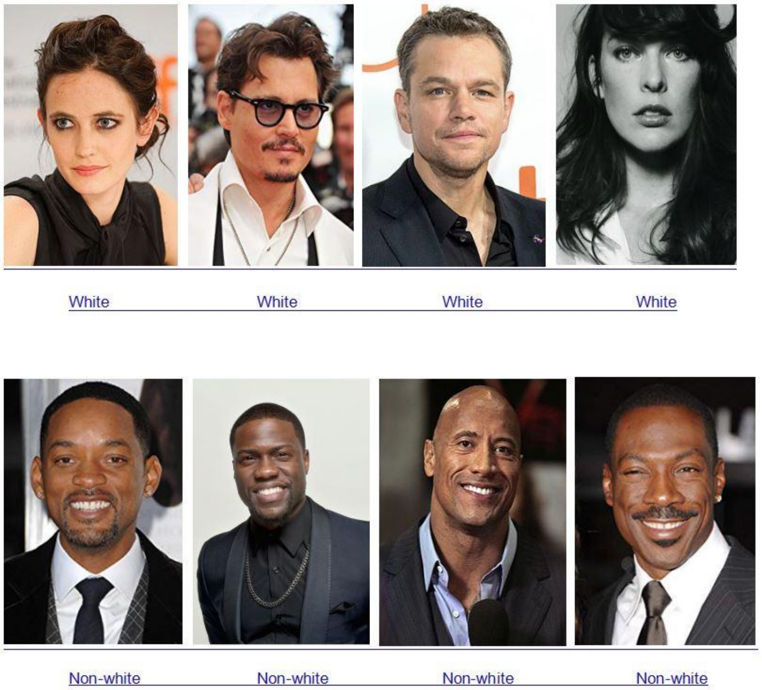


Figure C.1: Output of facial classification

C.4 OTHER FIGURES

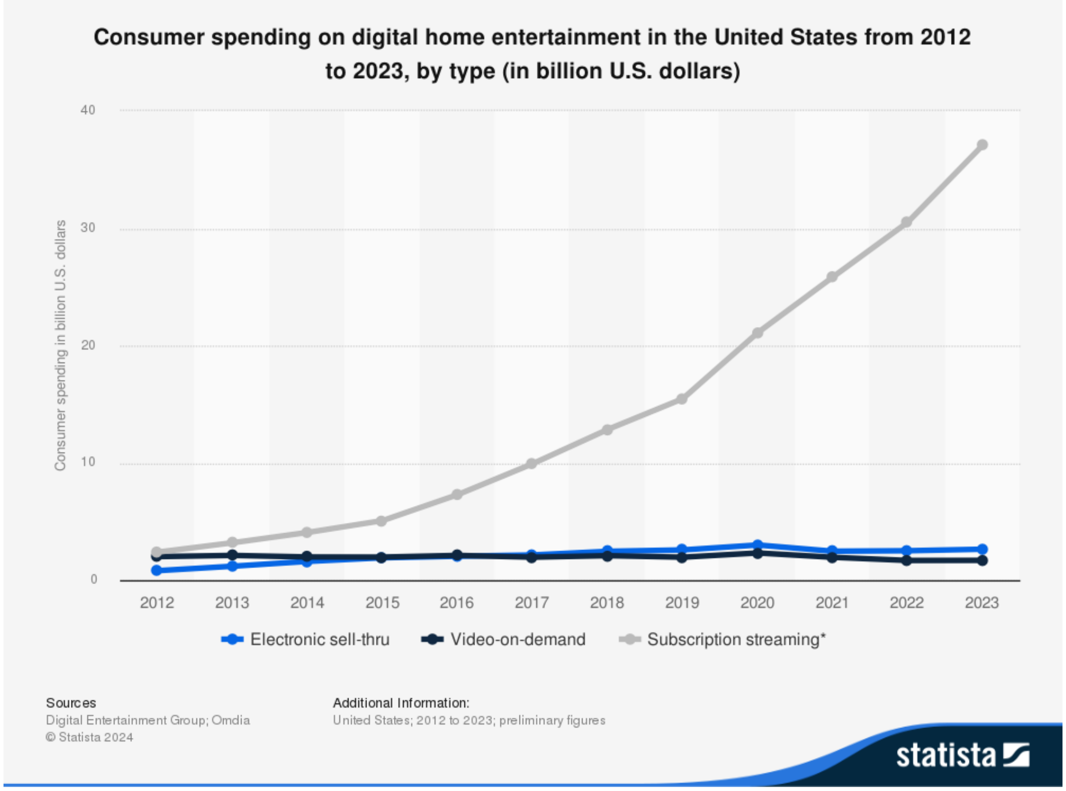


Figure D.1: Trend in consumer spending on digital home entertainment, by category

Source: Statista (link.)

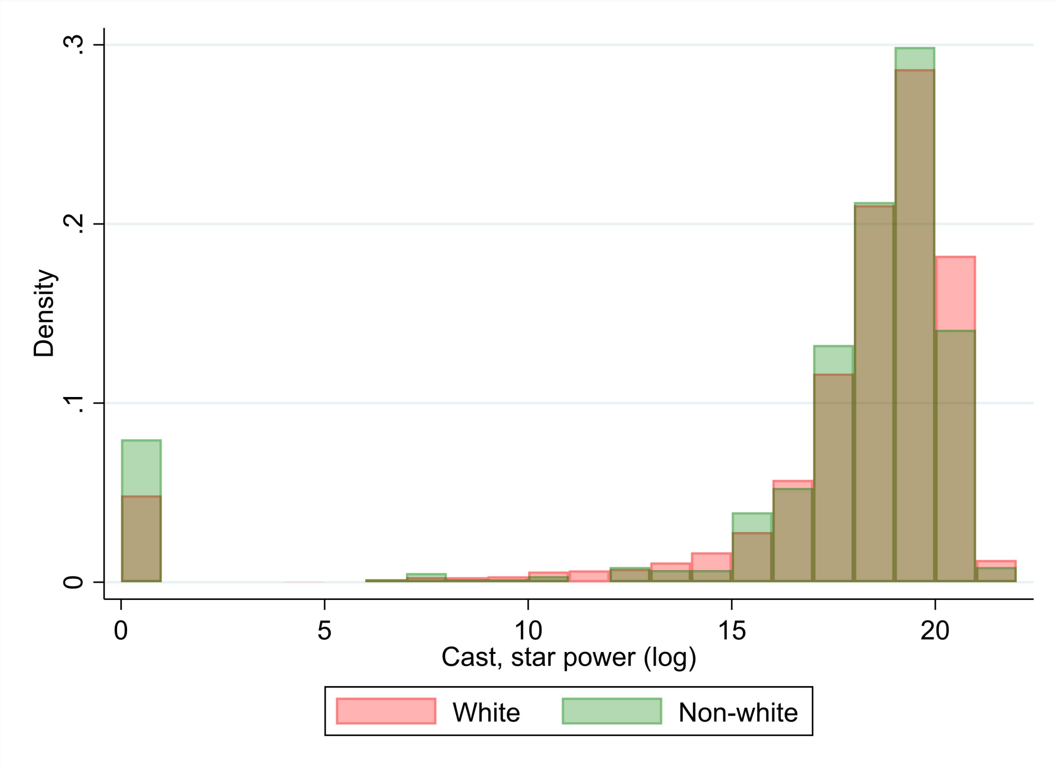


Figure D.2: Distribution of log cast star power across racial movie types

Note: A one-sided t-test on the means calculated off the full distributions fails to reject the null hypothesis that the white average is larger than the non-white average. Excluding the left tail of the distributions (i.e., truncating the distributions from below at 5) makes the means non significantly different.

BIBLIOGRAPHY

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2020). Sampling-based versus design-based uncertainty in regression analysis. *Econometrica*, *88*(1), pp. 265–296.

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2022). When Should You Adjust Standard Errors for Clustering?*. *The Quarterly Journal of Economics*, *138*(1), 1–35.

Adukia, A., Eble, A., Harrison, E., Runesha, H. B., & Szasz, T. (2023). What we teach about race and gender: Representation in images and text of childrens books. *Quarterly Journal of Economics*, *138*(4), 2225–2285.

Agan, A., & Starr, S. (2018). Ban the box, criminal records, and racial discrimination: A field experiment. *The Quarterly Journal of Economics*, *133*(1), 191–235.

Altonji, J. G., & Pierret, C. R. (2001). Employer learning and statistical discrimination. *The Quarterly Journal of Economics*, *116*(1), 313–350.

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects. *Journal of the American Statistical Association*, *103*(484), 1481–1495.

Andrews, D. W. (1997). A stopping rule for the computation of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, (pp. 913–931).

Angrist, J. D. (2014). The perils of peer effects. *Labour Economics*, *30*, 98–108.

Anwar, I., & Islam, N. U. (2017). Learned features are better for ethnicity classification. *Cybernetics and Information Technologies*, *17*(3), 152–164.

Anwar, S., & Fang, H. (2015). Testing for racial prejudice in the parole board release process: Theory and evidence. *The Journal of Legal Studies*, *44*(1), 1–37.

Arnold, D., Dobbie, W., & Hull, P. (2022). Measuring racial discrimination in bail decisions. *American Economic Review*, *112*(9), 2992–3038.

Arnoud, A., Guvenen, F., & Kleineberg, T. (2019). Benchmarking global optimizers. *NBER Working Paper*, (w26340).

Aronow, P. M. (2012). A general method for detecting interference between units in randomized experiments. *Sociological Methods & Research*, *41*(1), 3–16.

Aronow, P. M., Eckles, D., Samii, C., & Zonszein, S. (2020). Spillover effects in experimental data. https://arxiv.org/abs/2001.05444.

Aronow, P. M., & Samii, C. (2017). Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics*, *11*(4), 1912 – 1947.

Arrow, K. e. a. (1973). The theory of discrimination, discrimination in labor markets. In A. Achenfelter, & R. Ress (Eds.) *Discrimination in Labor Markets*. Princeton, New Jersey: Princeton University Press.

Åslund, O., Hensvik, L., & Skans, O. N. (2014). Seeking similarity: How immigrants and natives manage in the labor market. *Journal of Labor Economics*, *32*, 405441.

Athey, S., Eckles, D., & Imbens, G. W. (2018). Exact p-values for network interference. *Journal of the American Statistical Association*, *113*(521), 230–240.

Bar, R., & Zussman, A. (2017). Customer discrimination: Evidence from israel. *Journal of Labor Economics*, *35*(4), 1031–1059.

Barber, R. F., Candes, E. J., Ramdas, A., & Tibshirani, R. J. (2021). Predictive inference with the jackknife+. *The Annals of Statistics*, *49*(1), 486–507.

Basse, G., Ding, P., Feller, A., & Toulis, P. (2024). Randomization tests for peer effects in group formation experiments. *Econometrica*, *92*(2), 567–590.

Basse, G., & Feller, A. (2018). Analyzing two-stage experiments in the presence of interference. *Journal of the American Statistical Association*, *113*(521), 41–55.

Basse, G. W., & Airoldi, E. M. (2018). Limitations of design-based causal inference and a/b testing under arbitrary and network interference. *Sociological Methodology*, *48*(1), 136–151.

Basse, G. W., Feller, A., & Toulis, P. (2019). Randomization tests of causal effects under interference. *Biometrika*, *106*(2), 487–494.

Becker, G. S. (1957). *The economics of discrimination*. University of Chicago Press.

Bélisle, C. J. (1992). Convergence theorems for a class of simulated annealing algorithms on $\mathbb{R}^d$. *Journal of Applied Probability*, *29*(4), 885–895.

Benson, A., Board, S., & ter Vehn, M. M. (2024). Discrimination in hiring: Evidence from retail sales. *Review of Economic Studies*, *91*(4), 1956–1987.

Berry, S., Levinsohn, J., & Pakes, A. (1995). Automobile Prices in Market Equilibrium. *Econometrica*, *63*(4), 841.

Bertrand, M., & Duflo, E. (2017). Field experiments on discrimination. In *Handbook of economic field experiments*, vol. 1, (pp. 309–393).

Bharadwaj, P., Deb, R., & Renou, L. (2024). Statistical discrimination and the distribution of wages. Unpublished manuscript.

Bhatia, R. (2013). *Matrix Analysis*, vol. 169. Springer Science & Business Media.

Blattman, C., Green, D. P., Ortega, D., & Tobón, S. (2021). Place-Based Interventions at Scale: The Direct and Spillover Effects of Policing and City Services on Crime [Clustering as a Design Problem]. *Journal of the European Economic Association*, *19*(4), 2022–2051.

Bohren, J. A., Haggag, K., Imas, A., & Pope, D. G. (2023). Inaccurate statistical discrimination: An identification problem. *Review of Economics and Statistics*, (pp. 1–45).

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, *489*(7415), 295–298.

Bordalo, P., Coffman, K., Gennaioli, N., & Shleifer, A. (2016). Stereotypes. *The Quarterly Journal of Economics*, *131*(4), 1753–1794.

Borusyak, K., & Hull, P. (2023). Nonrandom exposure to exogenous shocks. *Econometrica : journal of the Econometric Society*, *91 6*, 2155–2185.

Bosch, A. (1986). The factorization of a square matrix into two symmetric matrices. *The American Mathematical Monthly*, *93*(6), 462–464.

Bowers, J., Fredrickson, M. M., & Panagopoulos, C. (2013). Reasoning about interference between units: A general framework. *Political Analysis*, *21*(1), 97124.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Breza, E., Chandrasekhar, A. G., McCormick, T. H., & Pan, M. (2020). Using aggregated relational data to feasibly identify network structure without network data. *American Economic Review*, *110*(8), 245484.

Brollo, F., Maria Kaufmann, K., & La Ferrara, E. (2020). Learning spillovers in conditional welfare programmes: Evidence from brazil. *The Economic Journal*, *130*(628), 853–879.

Brunner, D., Heiss, F., Romahn, A., & Weiser, C. (2017). *Reliable estimation of random coefficient logit demand models*. 267. DICE Discussion Paper.

Burdekin, R. C., & Idson, T. L. (1991). Customer preferences, attendance and the racial structure of professional basketball teams. *Applied Economics*, *23*(1), 179–186.

Cai, J., De Janvry, A., & Sadoulet, E. (2015). Social networks and the decision to insure. *American Economic Journal: Applied Economics*, *7*(2), 81108.

Canay, I. A., Mogstad, M., & Mountjoy, J. (2023). On the use of outcome tests for detecting bias in decision making. *The Review of Economic Studies*. Accepted for publication.

Caughey, D., Dafoe, A., Li, X., & Miratrix, L. (2023). Randomisation inference beyond the sharp null: bounded null hypotheses and quantiles of individual treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *85*(5), 1471–1491.

Chan, A. (2024). Customer discrimination and quality signals.

Charles, K. K., & Guryan, J. (2008). Prejudice and wages: an empirical assessment of becker's the economics of discrimination. *Journal of Political Economy*, *116*(5), 773–809.

Chernozhukov, V., & Hong, H. (2003). An mcmc approach to classical estimation. *Journal of Econometrics*, *115*(2), 293–346.

Cochrane, J. H. (2005). *Asset Pricing*. Princeton University Press.

Colacito, R., Croce, M., Ho, S., & Howard, P. (2018). Bkk the ez way: International long-run growth news and capital flows. *American Economic Review*, *108*(11), 3416–49.

Colella, F. (2021). Who benefits from support? the heterogeneous effects of supporters on athletes' performance by skin colour. Mimeo., Université de Lausanne.

Combes, P.-P., Decreuse, B., Laouenan, M., & Trannoy, A. (2016). Customer discrimination and employment outcomes: Theory and evidence from the french labor market. *Journal of Labor Economics*, *34*(1), 107–160.

Conlon, C., & Gortmaker, J. (2020). Best practices for differentiated products demand estimation with pyblp. *The RAND Journal of Economics*, *51*(4), 1108–1161.

Cox, D. R. (1958). *Planning of Experiments*. New York: Wiley.

Crimson Engine (2018). What does a producer actually do? https://www.youtube.com/watch?v=71Oh4gQ-1jM.

Cui, R., Li, J., & Zhang, D. J. (2020). Reducing discrimination with reviews in the sharing economy: Evidence from field experiments on airbnb. *Management Science*, *66*(3), 1071–1094.

de Paula, A., Richards-Shubik, S., & Tamer, E. (2018). Identifying preferences in networks with bounded degree. *Econometrica*, *86*(1), 263–288.

Dennis, J. E., & Schnabel, R. B. (1996). *Numerical methods for unconstrained optimization and nonlinear equations*. SIAM.

Deuflhard, P. (2005). *Newton methods for nonlinear problems: affine invariance and adaptive algorithms*, vol. 35. Springer Science & Business Media.

Ding, P., Feller, A., & Miratrix, L. (2016). Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *78*(3), 655–671.

Doleac, J. L., & Stein, L. C. (2013). The visible hand: Race and online market outcomes. *The Economic Journal*, *123*(572), F469–F492.

Donaldson, D. (2018). Railroads of the raj: Estimating the impact of transportation infrastructure. *American Economic Review*, *108*(4-5), 899–934.

Dufour, J.-M., & Khalaf, L. (2003). *Monte Carlo Test Methods in Econometrics*, chap. 23, (pp. 494–519). John Wiley & Sons, Ltd.

Dustmann, C., Glitz, A., & Schönberg, U. (2016). Referral-based job search networks. *Review of Economic Studies*, *83*, 514546.

Esponda, I., Oprea, R., & Yuksel, S. (2022). Discrimination without reason: Biases in statistical discrimination.

Fang, H., & Moro, A. (2011). Theories of statistical discrimination and affirmative action: A survey. In *Handbook of social economics*, vol. 1, (pp. 133–200).

Fang, K.-T., & Wang, Y. (1993). *Number-theoretic methods in statistics*, vol. 51. CRC Press.

Fisher, F. (1966). *The Identification Problem in Econometrics*. Economics handbook series. McGraw-Hill.

Fong, C. M., & Luttmer, E. F. (2011). Do fairness and race matter in generosity? evidence from a nationally representative charity experiment. *Journal of Public Economics*, *95*(5-6), 372–394.

Forneron, J.-J. (2023). Noisy, non-smooth, non-convex estimation of moment condition models. *arXiv preprint arXiv:2301.07196*.

Fowdur, L., Kadiyali, V., & Prince, J. (2012). Racial bias in expert quality assessment: A study of newspaper movie reviews. *Journal of Economic Behavior & Organization*, *84*(1), 292–307.

Frobenius, G. (1910). Über die mit einer matrix vertauschbaren matrizen. In *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften: Jahrgang 1910; Erster Halbband Januar bis Juni*, (pp. 3–15). Verlag der Königlichen Akademie der Wissenschaften.

Gallen, Y., & Wasserman, M. (2023). Does information affect homophily? *Journal of Public Economics*, *222*, 104876.

Gill, D., & Prowse, V. (2012). A structural analysis of disappointment aversion in a real effort competition. *American Economic Review*, *102*(1), 469–503.

Gladwell, M. (2006). The formula. Accessed: 2024-11-10, https://www.newyorker.com/magazine/2006/10/16/the-formula.

Goldenberg, A., Zheng, A. X., Fienberg, S. E., & Airoldi, E. M. (2009). A survey of statistical network models. *ArXiv*, *abs/0912.5410*.

Gourieroux, C., & Monfort, A. (1996). *Simulation-based econometric methods*. Oxford university press.

Gourieroux, C., Monfort, A., & Renault, E. (1993). Indirect inference. *Journal of applied econometrics*, *8*(S1), S85–S118.

Graham, B. S. (2017). An econometric model of network formation with degree heterogeneity. *Econometrica*, *85*(4), 1033–1063.

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, *78*(6), 1360–1380.

Guan, L. (2023). A conformal test of linear models via permutation-augmented regressions.

Guminov, S., Gasnikov, A., & Kuruzov, I. (2017). Accelerated methods for α-weakly-quasi-convex problems. *arXiv preprint arXiv:1710.00797*.

Guryan, J., & Charles, K. K. (2013). Tastebased or statistical discrimination: the economics of discrimination returns to its roots. *The Economic Journal*, *123*(572), F417–F432.

Hall, A. R., & Inoue, A. (2003). The large sample behaviour of the generalized method of moments estimator in misspecified models. *Journal of Econometrics*, *114*(2), 361–394.

Hansen, B. E., & Lee, S. (2021). Inference for iterated gmm under misspecification. *Econometrica*, *89*(3), 1419–1447.

Hedegaard, M. S., & Tyran, J.-R. (2018). The price of prejudice. *American Economic Journal: Applied Economics*, *10*, 4063.

Heid, P. (2023). A short note on an adaptive damped newton method for strongly monotone and lipschitz continuous operator equations. *Archiv der Mathematik*, *121*(1), 55–65.

Hennessy, J. P., Dasgupta, T., Miratrix, L. W., Pattanayak, C. W., & Sarkar, P. (2015). A conditional randomization test to account for covariate imbalance in randomized experiments. *Journal of Causal Inference*, *4*, 61 – 80.

Hinder, O., Sidford, A., & Sohoni, N. (2020). Near-optimal methods for minimizing star-convex functions and beyond. In *Conference on learning theory*, (pp. 1894–1938). PMLR.

Hoshino, T., & Yanagi, T. (2023). Randomization test for the specification of interference structure. https://arxiv.org/abs/2301.05580.

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

Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Jayachandran, S., de Laat, J., Lambin, E. F., Stanton, C. Y., Audy, R., & Thomas, N. E. (2017). Cash for carbon: A randomized trial of payments for ecosystem services to reduce deforestation. *Science*, *357*(6348), 267–273.

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics*, *40*(2), 633–643.

Kahn, L. M., & Sherer, P. D. (1988). Racial differences in professional basketball players' compensation. *Journal of Labor Economics*, *6*(1), 40–61.

Karimi, H., Nutini, J., & Schmidt, M. (2016). Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, (pp. 795–811). Springer.

Kelly, B., Lustig, H., & Van Nieuwerburgh, S. (2016). Too-systemic-to-fail: What option markets imply about sector-wide government guarantees. *American Economic Review*, *106*(6), 1278–1319.

Kelly, M. (2021). Persistence, Randomization, and Spatial Noise. Working Papers 202124, School of Economics, University College Dublin. https://ideas.repec.org/p/ucn/wpaper/202124.html.

Kline, P., Rose, E. K., & Walters, C. R. (2022). Systemic discrimination among large us employers. *The Quarterly Journal of Economics*, *137*(4), 19632036.

Knittel, C. R., & Metaxoglou, K. (2014). Estimation of random-coefficient demand models: two empiricists' perspective. *Review of Economics and Statistics*, *96*(1), 34–59.

Knowles, J., Persico, N., & Todd, P. (2001). Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy*, *109*(1), 203–229.

Komunjer, I. (2012). Global identification in nonlinear models with moment restrictions. *Econometric Theory*, *28*(4), 719–729.

Kuehn, J., & Lampe, R. (2023). Competition and product composition: Evidence from hollywood. *International Journal of Industrial Organization*, *91*, 102981.

Kuppuswamy, V., & Younkin, P. (2020). Testing the theory of consumer discrimination as an explanation for the lack of minority hiring in hollywood films. *Management Science*, *66*(3), 1227–1247.

Lagarias, J. C., Reeds, J. A., Wright, M. H., & Wright, P. E. (1998). Convergence properties of the nelder–mead simplex method in low dimensions. *SIAM Journal on optimization*, *9*(1), 112–147.

Lang, K., & Lehmann, J.-Y. K. (2012). Racial discrimination in the labor market: Theory and empirics. *Journal of Economic Literature*, *50*(4), 959–1006.

Lang, K., & Spitzer, A. K. (2020). Race discrimination: An economic perspective. *Journal of Economic Perspectives*, *34*(2), 68–89.

Lash, M. T., & Zhao, K. (2016). Early predictions of movie success. *Journal of Management Information Systems*, *33*(3), 874–903.

Lehmann, E. L. E. L., & Romano, J. P. (2005). *Testing statistical hypotheses*. Springer texts in statistics. New York: Springer, 3rd ed. ed.

Lemieux, C. (2009). *Monte Carlo and Quasi-Monte Carlo Sampling*. Springer Series in Statistics. Springer New York.

Leonard, J. S., Levine, D. I., & Giuliano, L. (2010). Customer discrimination. *The Review of Economics and Statistics*, *92*(3), 670–678.

Leung, M. P. (2020). Treatment and Spillover Effects Under Network Interference. *The Review of Economics and Statistics*, *102*(2), 368–380.

Leung, M. P. (2022). Causal inference under approximate neighborhood interference. *Econometrica*, *90*(1), 267–293.

Li, X., Ding, P., Lin, Q., Yang, D., & Liu, J. S. (2018). Randomization inference for peer effects. *Journal of the American Statistical Association*, *114*, 1651 – 1664.

Li, X., Ding, P., & Rubin, D. B. (2016). Asymptotic theory of rerandomization in treatmentcontrol experiments. *Proceedings of the National Academy of Sciences*, *115*, 9157 – 9162.

Lippens, L., Baert, S., Ghekiere, A., Verhaeghe, P.-P., & Derous, E. (2020). Is labour market discrimination against ethnic minorities better explained by taste or statistics? a systematic review of the empirical evidence. Tech. rep., IZA Discussion Paper No. 13523.

Lise, J., & Robin, J.-M. (2017). The macrodynamics of sorting between workers and firms. *American Economic Review*, *107*(4), 1104–35.

List, J. A. (2004). The nature and extent of discrimination in the marketplace: Evidence from the field. *The Quarterly Journal of Economics*, *119*(1), 49–89.

Lojasiewicz, S. (1963). A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, *117*(87-89), 2.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, *60*, 531–542.

McKinnon, K. I. (1998). Convergence of the nelder–mead simplex method to a nonstationary point. *SIAM Journal on optimization*, *9*(1), 148–158.

Meinshausen, N. (2008). Hierarchical testing of variable importance. *Biometrika*, *95*(2), 265–278.

Miguel, E., & Kremer, M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, *72*(1), 159–217.

Moran, P., & Queralto, A. (2018). Innovation, productivity, and monetary policy. *Journal of Monetary Economics*, *93*, 24–41.

Moretti, E. (2011). Social learning and peer effects in consumption: Evidence from movie sales. *The Review of Economic Studies*, *78*(1), 356–393.

Motion Picture Association (2022). 2021 theme report. Accessed: 2024-11-10, https://www.motionpictures.org/wp-content/uploads/2022/03/MPA-2021-THEME-Report-FINAL.pdf.

Nardinelli, C., & Simon, C. (1990). Customer racial discrimination in the market for memorabilia: The case of baseball. *The Quarterly Journal of Economics*, *105*(3), 575–595.

Nash, J. C. (1990). *Compact numerical methods for computers: linear algebra and function minimisation*. Routledge.

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The computer journal*, *7*(4), 308–313.

Nesterov, Y. (2018). *Lectures on convex optimization*. Springer optimization and its applications. Cham, Switzerland: Springer International Publishing, 2 ed.

Nesterov, Y., & Polyak, B. T. (2006). Cubic regularization of newton method and its global performance. *Mathematical programming*, *108*(1), 177–205.

Neumark, D., Bank, R. J., & Van Nort, K. D. (1996). Sex discrimination in restaurant hiring: An audit study. *The Quarterly Journal of Economics*, *111*(3), 915–941.

Nevo, A. (2001). Measuring market power in the ready-to-eat cereal industry. *Econometrica*, *69*(2), 307–342.

Newey, W., & McFadden, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, vol. 36:4, (pp. 2111–2234). North Holland.

Neyman, J., Iwaszkiewicz, K., & Koodziejczyk, S. (2018). Statistical Problems in Agricultural Experimentation. *Supplement to the Journal of the Royal Statistical Society*, *2*(2), 107–154.

Niederreiter, H. (1983). A quasi-monte carlo method for the approximate computation of the extreme values of a function. In *Studies in pure mathematics*, (pp. 523–529). Springer.

Nocedal, J., & Wright, S. (2006). *Numerical Optimzation*. Springer, second ed.

Onuchic, P. (2022). Recent contributions to theories of discrimination. *arXiv preprint*.

Owusu, J. (2023). Randomization inference of heterogeneous treatment effects under network interference.

Paluck, E., Shepherd, H., & Aronow, P. (2016). Changing climates of conflict: A social network experiment in 56 schools. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(3), 566–571.

Phelps, E. S. (1972). The statistical theory of racism and sexism. *The American Economic Review*, *62*(4), 659–661.

Pierson, E. (2020). Assessing racial inequality in covid-19 testing with bayesian threshold tests. *arXiv preprint*.

Pierson, E., Corbett-Davies, S., & Goel, S. (2018). Fast threshold tests for detecting discrimination. In *International conference on artificial intelligence and statistics*, (pp. 96–105). PMLR.

Pollmann, M. (2023). Causal inference for spatial treatments.

Polyak, B., & Tremba, A. (2020). New versions of newton method: step-size choice, convergence domain and under-determined equations. *Optimization Methods and Software*, *35*(6), 1272–1303.

Polyak, B. T. (1963). Gradient methods for minimizing functionals. *Zhurnal vychislitel'noi matematiki i matematicheskoi fiziki*, *3*(4), 643–653.

Pouliot, G. (2024). An exact t-test.

Powell, M. J. (1973). On search directions for minimization algorithms. *Mathematical programming*, *4*(1), 193–201.

Puelz, D., Basse, G., Feller, A., & Toulis, P. (2021). A Graph-Theoretic Approach to Randomization Tests of Causal Effects under General Interference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *84*(1), 174–204.

Rajkumar, K., Saint-Jacques, G., Bojinov, I., Brynjolfsson, E., & Aral, S. (2022). A causal test of the strength of weak ties. *Science*, *377*, 1304 – 1310.

Riley, E. (2024). Role models in movies: The impact of queen of katwe on students' educational attainment. *The Review of Economics and Statistics*, *106*(2), 334–351.

Ritzwoller, D. M., Romano, J. P., & Shaikh, A. M. (2025). Randomization inference: Theory and applications. https://arxiv.org/abs/2406.09521.

Rockafellar, R. T. (2015). *Convex Analysis*. Princeton Landmarks in Mathematics and Physics. Princeton, NJ: Princeton University Press,.

Rosenbaum, P. (2007). Interference between units in randomized experiments. *Journal of the American Statistical Association*, *102*(477), 191–200.

Rosenbaum, P. (2020). *Design of Observational Studies*. Springer Series in Statistics. Springer International Publishing.

Rosenbaum, S. (1961). Moments of a truncated bivariate normal distribution. *Journal of the Royal Statistical Society. Series B (Methodological)*, *23*(2), 405–408.

Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica: Journal of the Econometric Society*, (pp. 577–591).

Sacerdote, B. (2001). Peer Effects with Random Assignment: Results for Dartmouth Roommates*. *The Quarterly Journal of Economics*, *116*(2), 681–704.

Salanié, B., & Wolak, F. A. (2022). Fast, detail-free, and approximately correct: Estimating mixed demand systems.

Shirani, S., & Bayati, M. (2024). Causal message-passing for experiments with unknown and general network interference. *Proceedings of the National Academy of Sciences*, *121*(40).

Sieg, H., & Yoon, C. (2017). Estimating dynamic games of electoral competition to evaluate term limits in us gubernatorial elections. *American Economic Review*, *107*(7), 1824–1857.

Simoiu, C., Corbett-Davies, S., & Goel, S. (2017). The problem of infra-marginality in outcome tests for discrimination. *arXiv preprint*.

Snee, T. (2016). Predicting box office: Boffo or bomb? Accessed: 2024-11-10, https://now.uiowa.edu/news/2016/02/predicting-box-office-boffo-or-bomb.

Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate? *Journal of the American Statistical Association*, *101*(476), 1398–1407.

Solodov, M. V., & Svaiter, B. F. (2000). A truly globally convergent newton-type method for the monotone nonlinear complementarity problem. *SIAM Journal on Optimization*, *10*(2), 605–625.

Spall, J. C. (2005). *Introduction to stochastic search and optimization: estimation, simulation, and control*. John Wiley & Sons.

Sperling, N. (2020a). Academy sets new diversity requirements for oscars best picture eligibility. *The New York Times*. Accessed: 2024-11-10.

Sperling, N. (2020b). The oscars will add a diversity requirement for eligibility. *The New York Times*. Accessed: 2024-11-10.

Stone, E. W., & Warren, R. S. (1999). Customer discrimination in professional basketball: Evidence from the trading-card market. *Applied Economics*, *31*(6), 679–685.

Taylor, S. J., & Eckles, D. (2018). *Randomized Experiments to Detect and Estimate Social Influence in Networks*, (pp. 289–322). Cham: Springer International Publishing.

Thompson, B. (2013). Solving equation of a hit film script, with data. Accessed: 2024-11-10, https://www.nytimes.com/2013/05/06/business/media/solving-equation-of-a-hit-film-script-with-data.html.

Toulis, P., & Kao, E. (2013). Estimation of causal peer influence effects. In S. Dasgupta, & D. McAllester (Eds.) *Proceedings of the 30th International Conference on Machine Learning*, vol. 28 of *Proceedings of Machine Learning Research*, (pp. 1489–1497). Atlanta, Georgia, USA: PMLR. https://proceedings.mlr.press/v28/toulis13.html.

Vazquez-Bare, G. (2023). Identification and estimation of spillover effects in randomized experiments. *Journal of Econometrics*, *237*(1), 105237.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, vol. 47. Cambridge university press.

Viviano, D. (2022). Experimental design under network interference. https://arxiv.org/abs/2003.08421.

Vovk, V., Nouretdinov, I., Manokhin, V., & Gammerman, A. (2018). Cross-conformal predictive distributions. In A. Gammerman, V. Vovk, Z. Luo, E. Smirnov, & R. Peeters (Eds.) *Proceedings of the Seventh Workshop on Conformal and Probabilistic Prediction and Applications*, vol. 91 of *Proceedings of Machine Learning Research*, (pp. 37–51). PMLR. https://proceedings.mlr.press/v91/vovk18a.html.

Wang, L. B., Bedant, O. P., Jiao, Z., & Wang, H. (2024). From friendship networks to classroom dynamics: Leveraging neural networks, instrumental variable and genetic algorithms for optimal educational outcomes. https://arxiv.org/abs/2404.02497.

Wang, Y., Samii, C., Chang, H., & Aronow, P. M. (2023). Design-based inference for spatial experiments under unknown interference.

Weaver, A. J. (2011). The role of actors' race in white audiences' selective exposure to movies. *Journal of Communication*, *61*(2), 369–385.

Weinstein, M. (1998). Profit-sharing contracts in hollywood: Evolution and analysis. *The Journal of Legal Studies*, *27*(1), 67–112.

Wen, K., Wang, T., & Wang, Y. (2023). Residual permutation test for high-dimensional regression coefficient testing.

Wu, J., & Ding, P. (2021). Randomization tests for weak null hypotheses in randomized experiments. *Journal of the American Statistical Association*, *116*(536), 1898–1913.

Yahr, E. (2016). It's hard to predict a movie's profitability, but you learn some lessons along the way. Accessed: 2024-11-10, https://www.washingtonpost.com/news/arts-and-entertainment/wp/2016/05/16/.

Zhang, Y., & Zhao, Q. (2021). Multiple conditional randomization tests. *arXiv: Statistics Theory*.

Zhang, Y., & Zhao, Q. (2023). What is a randomization test? *Journal of the American Statistical Association*, *118*(544), 2928–2942.

Zhao, A., & Ding, P. (2020). Covariate-adjusted fisher randomization tests for the average treatment effect. *Journal of Econometrics*.

Zussman, A. (2013). Ethnic discrimination: Lessons from the israeli online market for used cars. *The Economic Journal*, *123*(572), F433–F468.

Zwick, E. (2024). *Hits, flops, and other illusions: My Fortysomething Years in Hollywood*. Gallery Books.

CURRICULUM VITAE

Liang Zhong, MA

Friday 18th April, 2025

85 Brainerd Rd
Allston MA 02134 USA
(617) 369-2010
610466047@qq.com

264 Bay state Rd
Boston University
Boston, MA 02215
samzl@bu.edu

Academic Training:

05/2025(expected)	PhD Boston University, Boston, MA; Economics
05/2019	MA Boston University, Boston, MA; Econometrics and Quantitative Economics
05/2017	BS Zhejiang University, Hang Zhou, China; Mathematics and Applied Mathematics

Doctoral Research:

Title:	Essays on Causal Inference, Structural Estimation, and their Applications
Thesis advisor:	Hiroaki Kaido, PhD
Defense date:	April 2, 2025
Summary:	This dissertation comprises three chapters that explore two interconnected areas: the development of innovative econometric tools to reduce computational complexities and the analysis of strategic behaviors for actionable policy insights. The first two chapters introduce new statistical approaches that link advanced econometric methods with empirical research, while the third chapter connects economic theory to practical applications by leveraging big data techniques.