The Instrumental Variable Model with Categorical Instrument, Exposure, and Outcome: Characterization, Partial Identification, and Statistical Inference

Yilin Song

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

Doctor of Philosophy

University of Washington

2025

Reading Committee:
Thomas Richardson, Chair
K.C. Gary Chan, Chair
Ting Ye

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

The Instrumental Variable Model
with Categorical Instrument, Exposure, and Outcome:
Characterization, Partial Identification, and Statistical Inference

Yilin Song

Co-Chairs of the Supervisory Committee:
Thomas Richardson
Department of Statistics

K.C. Gary Chan
Department of Biotatistics

Instrumental variable (IV) analysis is a crucial tool in estimating causal relationships that addresses the issue of confounding variables that may lead to bias. Under certain IV assumptions, the causal effect may be partially identified. The binary IV model has been well studied in economics, statistics, and epidemiology, while IV models for general categorical exposure and outcome are less explored. This dissertation studies several aspects of the instrumental variable model with categorical instrument, exposure, and outcome including giving a characterization of the model (Chapters 2, 3 and 5), methods for statistical inference (Chapter 4), and a study of the variation independence properties of the marginal counterfactual distributions (Chapter 6).

In Chapter 2, we first give a simple closed-form characterization of the set of joint distributions of the potential outcomes compatible with a given observed probability distribution via a set of inequalities. In Chapter 3, we further derive conditions for the inequalities in Chapter 2 to be non-redundant and construct the minimal set. To handle sampling variability, we provide an algorithm in Chapter 4 to construct confidence regions for any convex functional of the joint counterfactual distribution, such as the average causal effect (ATE),

using a finite-sample tail bound for the KL-divergence due to Guo and Richardson [2021]. We also illustrate our methods in Chapters 2 and 4 using data from the Minneapolis Domestic Violence Experiment. In Chapter 5, we study falsification tests for the categorical IV model through simulations. We explore the variation dependence property of the marginal counterfactual distributions and discuss its practical implications in Chapter 6. We conclude with a discussion and directions for future work in Chapter 7.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

ATE: Average treatment effect

RCT: Randomized controlled trials

ITT: Intention to treat

PP: Per protocol

LRT: Likelihood ratio test

FWER: Family-wise error rate

ACKNOWLEDGMENTS

Over the past five years, there were so many moments when I thought, "This is going in the acknowledgments for sure." I only wish I had written them all down. It still feels unreal that I'm done with my PhD – and that I'm finally writing this section, which, true to form, I saved for last (procrastination at its most sentimental). This is where I get to thank the people and moments that got me through it all – with caffeine, stress, and a lot of love.

I would like to thank my dissertation advisors, Dr. Thomas Richardson and Dr. Gary Chan, for their patience, insight, and unwavering support. I've always felt incredibly lucky to work with both of them – and to soak in their wisdom, whether it's about research, career, or life, during our weekly meetings. I'm still amazed by how they seem to know (and remember!) everything. Thomas is brilliant, rigorous, optimistic, and has a fantastic sense of humor. A couple of my all-time favorite Thomas quotes: "Epidemiologists always say biking is good for your cardiovascular health – but first, you have to survive biking;" and "I'm not one of those people who grade papers by sliding them down the stairs and seeing which one goes the furthest." Gary, on the other hand, has a superpower for intuition – no matter the topic, he always seems to know the answer. He's calm, thoughtful, and constantly radiates positivity. I'm deeply grateful for the times he listened patiently to my panicked thoughts and offered steady, reassuring guidance. I always left our conversations feeling more grounded and hopeful. This past year hasn't been easy in many ways, and I feel incredibly blessed to have had Thomas and Gary by my side – understanding, kind, and truly the best mentors I could have asked for.

I would also like to thank Dr. Ting Ye, who has been my RA advisor for the past four years and generously supported me – both intellectually and financially. She's not

only my advisor and academic role model, but also a friend, a sister, and someone I deeply admire. Her passion for research and meticulous attention to detail never fail to inspire me to aim higher (and double-check my work). Ting has always been open in sharing her own journey and offering thoughtful, honest guidance. Whenever I'm at a crossroads, I feel genuinely unsettled if I haven't talked to her. Her presence in my PhD life has been steady, empowering, and irreplaceable.

I'm grateful to Dr. Richard Guo for reading and collaborating on my paper, and to Dr. Yanqin Fan for serving on my committee and offering many valuable insights from an econometrician's perspective. Beyond my advisors and committee members, I also want to thank the wonderful scholars I had the privilege of interacting with at UW: Dr. Andrea Rotnitzky, Dr. Lyn Brumback, Dr. Guanghao Qi, and Dr. Lurdes Inoue – for many insightful and inspiring conversations. My graduate school journey wouldn't have gone nearly as smoothly without the incredible support of the department's staff. A heartfelt thank you to Gitana Garofalo, Minh Vo, Deb Nelson, Maggie Tarnawa, and Thomas Nelson for keeping everything running seamlessly – and for always being kind, patient, and generous with their help.

I want to thank my long-time collaborator, Dr. Stacey Winham at the Mayo Clinic, who first introduced me to the field of Biostatistics. I started working with her as an intern during my junior year of college, and we've been collaborating ever since. Back then, I was torn between going to graduate school right after college or taking a couple of years to work. We had countless conversations, and I shadowed her through most of her meetings until I realized her job was my dream job (and it still is!). Stacey has always gone out of her way to support me, often anticipating what I need before I even say a word. I truly wouldn't be where I am today without her guidance, encouragement, and belief in me. I'm also deeply grateful to my other wonderful collaborators at Mayo – Dr. Joanna Biernacka, Dr. Ada Ho, and Dr. Victor Karpyak – whom I've had the pleasure and privilege to work with over the

years.

I arrived in Seattle in June 2020 with the sadness of not being able to return home to enjoy my free summer after graduating from college. Jiaqi Yin (Shuashua), who graduated from our department, was looking for a roommate so I moved in with her, one of the best decisions I ever made. She became a beloved friend and sister ever since, sharing all my ups and downs, traveling together, and pushing me down the road of academia for the sake of her kid's summer research opportunity. She gave me so many suggestions, some I listened to and some I didn't, including talking to Thomas and being his student, getting a cat, joining the stock market, buying a house, and signing up for dating apps. She then introduced me to other friends Xiaoxiao Wu, whom I consult for fashion tips, and Ningxin Ma, who suffered Linear Models with me and is my eating, hiking, and badminton mate. I also feel grateful to get to know Shuxian Chen and Zhongdi Chu through Shuashua and enjoyed many of our gossiping and venting sessions.

I couldn't have done this without my friends in the Biostatistics department, who kept me sane and made coming into the office enjoyable. This includes my "Song Family" members Abby He, Zhaoheng Li, Rui Wang, Yimin Zhao, and Yunji Zhou. We shared endless laughter and created memories I'll always treasure (some of which still make me laugh out loud at random times). In addition, I would like to thank Erli, namely Mali (Xingyu Wang) and Haili (Congyu Hang) and Turbo Du for many late (pretended to be productive) nights at HRC. I thank Zichun Xu for sharing the desk space across from mine so that a random conversation begins with just a glance. To my cohort mates—Angela Daul, Anand Hemmady, Victoria Knutson, Nobu Masaki, and Gabby Vasconcelos—thank you for starting this wild ride of grad school with me during the pandemic, and for all the Zoom coffee chats that helped us feel a little less isolated. I also want to thank Sijia Li and Fan Xia for generously sharing career advice and helping me navigate the uncertainty of what comes next. And to my friends outside of school – Waterlily Huang, Stella Song, Zhiying Xie, and Siyu Zhang –

thank you for exploring Seattle with me and filling my weekends with fun and friendship. I've also been lucky to live in the same apartment complex as Wenbo Fei and Zhen Miao, which made it all too easy to organize spontaneous potlucks and board game nights (and maybe a few too many hot pot sessions). Lastly, I'm grateful to be part of such a warm and welcoming department and university community, especially fellow students like Nina Galanter, Leah Andrews, Ellen Graham, and Kenny Zhang.

I interned at Genentech in the summer of 2023, during which I learned so much from my mentors Ali Valcarcel and Michel Friesenhahn and also made a group of great friends. I met two of my best friends Wei Qiu and Cady Fu during the internship. They are caring, brilliant, passionate, and sweet—and most importantly, they embrace my goofiness as much as I embrace theirs. Though we only met two years ago, it feels like we've known each other forever. They've become the beloved aunties to my cats, surprise-givers (and surprise-lovers), and just all-around wonderful women who, frankly, any man would be lucky to have in their life. I'm still amazed by how many close friendships I formed during just a short internship. I feel so lucky to have bonded with Ziyi Song, Xinming Tu, Jiahui Peng, Wancen Mu, Yiwei Gong, Xin Huang, Vincent Guo, and Yingying Wei. Every trip with them was filled with joy, laughter, and the kind of moments that remind you how fun life can be when shared with the right people.

I would also like to thank my college friends – Xiaotian Fan, Rui Li, and Jiashan Song – for their companionship and for sharing those unforgettable quarantine days in Northfield, MN. I'm grateful to my childhood friends – Haoyu Tian, Zitong Shen, Xiangyu Zhang, Tianyu Gao, Ziyi Yang, Han Diao, Ziyue Wu, and Xinnan Li – for always welcoming me with warmth whenever I return home. Though we've each taken different paths, we continue to hold each other close in our hearts. When I started graduate school, I set a quiet goal for myself: to finish in five years – because this year, the last day of school happens to fall on my birthday. I thought it would be the perfect way to end my student life, and I'm

proud (and slightly stunned) that I actually pulled it off. One of the sweetest traditions that kept me grounded along the way was sharing a birthday with my best friend, Xueying Kong. Technically, her birthday is one day after mine, but because of the time difference, we've celebrated on the same day for the past nine years. She's the person I turn to when I'm overwhelmed or down, and I'm constantly inspired by her compassion and her uncanny ability to find silver linings in everything. Thankfully, we've both long forgotten whatever caused our dramatic teenage fights – and I feel incredibly lucky to still have her just a phone call away.

Special thanks to all the kittens and cats I handled while volunteering at Paws Cat City and to everyone I met on the badminton court for the games and the shared love of smashing stress away. I would especially like to thank Chen Yufei, the Olympic badminton champion, for inspiring fans like me with her resilience—the courage to pause, reflect, and begin again. Thanks Jia Yifan, another badminton Olympic champion, for her long vlogs which cheered me up and served as my melatonin in restless nights and my companion at countless meals. Huge thanks to BWF for live-streaming tournaments that kept me company through countless late nights of wrestling with theorems, simulations, and writing. Your presence, even through a screen, made the academic grind a little lighter.

Last but certainly not least, I want to thank my family – my strongest and most unwavering support system. My parents, Xiaoxia Zou and Qingxin Song, and my grandparents, Yingfen Lin and Shuli Zou, raised me with unconditional love and have always stood firmly behind me. I'm forever grateful to have them cheering me on through every step of this journey. My aunt, Jiao Lin (who is only six years older than me!) holds a special place in my heart. She's both family and my best friend, which means I get to share everything with her – from everyday joys to life's biggest questions. Her humor and optimism never fail to lift my spirits. During the pandemic, when I couldn't go home, staying with her in London (the one in Canada!) brought me so much comfort and joy. I'm also thankful to my uncle,

Xiaotao Zou, for his constant support throughout my time in the U.S. And to my boyfriend, Zhaoqi Li – thank you for your kindness, patience, encouragement, and love. You're the calm in my chaos and the voice that always tells me "You can do it," especially when imposter syndrome tries to say otherwise. Raising our two cats, Dobby and Lottie, together has been one of the brightest parts of graduate school. With you (and the cats), this journey has felt a little less stressful, and a lot more like home.

DEDICATION

to my grandparents, 林应芬 and 邹树礼,

whose wisdom shaped my first steps,

whose faith in my dreams made the impossible attainable,

and whose deepest desire is simply to see me healthy and joyful

Chapter 1

INTRODUCTION

Instrumental variable (IV) analysis is a crucial tool in estimating causal relationships by addressing the issue of unmeasured confounders that may bias the results. There are several questions that researchers are naturally interested in when given data that is suitable for an IV analysis, such as randomized controlled trials (RCTs) with non-compliance, encouragement design experiments, Mendelian Randomization studies, etc. Is my data actually compatible with an IV model? What do I know about the average treatment effect (ATE) from my data? Can I make a statistical inference on the ATE? What is the best way to analyze my data? These questions are well-studied when the instrument, exposure, and outcome are all binary, i.e. the binary IV model, across the economics, statistics, and epidemiology literature. However, it lacks a systematic and general exploration when the instrument, exposure, and outcome are categorical.

In this dissertation, we examine several aspects of the instrumental variable model with categorical instrument, exposure, and outcome, including characterization of the IV model and partial identification in Chapters 2 and 3, statistical inference on partial identification bounds in Chapter 4, falsification of the IV model in Chapter 5, and variation independence property of the marginal counterfactual distribution in Chapter 6. We argue that our work contributes to the IV literature in the following aspects.

Firstly, we give a simple closed-form characterization of the joint counterfactual probability distribution with categorical instrument, exposure, and outcome. The characterization consists of necessary and sufficient inequalities relating the joint counterfactual probabilities to the observed probabilities. Since our inequalities are necessary and sufficient, they can be used simultaneously to perform falsification tests of the categorical IV model and to give

partial identification bounds on the ATE.

Secondly, we derive the smallest set of inequalities needed for characterization and partial identification, which corresponds to the facets of the polytope defining the joint counterfactual probability distribution. This significantly improves the computational feasibility of our results.

Thirdly, we aim to develop our results such that they hold under various versions of the assumptions for the IV model that appear in the previous literature. This is especially crucial since the assumptions for IV models are not individually testable.

Lastly, we provide an algorithm to construct confidence regions for any convex functional of the joint counterfactual distribution, using a finite-sample tail bound for the KL-divergence due to Guo and Richardson [2021]. We also discuss the variation independence property for the marginal distribution, which offers a crucial practical guideline on how researchers should obtain partial identification bounds on the ATE.

To the best of our knowledge, this is the first work that studies all aspects above as a whole.

1.1 Motivating Example: the Minneapolis Domestic Violence Experiment

We consider the Minneapolis domestic violence experiment in Sherman and Berk [1984] in which the Minneapolis Police Department and the Police Foundation conducted an experiment from early 1981 to mid-1982 testing police responses to domestic violence and suspect's re-offence status. When the police officers responded to a domestic violence case, they were randomly recommended that the suspects would be arrested, sent from the scene of the assault for eight hours, or given some form of advice, determined by a lottery selection. In the experiment, the responses were named as arrest, send, and advice, while we will name them as arrest, separate, and advice to be consistent with Sherman and Berk [1984] and Angrist [2006]. The study followed up all cases after a 6-month period to see the suspects' re-offense status, which is our primary outcome of interest in this study, through self-reports and the police database. There were a total of 314 cases in the experiment with 92 randomly

assigned to arrest, 108 to advise, and 114 to separate.

Non-compliance is one of the common issues in randomized controlled trials (RCTs), where participants may not adhere to their assigned treatment. Intention-to-treat (ITT) analysis is typically conducted, which includes all participants as originally assigned, regardless of whether they completed the treatment according to the protocol. Such an analysis preserves the benefits of randomization and provides an unbiased estimate of the effect of treatment assignment. However, per protocol (PP) analysis, which considers only those participants who fully adhered to the treatment assignment protocol, can also be of interest as it can offer insights into the treatment's efficacy under ideal conditions. Nonetheless, PP analysis is more challenging to interpret due to potential biases introduced by non-compliance. Instrumental variable analysis is one of the tools that can be used to address the issue of non-compliance and estimate the true efficacy of the treatment with the random assignment being the instrumental variable Z and the actual treatment taken by the participants as the exposure X. In the Minneapolis domestic violence experiment, police officers could deliver a different response than what they were randomly assigned to, resulting in non-compliance in the experiment.

The Minneapolis domestic violence data is shown below in Table 1.1. We use Z to denote the random assignment to Arrest ($Z = Arr$), Advice ($Z = Adv$), and Separate ($Z = Sep$). We use X to denote the actual response taken by the police officers reflecting the differential response recommendation acceptance of Arrest ($X = Arr$), Advice ($X = Adv$), and Separate ($X = Sep$). We use $Y = 2$ to denote the repeated violence of the suspect within the 6-month follow-up period.

Consider the hypothetical (and hopefully not real!) situation in which the researchers are interested in estimating the average causal effect of the two coddled responses, namely Advice vs. Separate, and thought Arrest was not their question of interest. In addition, they thought that problems with binary treatment and outcome were well-studied and easy to analyze. Therefore, they threw away all data related to either being randomly assigned to the Arrest arm or taking the Arrest response (i.e. the first two columns and the first row

Table 1.1: Minneapolis Domestic Violence Experiment

	X=Arr, Y=1	X=Arr, Y=2	X=Adv, Y=1	X=Adv, Y=2	X=Sep, Y=1	X=Sep, Y=2
Z=Arr	81	10	0	0	1	0
Z=Adv	15	3	69	15	3	3
Z=Sep	21	5	4	1	62	20

in Table 1.1) as shown in Table 1.2 and proceeded with using the well-known Balke-Pearl bounds.

Table 1.2: Reduced Minneapolis Domestic Violence Experiment

	X=Adv, Y=1	X=Adv, Y=2	X=Sep, Y=1	X=Sep, Y=2
Z=Adv	69	15	3	3
Z=Sep	4	1	62	20

This approach is quite problematic. Swanson et al. [2015] showed that performing an IV analysis only using a subset of people who are assigned to their treatments of interest and ignoring other possible assignment options may lead to biased results. Conceptually, this is wrong because we are throwing out a subpopulation of people with a certain combination of treatment assignment and treatment taken so that the estimand is changed and restricted to the average treatment effect for the remaining population. In our case, deleting data on treatment group $X = Arr$ and instrument arm $Z = Arr$ means we ignore police officers who would choose the Arrest response even if they were assigned to the Advice response or the Separate response. This is no longer a proper causal estimand since it is defined based on the treatment taken of interest.

Other approaches include combining the treatment groups such that it is binary. For example, Angrist [2006] combined the Advice and Separate into 'Coddled' and performed an analysis comparing Arrest vs. Coddled treatments. However, there is a lack of methods that directly estimate pairwise average treatment effects without any manipulation of the data when the exposure, outcome, or instrument takes multiple levels. This dissertation provides a framework to achieve such estimation.

1.2 Notation

Throughout this dissertation, we use X to denote the treatment/exposure, Y to denote the outcome, and Z to denote the instrument with $|X| = K$, $|Y| = M$, and $|Z| = Q$.

1.3 Organization

The rest of this dissertation is organized as follows. In chapter 2, we provide a set of necessary and sufficient inequalities that characterize the joint counterfactual probability distribution under the categorical IV model. In Chapter 3, we give conditions to identify the redundant inequalities in the characterization set introduced in Chapter 2, such that we can obtain inequalities that correspond to the facets of the polytope for the joint-counterfactual probability distribution. We also compare our results to Russell [2021]. In Chapter 4, we provide an algorithm to obtain confidence regions on partial identification bounds for any convex functionals of the joint counterfactual probability distribution. We implement our methods in the Minneapolis Domestic Violence Experiment data. We study the falsification test and volumes of the IV model when the instrument and exposure take different levels in Chapter 5. In Chapter 6, we explore the variation independence property of the marginal counterfactual probability when the instrument, exposure, and outcome take different levels.

Chapter 2

A CHARACTERIZATION OF THE IV MODEL WITH CATEGORICAL INSTRUMENT, EXPOSURE, AND OUTCOME

2.1 Introduction

This chapter studies the characterization of the IV model and partial identification using instrumental variable models when the instrument, treatment, and outcome are categorical. Let X and Y denote the exposure and outcome of interest respectively. Generally speaking, a variable Z is a valid instrumental variable if certain versions of the following two assumptions hold: (1) an independence condition (or also known as exchangeability condition in the literature): Z is independent of any unmeasured confounder, U, of the exposure-outcome relationship; (2) an exclusion restriction: there is no direct effect of Z on the outcome Y that is not completely through the exposure of interest X. Both of these assumptions are individually untestable. A third relevance assumption is also often referred to in the literature, which states that Z is associated with the exposure X. However, it only comes into play with whether the bound derived is informative and is not useful for other aspects of partial identification. In this work, we will not invoke any monotonicity assumption, which would permit point identification of the effect among the 'compliers' [Imbens and Angrist, 1994, Angrist and Imbens, 1995, Angrist et al., 1996]. A directed acyclic graph (DAG) representing the assumptions on instrumental variables is shown in Figure 2.1.

Instrumental models with instrument, exposure, and outcome all being binary have been well-studied across the economics, statistics, and epidemiology literature. Robins et al. [1989], Manski [1990], Balke and Pearl [1997], and Richardson and Robins [2014] derived sharp lower and upper bounds on the average treatment effect (ATE) under different versions

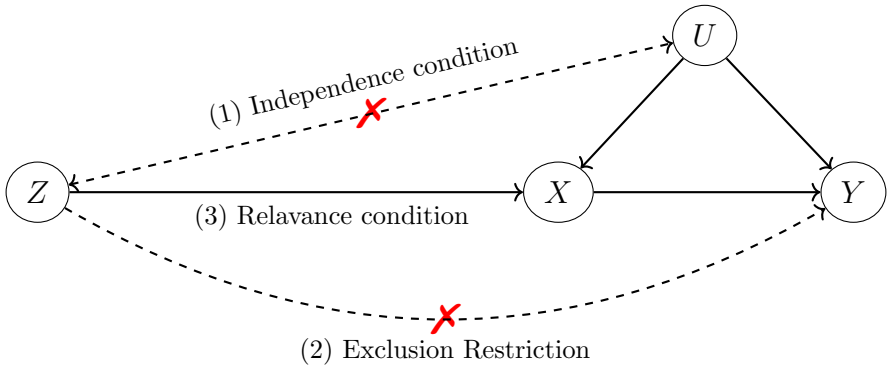


Figure 2.1: Directed acyclic graph (DAG) representing the assumptions of a valid instrumental variable

of the independence and exclusion restriction conditions. Swanson et al. [2018] provided a comprehensive discussion of the underlying assumptions and results of those methods. However, IV models beyond binary instrument, exposure, and outcome are less investigated. Below, we discuss a few works that explored partial identification or falsification tests beyond binary IV models. Bonet [2001] used an optimization program to check that the IV inequalities proposed by Pearl [1995] are necessary but not sufficient in general. In particular, Pearl's IV inequalities are sufficient when the instrument, exposure, and outcome are all binary but not when the instrument takes 3 levels while the exposure and outcome are binary. Richardson and Robins [2014] derived necessary and sufficient bounds on the ATE when the instrument is discrete and takes finite states with binary exposure and outcome. Kédagni and Mourifié [2020] proposed a generalized set of IV inequalities for the joint independence assumption (which will be precisely defined later) and exclusion restriction, and showed they are necessary and sufficient for an observed distribution to be compatible with the IV model with binary outcome. Cheng and Small [2006] derived sharp bounds for causal effects within principal strata in a three-arm trial under the assumption of the monotonicity of compliance. Beresteanu et al. [2012] gave a characterization of a joint potential outcome distribution using random set theory under the joint independence assumption while allowing the outcome to be continuous. However, the practical implementation of random set theory without closed-

form expressions may limit its applicability, and it may require substantial computational resources. Hence, a crucial gap in the literature is a simple closed-form characterization of the IV model beyond binary instrument, exposure, and outcome.

To the best of our knowledge, our work is the first to provide a simple closed-form characterization of the joint counterfactual probability distribution while allowing the instrument, exposure, and outcome to be all discrete, taking finitely many states. The bounds we derived are necessary and sufficient, and they hold under various versions of the independence and exclusion restriction assumptions that have been discussed in the literature. We argue that these results offer valuable insights into the understanding of the fundamentals of IV models and aid the computational development of causal inference.

2.2 Assumptions and Previous Results

2.2.1 Assumptions

Let Z be an instrument variable with Q states, X be an exposure (or treatment) with K states, and Y be an outcome with M states. All of the models that we consider will assume the existence of potential outcomes $Y(x = i, z = q)$ for $i \in \{1, \ldots, K\}$, $q \in \{1, \ldots, Q\}$, corresponding to the value of Y for a subject who (possibly counter-to-fact) receives $Z = j$ and $X = i$. In addition, some of the models assume the existence of potential outcomes $X(z = q)$, giving the value of the exposure X that a subject receives when $Z = q$. We will often use the shorthand notation $Y(x_i, z_q) \equiv Y(x = i, z = q)$ and $X(z = q) \equiv X(z_q)$.

The observed data and potential outcomes are related via the usual consistency relation: $Y = Y(X, Z)$; for models with $X(z)$ potential outcomes, we also have $X = X(Z)$. We also define $Y(x) = Y(x, Z)$ to be the potential outcome for Y arising from an intervention on X alone.

We will also consider a latent variable model, which posits the existence of an unmeasured variable U (with unknown state-space) that represents all variables giving rise to confounding between X and Y.

The instrumental variable model is based on an Exclusion assumption and an Independence assumption. Different forms of these have been considered in the literature:

Assumption 1 (Versions of Exclusion assumption)**.**

(A1-1) Individual-level Exclusion Restriction:

$$Y(x_i, z) = Y(x_i, \tilde{z}) \text{ for all } z, \tilde{z} \in \{1, \dots, Q\}, i \in \{1, \dots, K\}, \text{ and } q \in \{1, \dots, Q\}. \tag{2.1}$$

(A1-2) Joint Stochastic Exclusion:

$$P(Y(x_1, z) = y^1, \dots Y(x_K, z) = y^K) = P(Y(x_1, \tilde{z}) = y^1, \dots Y(x_K, \tilde{z}) = y^K) \tag{2.2}$$

for all $z, \tilde{z} \in \{1, \dots, Q\}$*, and* $y^1, \dots, y^K \in \{1, \dots, M\}$*.*

(A1-3) Latent Exclusion:

$$P(Y(x, z) = y \mid U = u) = P(Y(x, \tilde{z}) = y \mid U = u) \tag{2.3}$$

for all $z, \tilde{z} \in \{1, \dots, Q\}$*,* $x \in \{1, \dots, K\}$ *and* $y \in \{1, \dots, M\}$ *and latent states* u*.*

The strongest version (A1-1) requires that there be no direct effect of Z on Y relative to X at the individual level. The weaker versions (A1-2) and (A1-3) restrict the effect of Z on Y relative to X at the population level. Specifically, version (A1-3) means that the direct effect of Z on Y holding X and a latent variable U fixed are zero at the population level. Hirano et al. [2000] considered (A1-2), the joint stochastic exclusion, in their paper.

Different versions of the independence assumptions have also been considered in different methods and areas (see Swanson et al. [2018] for a review).

Assumption 2 (Versions of Independence assumption)**.**

(A2-1) Random assignment:

$$Z \perp\!\!\!\perp \{Y(x,z), X(z), \text{ for all } x \in \{1,\ldots,K\}, z \in \{1,\ldots,Q\}\}. \tag{2.4}$$

(A2-2) Joint independence:

$$Z \perp\!\!\!\perp \{Y(x,z) \text{ for all } x \in \{1,\ldots,K\}, z \in \{1,\ldots,Q\}\} \tag{2.5}$$

(A2-3) Single-world independence: for $z \in \{1,\ldots,Q\}$, $x \in \{1,\ldots,K\}$,

$$Z \perp\!\!\!\perp \{X(z), Y(x,z)\}. \tag{2.6}$$

(A2-4) Latent-variable exogeneity: there exists U *such that* $U \perp\!\!\!\perp Z$, *and*

$$Y(x,z) \perp\!\!\!\perp (X,Z) \mid U, \tag{2.7}$$

for $z \in \{1,\ldots,Q\}$, $x \in \{1,\ldots,K\}$.

DAGs and Single-World Intervention Graphs (SWIGs) for the independence assumptions listed above can be found in Richardson and Robins [2014]. The Balke-Pearl bounds were derived under (A2-1) but are shown to also hold under (A2-2), (A2-3), and (A2-4) [Richardson and Robins, 2014]. Kitagawa [2021] analyzed the IV model under (A2-2). Some works formulated the IV model with the presence of an unmeasured confounder U between X and Y as defined in (A2-4) including Dawid [2003]. Richardson and Robins [2014] developed a sharp characterization of the joint counterfactual probability distribution $P(Y(x_1), P(Y(x_2))$ given an observed conditional probability $P(X,Y \mid Z)$, which hold under any of the independence conditions (A2-1)–(A2-4). We consider five models $\mathcal{M}_1,\ldots,\mathcal{M}_5$ in this paper given the versions of the exclusion and independence assumption as shown in Table 2.1. We state the relationship among the models, $\mathcal{M}_1,\ldots,\mathcal{M}_5$, in Figure 2.2 and in the Lemma below.

Table 2.1: Instrumental variable models considered in this paper

Model Name		Exclusion	Independence
Randomization	$\mathcal{M}_1$	Individual-level	Random assignment
Joint Ind. & Indiv. Excl.	$\mathcal{M}_2$	Individual-level	Joint independence
Joint Ind. & Stoch. Excl.	$\mathcal{M}_3$	Joint stochastic exclusion	Joint independence
SWIG	$\mathcal{M}_4$	Individual-level	Single-world independence
Latent Model	$\mathcal{M}_5$	Latent Exclusion	Latent-variable exogeneity

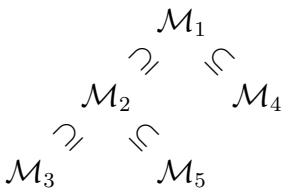


Figure 2.2: Nested structure between models $\mathcal{M}_1$–$\mathcal{M}_5$

Lemma 1. *We have $\mathcal{M}_1 \subseteq \mathcal{M}_2 \subseteq \mathcal{M}_3$, $\mathcal{M}_1 \subseteq \mathcal{M}_4$, $\mathcal{M}_1 \subseteq \mathcal{M}_5$, and $\mathcal{M}_2 \subseteq \mathcal{M}_5$; See Figure 2.2.*

2.2.2 Results from Richardson and Robins [2014]

In this subsection, we review the special case in Richardson and Robins [2014] in which there is a binary exposure X, a binary outcome Y, and a categorical instrument Z that takes Q states.

Theorem 1 (Richardson and Robins [2014])**.** *Consider a categorical IV model with $K = M = 2$ and $Q \geq 2$. Under M_1, M_2, M_4 and M_5, the joint distribution $P(Y(x_1), Y(x_2))$ is characterized by the following $8Q$ inequalities:*

$$P(Y(x_i) = y) \leq P(X = i, Y = y \mid Z = z) + P(X = 3 - i \mid Z = z), \tag{2.8}$$

$$P\left(Y(x_1) = y^1, Y(x_2) = y^2\right) \leq P\left(X = 1, Y = y^1 \middle| Z = z\right) + P\left(X = 2, Y = y^2 \middle| Z = z\right), \tag{2.9}$$

with $y, i, y^1, y^2 \in \{1, 2\}$ *and* $z \in \{1, \ldots, Q\}$.

Given each $Z = z$, since $y, i, y^1, y^2 \in \{1, 2\}$, we have 4 inequalities in the form of (2.8) and (2.9) each, and thus in total $8Q$ inequalities. The $8Q$ inequalities are sufficient and necessary. This result can be shown as a special case of Theorem 2 presented in the next section. In this chapter, we aim to generalize these inequalities to achieve a sharp characterization of a joint counterfactual probability distribution $P(Y(x_1), \ldots, Y(x_K))$ given an observed conditional probability distribution $P(X, Y \mid Z)$ while allowing X, Y, Z to be generically categorical.

2.3 Main Results

We give a characterization of the joint counterfactual distribution for the potential outcomes of Y:

Theorem 2. *Under each of the models* $\mathcal{M}_1, \ldots, \mathcal{M}_5$*, the relationship between the observed distribution* $P(X, Y \mid Z)$ *and the joint counterfactual probability distribution* $P'(Y(x_1), \ldots, Y(x_K))$ *is characterized by the following set of inequalities:*

$$P'\left(Y(x_1) \in \mathcal{V}^{(1)}, \ldots, Y(x_K) \in \mathcal{V}^{(K)}\right) \leq \sum_{i=1}^{K} P\left(X = i, Y \in \mathcal{V}^{(i)} \middle| Z = z\right), \tag{2.10}$$

where $z \in \{1, \ldots, Q\}$, $\mathcal{V}^{(1)}, \ldots, \mathcal{V}^{(K)}$ *are non-empty subsets of* $\{1, \ldots, M\}$ *and for at least one* k, $\mathcal{V}^{(k)}$ *is a strict subset of* $\{1, \ldots, M\}$.

There are $Q((2^M - 1)^K - 1)$ inequalities in this set: here $2^M - 1$ counts the non-empty subsets of $\{1, \ldots, M\}$; the second '-1' arises from the requirement that at least one $\mathcal{V}^{(k)}$ be a strict subset (otherwise the inequality becomes trivial, since both sides are 1). We further note that the left- and right-hand side of all bounds in the form of (2.10) are linear summations of $P'(Y(x_1) = y^1, \ldots, Y(x_K) = y^K)$ and $P(Y = y, X = x \mid Z = z)$, which makes the practical implementation of our bounds efficient.

The inequalities (2.10) are *necessary* in that they are implied by each of the models $\mathcal{M}_1, \ldots, \mathcal{M}_5$. They are also *jointly sufficient*: given a counterfactual distribution

$P'(Y(x_1),\ldots,Y(x_K))$ and an observed distribution $P(X,Y\,|\,Z)$ obeying (2.10), there exists a joint distribution $P''(Z,X,Y(x_1),\ldots,Y(x_K))$ that has margins P' and P and is compatible with each of the models $\mathcal{M}_1$, …, $\mathcal{M}_5$.

Remark 1. *For a given observed distribution $P(X,Y\,|\,Z)$, since the inequalities (2.10) are necessary, if the intersection of the corresponding half-spaces is empty, then the categorical IV model is falsified.*

The proof of sufficiency here uses Strassen's Theorem. Note however, that in this setting a direct application of Strassen's Theorem results in the following larger set of inequalities:

$$P'\left((Y(x_1),\ldots,Y(x_K))\in\widetilde{\mathcal{V}}\right)\leq\sum_{i=1}^{K}P\left(X=i,Y\in\widetilde{\mathcal{V}}^{(i)}\,\middle|\,Z=z\right) \tag{2.11}$$

where $\widetilde{\mathcal{V}}\subsetneq\{1,\ldots,M\}^K$ and $\widetilde{\mathcal{V}}^{(i)}$ is the set of values v^* such that for some $\widetilde{\mathbf{v}}\in\widetilde{\mathcal{V}}$, $v^*=(\widetilde{\mathbf{v}})_i$, the i-th entry in the vector $\widetilde{\mathbf{v}}$. There are $Q(2^{M^K}-1)$ inequalities of the form (2.11). However, since $\widetilde{\mathcal{V}}\subseteq\widetilde{\mathcal{V}}^{(1)}\times\cdots\times\widetilde{\mathcal{V}}^{(K)}$,

$$P'\left((Y(x_1),\ldots,Y(x_K))\in\widetilde{\mathcal{V}}\right)\leq P'\left(Y(x_1)\in\widetilde{\mathcal{V}}^{(1)},\ldots,Y(x_K)\in\widetilde{\mathcal{V}}^{(K)}\right),$$

so every inequality in (2.11) is implied by an inequality of the form (2.10).

In fact, we will show in Theorem 4 in Chapter 3 that when Y has more than two states, so $M>2$, there is still some redundancy in the set (2.10).

Corollary 1. *Let $\{i_{(1)},\ldots,i_{(k)}\}\subseteq\{1,\ldots,K\}$ and $(j_{(1)},\ldots,j_{(k)})$ be a sequence of (not necessarily distinct) indices from $\{1,\ldots,M\}$. Bounds on the marginalized counterfactual prob-*

abilities are given as:

$$P'(Y(x_i) = j) \leq 1 - P(X = i, Y \neq j \mid Z = z), \tag{2.12}$$

$$P'\left(Y(x_i) = j, Y(x_{i'}) = j'\right) \leq 1 - P(X = i, Y \neq j \mid Z = z) - P\left(X = i', Y \neq j' \middle| Z = z\right),$$

$$\vdots \qquad\qquad \vdots$$

$$P'\left(Y(x_{i_{(1)}}) = j_{(1)}, \ldots, Y(x_{i_{(k)}}) = j_{(k)}\right) \leq 1 - P\left(X = i_{(1)}, Y \neq j_{(1)} \middle| Z = z\right) - \ldots -$$
$$P\left(X = i_{(k)}, Y \neq j_{(k)} \middle| Z = z\right).$$

In the special case of $M = 2$ (Y is binary), this is the same set of inequalities as in (2.10). However, this is generally a subset of inequalities given by (2.10).

Example 1. Suppose $K = 2$, $M = 3$, and $\mathcal{V}_1^{(1)} = \{1, 2, 3\}$ and $\mathcal{V}_1^{(2)} = \{1, 2\}$. We also have $\mathcal{V}_1 = \mathcal{V}_1^{(1)} \times \mathcal{V}_1^{(2)} = \{(1,1), (1,2), (2,1), (2,2), (3,1), (3,2)\}$. By Theorem 2, the left-hand side of the bound in the form of (2.10) is

$$\begin{aligned}
&P'(Y(x_1) \in \mathcal{V}^{(1)}, Y(x_2) \in \mathcal{V}^{(2)}) \\
&= P'(Y(x_1) = 1, Y(x_2) = 1) + P'(Y(x_1) = 2, Y(x_2) = 1) + P'(Y(x_1) = 3, Y(x_2) = 1) \\
&\quad + P'(Y(x_1) = 1, Y(x_2) = 2) + P'(Y(x_1) = 2, Y(x_2) = 2) + P'(Y(x_1) = 3, Y(x_2) = 2) \\
&= P'(Y(x_2) = 1) + P'(Y(x_2) = 2) \\
&= P'(Y(x_2) \neq 3).
\end{aligned}$$

The right-hand side of the inequality given an instrument arm $Z = z$ is

$$\begin{aligned}
&\sum_{i=1}^{K} P\left(X = i, Y \in \mathcal{V}^{(i)} \middle| Z = z\right) \\
&= P(X = 1, Y = 1 \mid Z = z) + P(X = 1, Y = 2 \mid Z = z) + P(X = 1, Y = 3 \mid Z = z) \\
&\quad + P(X = 2, Y = 1 \mid Z = z) + P(X = 2, Y = 2 \mid Z = z) \\
&= 1 - P(X = 2, Y = 3 \mid Z = z).
\end{aligned}$$

Together, we have

$$P'(Y(x_2) \neq 3) \leq 1 - P(X = 2, Y = 3 \mid Z = z). \tag{2.13}$$

Similarly, suppose $\mathcal{V}_2^{(1)} = \{1, 2, 3\}$ and $\mathcal{V}_2^{(2)} = \{2, 3\}$. We have $\mathcal{V}_2 = \mathcal{V}_2^{(1)} \times \mathcal{V}_2^{(2)} = \{(1, 2), (2, 2), (3, 2), (1, 3), (2, 3), (3, 3)\}$. Then we have an inequality

$$P'(Y(x_2) \neq 1) \leq 1 - P(X = 2, Y = 1 \mid Z = z). \tag{2.14}$$

□

We want to point out that even though we focused on the setting of instrumental variable analysis and presented our results under IV models, our results still hold under an observational model without any IV. Specifically, we can think of the observational model as an "IV model" with only 1 Z-arm ($Q = 1$). All of our results above still hold with $Q = 1$.

2.3.1 Strassen's Theorem and applications

Strassen's theorem [Strassen, 1965, Friedland et al., 2019] is a fundamental result in probability theory. While the full technical statement of the theorem can be quite complex, involving detailed mathematical notations and conditions, it essentially provides necessary and sufficient conditions on the existence of a probability measure with a given support and marginal. Koperberg [Koperberg, 2024] stated a simplified version of Strassen's theorem for finite sets. We restate their results below.

Definition 1 (Neighbors)**.** *Let $\mathcal{A}$ and $\mathcal{B}$ be sets and $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{B}$ a relation. Then for each $U \subseteq \mathcal{A}$, the set of neighbors of U in $\mathcal{R}$ is denoted by*

$$\mathcal{N}_{\mathcal{R}}(U) = \{\boldsymbol{v} \in \mathcal{B} : (U \times \{\boldsymbol{v}\}) \cap \mathcal{R} \neq \emptyset\}.$$

Definition 2 (Coupling)**.** *Let $\mathcal{A}$ and $\mathcal{B}$ be finite sets, $P_{\mathcal{A}}$ and $P_{\mathcal{B}}$ probability measures on $\mathcal{A}$*

and $\mathcal{B}$ respectively. Then a coupling of $P_{\mathcal{A}}$ and $P_{\mathcal{B}}$ is a probability measure $\hat{P}$ on $\mathcal{A}\times\mathcal{B}$, such that the marginal distributions of $\hat{P}$ correspond to $P_{\mathcal{A}}$ and $P_{\mathcal{B}}$.

Theorem 3 (Strassen's theorem for finite sets [Koperberg, 2024])**.** *Let $\mathcal{A}$ and $\mathcal{B}$ be finite sets, $P_{\mathcal{A}}$ and $P_{\mathcal{B}}$ probability measures on $\mathcal{A}$ and $\mathcal{B}$ respectively and $\mathcal{R}\subseteq\mathcal{A}\times\mathcal{B}$ a relation. Then, there exists a coupling $\hat{P}$ of $P_{\mathcal{A}}$ and $P_{\mathcal{B}}$ that satisfies $\hat{P}(\mathcal{R})=1$ if and only if*

$$P_{\mathcal{A}}(U)\leq P_{\mathcal{B}}(\mathcal{N}_{\mathcal{R}}(U)),\ \textit{for all } U\subseteq\mathcal{A}.$$

Remark 2. *It is easy to show that the set of inequalities $P_{\mathcal{A}}(U)\leq P_{\mathcal{B}}(\mathcal{N}_{\mathcal{R}}(U))$, for all $U\subseteq\mathcal{A}$ is the same as the set of inequalities $P_{\mathcal{B}}(U')\leq P_{\mathcal{A}}(\mathcal{N}_{\mathcal{R}}(U'))$, for all $U'\subseteq\mathcal{B}$, since the probabilities assigned to the outcomes in $\mathcal{A}$ and $\mathcal{B}$ both sum to 1. Consequently, we have, for a given $U\subseteq A$, $P_{\mathcal{A}}(U)\leq P_{\mathcal{B}}(\mathcal{N}_{\mathcal{R}}(U))$ is the same as $P_{\mathcal{B}}(\overline{\mathcal{N}_{\mathcal{R}}(U)})\leq P_{\mathcal{A}}(\overline{U})$ where $\overline{\mathcal{N}_{\mathcal{R}}(U)}\subseteq\mathcal{B}$. Hence, even though Theorem 2 is formulated as upper-bounding the joint counterfactual probabilities using conditional observed probabilities, the set of inequalities actually includes both the upper and the lower bound for a given joint counterfactual probability on the left-hand side of (2.10).*

The next proof uses a simple form of Strassen's theorem for variables taking finitely many values Koperberg [2024]. We first introduce some notation.

For a fixed z, we use $\mathcal{A}_z$ to denote the set of mutually exclusive events

$$\mathcal{A}_z=\{(X(z)=i,Y(x_i)=y):i\in\{1,\dots,K\},y\in\{1,\dots,M\}\},$$

Note that *which* Y potential outcome is present in each event in $\mathcal{A}_z$ depends on the value taken by $X(z)$. Under the the individual level exclusion assumption and consistency when $Z=z$ we have the following equivalence:

$$(X(z)=i,Y(x_i)=y)\quad\Leftrightarrow\quad(X=i,Y=y).$$

Consequently, given $Z = z$, there is a one to one correspondence between observed values for (X, Y) and the events in $\mathcal{A}_z$. Furthermore, under $\mathcal{M}_1$ the observed distribution $P(X, Y|Z = z)$ induces a distribution on the events in $\mathcal{A}_z$; we will use $\widetilde{P}_z$ to denote this induced distribution on the events in $\mathcal{A}_z$. Note that this is *not* a joint distribution on the variables $(X(z), Y(x_1), \ldots, Y(x_K))$, rather it is a distribution on the set of mutually exclusive events $\mathcal{A}_z$ that are defined in terms of these variables.

We use $\mathcal{B}$ to denote the space of potential outcomes $(Y(x_1), \ldots, Y(x_K))$ and we have

$$\mathcal{B} \equiv \{1, \ldots, M\}^K.$$

Subsets of $\mathcal{B}$ describe events of potential outcomes. For example, when $K = 3$, we use $\{(1, 1, 1)\} \subset \mathcal{B}$ to denote the event $\{(Y(x_1) = 1, Y(x_2) = 1, Y(x_3) = 1)\}$. It is obvious that $|\mathcal{A}_z| = MK$ and $|\mathcal{B}| = M^K$.

Given an instrument arm $Z = z$, events in $\mathcal{A}_z$ are in the form of $(X(z) = x, Y(x) = y)$ and events in $\mathcal{B}$ are the form of $\left(Y(x_1) = y^1, \ldots, Y(x_K) = y^K\right)$. Given two outcomes $\boldsymbol{w} \in \mathcal{A}_z$ and $\boldsymbol{v} \in \mathcal{B}$, we will say that $\boldsymbol{w}$ and $\boldsymbol{v}$ are coherent if and only if they assign the same values to any variables that are in common. In other words, they are pairs of (observed event, counterfactual event) where consistency is not violated. Thus for example, $(X(z) = 1, Y(x_1) = 1)$ and $(Y(x_1) = 1, Y(x_2) = 1)$ are coherent, but $(X(z) = 1, Y(x_1) = 1)$ and $(Y(x_1) = 2, Y(x_2) = 1)$ are not.

Let

$$\mathfrak{C} \equiv \{(\boldsymbol{w}, \boldsymbol{v}) \mid \boldsymbol{w} \in \mathcal{A}_z, \boldsymbol{v} \in \mathcal{B}, \boldsymbol{w} \text{ and } \boldsymbol{v} \text{ are coherent}\} \subseteq \mathcal{A}_z \times \mathcal{B}.$$

We may also view $\mathfrak{C}$ as specifying a set of edges in a bi-partite graph; see Figure 2.3 for binary exposure and outcome. Notice that if $(\boldsymbol{w}, \boldsymbol{v}) \in \mathfrak{C}$, then the conjunction of $\boldsymbol{w}$ and $\boldsymbol{v}$ corresponds to an assignment to all three variables $(X(z), Y(x_1), Y(x_2))$ given $Z = z$; see Figure 2.4.

The bipartite graph $\mathcal{G} = (\mathcal{A}_z, \mathcal{B}, \mathfrak{C})$ illustrating pairs of $(\boldsymbol{w}, \boldsymbol{v}) \in \mathcal{A}_z \times \mathcal{B}$ that are coherent like Figure 2.3, which is on a special case of binary X and Y, can be easily generalized.

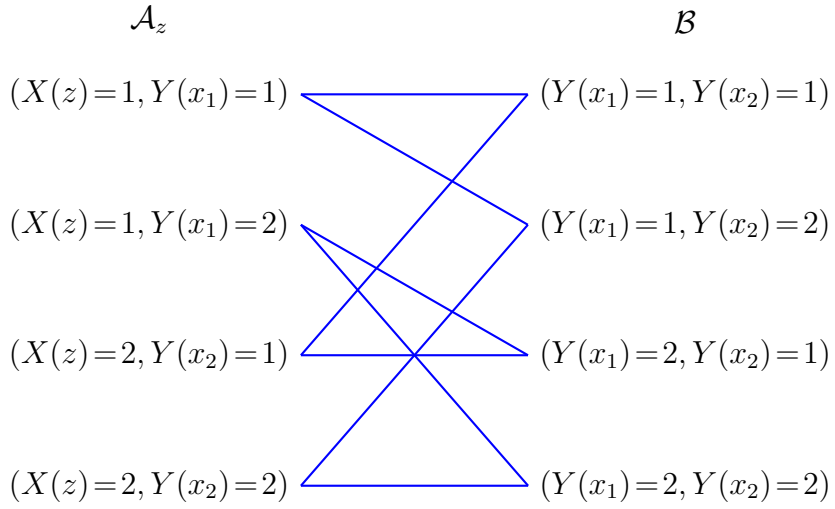


Figure 2.3: Illustration of pairs $(a, b) \in \mathcal{A}_z \times \mathcal{B}$ that are coherent conditional on a given instrument arm $Z = z$. Each edge corresponds to a coherent pair.

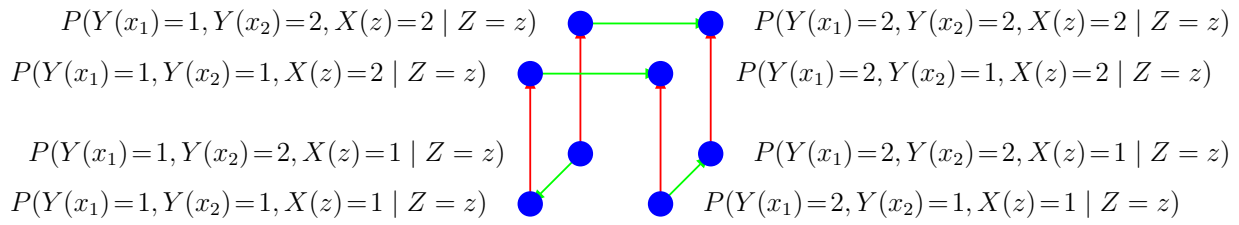


Figure 2.4: A graph in which points in $\mathfrak{C}$ are depicted as blue points; $\mathcal{A}_z$ corresponds to the set of green edges, while $\mathcal{B}$ corresponds to the red edges.

Neighbors are connected by blue lines given coherence. For example, the events $(Y(x_1) = 1, Y(x_2) = 1), (Y(x_1) = 1, Y(x_2) = 2) \in \mathcal{B}$ are both neighbors of $(X(z) = 1, Y(x_1) = 1) \in \mathcal{A}_z$. Each of the MK events in $\mathcal{A}_z$ is connected to M^{K-1} events/neighbors in $\mathcal{B}$, while each of the M^K events in $\mathcal{B}$ is connected to K events/neighbors in $\mathcal{A}_z$. The total number of edges in $\mathcal{G}$ is KM^K. Note that the neighborhood of any event $\mathcal{V} \subseteq \mathcal{B}$, $\mathcal{N}_{\mathfrak{C}}(\mathcal{V})$, contains at least one of the K observed events for each level of X in $\mathcal{A}_z$.

Now we show how using Strassen's theorem can re-establish a key part of the proof in

Richardson and Robins [2024] for binary X and Y. Specifically, it can be used to show the bounds, in the form of (2.8) and (2.9) are sufficient in relating the observed distribution $P(X, Y|Z = z)$ and the potential outcome distribution $P(Y(x_0), Y(x_1))$ when X and Y are binary taking states $\{1, 2\}$. Again, we let $\widetilde{P}_z$ be a probability measure on $\mathcal{A}_z$, and P' be a probability measure on $\mathcal{B}$. We have $\widetilde{P}_z$ specifies a distribution over the following four events in $\mathcal{A}_z$.

$$\begin{aligned}\mathcal{A}_z \equiv \{&(X(z) = 1, Y(x_1) = 1), (X(z) = 1, Y(x_1) = 2),\\ &(X(z) = 2, Y(x_2) = 1), (X(z) = 2, Y(x_2) = 2)\}.\end{aligned}$$

Note that though the outcomes in $\mathcal{A}_z$ do not form a product space, the probabilities assigned to these outcomes by P sum to 1 since:

$$\begin{aligned}\widetilde{P}_z(X(z) = 1, Y(x_1) = 1) + \widetilde{P}_z(X(z) = 1, Y(x_1) = 2) = \widetilde{P}_z(X(z) = 1),\\ \widetilde{P}_z(X(z) = 2, Y(x_2) = 1) + \widetilde{P}_z(X(Z) = 2, Y(x_2) = 2) = \widetilde{P}_z(X(z) = 2).\end{aligned}$$

This is to be expected since, via consistency, the four outcomes in $\mathcal{A}_z$ correspond to observing

$$\{(X = 1, Y = 1), (X = 1, Y = 2), (X = 2, Y = 1), (X = 2, Y = 2)\},$$

in instrument arm $Z = z$.

At the same time P' specifies a distribution over the following four events:

$$\begin{aligned}\mathcal{B} \equiv \{&(Y(x_1) = 1, Y(x_2) = 1), (Y(x_1) = 1, Y(x_2) = 2),\\ &(Y(x_1) = 2, Y(x_2) = 1, (Y(x_1) = 2, Y(x_2) = 2)\}.\end{aligned}$$

Strassen's theorem can thus be used to address the question of whether for a fixed $Z = z$, $\widetilde{P}_z$ and P' are compatible in that there exists a single joint distribution of $\{X(z), Y(x_1), Y(x_2)\}$ that agrees with the distributions on $\mathcal{A}_z$ and $\mathcal{B}$ implied by $\widetilde{P}_z$ and P'. Given the simple char-

acterization given by Koperberg, the necessary and sufficient condition is that:

$$P'(\mathcal{V}) \leq \widetilde{P}_z(\mathcal{N}_{\mathfrak{C}}(\mathcal{V})), \text{ for all } \mathcal{V} \subseteq \mathcal{B}, \tag{2.15}$$

where $\mathcal{N}_{\mathfrak{C}}(\mathcal{V}) \equiv \{\boldsymbol{w} \mid \text{for some } \boldsymbol{v} \in \mathcal{V}, (\boldsymbol{w}, \boldsymbol{v}) \in \mathfrak{C}\}$, is the set of outcomes in $\mathcal{A}_z$ that are neighbors of at least one outcome in $\mathcal{V}$ under $\mathfrak{C}$. Observe that

$$\begin{aligned} &P(X = i, Y = y \mid Z = z) \\ &\quad = P(X(z) = i, Y(x_i, z) = y \mid Z = z) \qquad \text{by consistency} \\ &\quad = \widetilde{P}_z(X(z) = i, Y(x_i) = y), \qquad \text{by assumption A2-1} \end{aligned} \tag{2.16}$$

where the last event is in $\mathcal{A}_z$. The first equality is by consistency; and the second equality is by assumption and A2-1; and the last equality is by summing over $Y(x_{3-i})$. Consequently, the compatibility question addressed by Strassen's theorem regarding $\mathcal{A}_z$ and $\mathcal{B}$ is equivalent to the question of whether assuming A2-1, for a fixed $Z = z$, the observed distribution $P(X, Y \mid Z = z)$ is compatible with a given distribution $P'(Y(x_1), Y(x_2))$.

Finally, we will show in Section 2.4.3 in Lemma 4 that if for all z, $P'(Y(x_1), Y(x_2))$ is compatible with $\widetilde{P}_z(X(z) = i, Y(x_i) = y)$, then there exists a single joint distribution, $\widetilde{P}_z(X(z_1), \ldots, X(z_Q), Y(x_1), Y(x_2))$, that is compatible with the observed distributions in all of the Z arms and $P'(Y(x_1), Y(x_2))$.

We now derive bounds in the form of (2.15) by Strassen's theorem which gives us sufficient conditions for the existence of a single joint distribution of $\{X(z), Y(x_1), Y(x_2)\}$ that satisfies assumptions A1 and A2-1, and agrees with $\widetilde{P}_z$ on set $\mathcal{A}_z$ and P' on set $\mathcal{B}$.

Considering sets $\mathcal{V}$ of cardinality 1 leads to the inequalities:

$$\begin{aligned} &P'\left(Y(x_1) = y^1, Y(x_2) = y^2\right) \\ &\quad \leq \widetilde{P}_z\left(X(z) = 1, Y(x_1) = y^1\right) + \widetilde{P}_z\left(X(z) = 2, Y(x_2) = y^2\right). \end{aligned}$$

Applying (2.16) to the RHS we then obtain the inequality

$$P'\left(Y(x_1) = y^1, Y(x_2) = y^2\right) \\ \leq \widetilde{P}_z\left(X = 1, Y = y^1 \middle| Z = z\right) + \widetilde{P}_z\left(X = 2, Y = y^2 \middle| Z = z\right). \tag{2.17}$$

Given a set $\mathcal{V}$ of cardinality 2, the cardinality of $\mathcal{N}_{\mathfrak{C}}(\mathcal{V})$ is 3 (respectively 4), depending on whether or not the two outcomes do (respectively, do not) share an assignment to a variable. If $|\mathcal{N}_{\mathfrak{C}}(\mathcal{V})| = 4$, then we obtain a trivial inequality. If $|\mathcal{N}_{\mathfrak{C}}(\mathcal{V})| = 3$, then we obtain the following non-trivial inequality:

$$\begin{aligned} &P'(Y(x_i) = y) \\ &\quad = P'(Y(x_i) = y, Y(x_{3-i}) = 1) + P'(Y(x_i) = y, Y(x_{3-i}) = 2) \\ &\quad \leq \widetilde{P}_z(X(z) = i, Y(x_i) = y) + \widetilde{P}_z(X(z) = 3-i, Y(x_{3-i}) = 1) + \widetilde{P}_z(X(z) = 3-i, Y(x_{3-i}) = 2). \end{aligned}$$

Applying (2.16) to the RHS, we get

$$\begin{aligned} &P'(Y(x_i) = y) \\ &\quad \leq P(X = i, Y = y \mid Z = z) + P(X = 3-i, Y = 1 \mid Z = z) + P(X = 3-i, Y = 2 \mid Z = z) \\ &\quad = P(X = i, Y = y \mid Z = z) + P(X = 3-i \mid Z = z). \end{aligned} \tag{2.18}$$

Every set $\mathcal{V}$ of cardinality 3 (or 4) will lead to a trivial inequality since any set of three outcomes in $\mathcal{B}$ will have $|\mathcal{N}_{\mathfrak{C}}(\mathcal{V})| = 4$, so $\widetilde{P}_z(\mathcal{N}_{\mathfrak{C}}(\mathcal{V})) = 1$.

The inequalities (2.17) and (2.18) are exactly (2.8) and (2.9) which define the polytope for the binary potential outcome model in Richardson and Robins [2014].

2.4 Proofs

For simplicity of notation, we use P to denote all probability distributions in this section except for in Lemma 4.

2.4.1 Independence of Z from $Y(x)$

We formulated independence assumptions in terms of potential outcomes $Y(x, z)$. Here we show that, in conjunction with the appropriate exclusion restrictions, these imply independence from $Y(x)$.

Proposition 1. *If Individual level exclusion holds, then:*

- *Random assignment (2.4) is equivalent to*

$$Z \perp\!\!\!\perp \{Y(x), X(z) \text{ for all } x, z\}. \tag{2.19}$$

- *Similarly, Single-world independence (2.6) implies that for all x, z:*

$$Z \perp\!\!\!\perp \{X(z), Y(x)\}. \tag{2.20}$$

Lemma 2. *Joint Stochastic Exclusion (2.2) and Joint Independence (2.5) imply*

$$Z \perp\!\!\!\perp Y(x_1), \ldots, Y(x_K).$$

Proof.

$$\begin{aligned}
&P\left(Y(x_1) = y^1, \ldots, Y(x_K) = y^K \middle| Z = z\right) \\
&\quad = P\left(Y(x_1, z) = y^1, \ldots, Y(x_K, z) = y^K \middle| Z = z\right) \\
&\quad = P\left(Y(x_1, z) = y^1, \ldots, Y(x_K, z) = y^K\right) \\
&\quad = P\left(Y(x_1, \tilde{z}) = y^1, \ldots, Y(x_K, \tilde{z}) = y^K\right) \\
&\quad = P\left(Y(x_1, \tilde{z}) = y^1, \ldots, Y(x_K, \tilde{z}) = y^K \middle| Z = \tilde{z}\right) \\
&\quad = P\left(Y(x_1) = y^1, \ldots, Y(x_K) = y^K \middle| Z = \tilde{z}\right).
\end{aligned}$$

Here the first and fifth equations are by consistency, the second and fourth follow from (2.5), the third is Joint Stochastic Exclusion (2.2). □

Lemma 3. *Latent Stochastic Exclusion (2.3) and Latent Exogeneity (2.7) imply*

$$Y(x) \perp\!\!\!\perp X, Z \mid U.$$

Proof.

$$\begin{aligned}
&P\left(Y(x) = y | \, X = x^*, Z = z^*, U = u\right) \\
&\quad = P(Y(x, z^*) = y \mid X = x^*, Z = z^*, U = u) \\
&\quad = P(Y(x, z^*) = y \mid U = u) \\
&\quad = P(Y(x, z^{**}) = y \mid U = u) \\
&\quad = P(Y(x, z^{**}) = y \mid X = x^{**}, Z = z^{**}, U = u) \\
&\quad = P(Y(x) = y \mid X = x^{**}, Z = z^{**}, U = u)
\end{aligned}$$

Here the first and fifth equations are by consistency, the second and fourth follow from (2.7), the third is Latent Exclusion (2.3). □

2.4.2 Proof of Lemma 1

Proof. Firstly, it is obvious to see that $\mathcal{M}_1$ is the smallest model since assumptions A1-1 and A2-1 are the strongest. Hence, we have $\mathcal{M}_1 \subseteq \mathcal{M}_i$ for $i = 2, 3, 4, 5$. Further, since individual-level restriction implies joint stochastic exclusion, we have $\mathcal{M}_2 \subseteq \mathcal{M}_3$. Moreover, by Proposition 1, Lemma 2 and 3 and Richardson and Robins [2024] (Lemma 1), we know $\mathcal{M}_2 \subseteq \mathcal{M}_5$. □

2.4.3 Proof of Theorem 1

We use $\mathcal{M}_i$ to denote the possible joint distributions $P(Z, X, Y(x_1), \ldots, Y(x_K))$ implied by the exclusion assumption A1 and the independence assumption A2 as shown in Table 2.1. Also, define

$$\phi(\mathcal{M}_i) : P(Z, X, Y(x_1), \ldots, Y(x_K)) \to (P(Y(x_1), \ldots, Y(x_K)), P(X, Y \mid Z)),$$

which maps the joint distribution of Z, X and the potential outcomes of Y to the marginal probabilities over the potential outcomes of Y and the observed distribution of X, Y given Z. We use $\phi(\mathcal{M}_i)$ to denote the image of ϕ of the joint distributions in $\mathcal{M}_i, i = 1, \ldots, 5$. Let $\mathcal{T}$ denote the set of pairs of distributions $(P(Y(x_1), \ldots, Y(x_K)), P(X, Y \mid Z))$ that obey (2.10). We make the observation that Theorem 2 is equivalent to $\phi(\mathcal{M}_i) = \mathcal{T}, i = 1, \ldots, 5$.

To prove $\phi(\mathcal{M}_i) = \mathcal{T}\ \forall i = 1, \ldots, 5$, we prove the following: (i) $\phi(\mathcal{M}_3) \subseteq \mathcal{T}$, (ii)$\phi(\mathcal{M}_4) \subseteq \mathcal{T}$, (iii)$\phi(\mathcal{M}_5) \subseteq \mathcal{T}$, and (iv) $\mathcal{T} \subseteq \phi(\mathcal{M}_1)$. Since we have $\mathcal{M}_1 \subseteq \mathcal{M}_2 \subseteq \mathcal{M}_3$, $\mathcal{M}_1 \subseteq \mathcal{M}_5$, $\mathcal{M}_1 \subseteq M_4$, and $\mathcal{M}_2 \subseteq \mathcal{M}_5$, this is sufficient.

Specifically, (i), (ii) and (iii) prove that the bounds described in equation (2) are necessary, while (iv) shows they are sufficient.

(i) Proof of necessity for joint independence and joint stochastic exclusion model

Claim: $\phi(M_3) \subseteq \mathcal{T}$.

$$\begin{aligned}
&\sum_{i=1}^{K} P\left(X = i, Y \in \mathcal{V}^{(i)} \middle| Z = z\right) \\
&\quad = \sum_{i=1}^{K} P\left(X = i, Y(x_i) \in \mathcal{V}^{(i)} \middle| Z = z\right) \\
&\quad \geq \sum_{i=1}^{K} P\left(X = i, Y(x_1) \in \mathcal{V}^{(1)}, \ldots, Y(x_K) \in \mathcal{V}^{(K)} \middle| Z = z\right) \\
&\quad = P\left(Y(x_1) \in \mathcal{V}^{(1)}, \ldots, Y(x_K) \in \mathcal{V}^{(K)} \middle| Z = z\right) \\
&\quad = P\left(Y(x_1) \in \mathcal{V}^{(1)}, \ldots, Y(x_K) \in \mathcal{V}^{(K)}\right)
\end{aligned}$$

The first equality is by consistency; the second inequality follows by probability; the third equality is by definition of marginalization; the fourth is by Proposition 2. □

(ii) Proof of necessity for the SWIG model

CLAIM: $\phi(M_4) \subseteq \mathcal{T}$.

$$
\begin{aligned}
&\sum_{i=1}^{K} P\left(X = i, Y \in \mathcal{V}^{(i)} \middle| Z = z\right) \\
&\quad = \sum_{i=1}^{K} P\left(X(z) = i, Y(x) \in \mathcal{V}^{(i)} \middle| Z = z\right) \\
&\quad = \sum_{i=1}^{K} P\left(X(z) = i, Y(x_i) \in \mathcal{V}^{(i)}\right) \\
&\quad \geq \sum_{i=1}^{K} P\left(X(z) = i, Y(x_1) \in \mathcal{V}^{(1)}, \ldots, Y(x_i) \in \mathcal{V}^{(i)}, \ldots, Y(x_K) \in \mathcal{V}^{(K)}\right) \\
&\quad = P\left(Y(x_1) \in \mathcal{V}^{(1)}, \ldots, Y(x_i) \in \mathcal{V}^{(i)}, \ldots, Y(x_K) \in \mathcal{V}^{(K)}\right)
\end{aligned}
$$

The first equality is by consistency; the second by Proposition 1; the third inequality follows by probability; the fourth by marginalization over $X(z) = i$. □

(iii) Proof of necessity for the latent model

CLAIM: $\phi(M_5) \subseteq \mathcal{T}$.

$$
\begin{aligned}
&\sum_{i=1}^{K} P\left(X=i, Y \in \mathcal{V}^{(i)} \middle| Z=z\right) \\
&\quad= \sum_{u}\left(\sum_{i=1}^{K} P\left(X=i, Y \in \mathcal{V}^{(i)}, U=u \middle| Z=z\right)\right) \\
&\quad= \sum_{u}\left(\left(\sum_{i=1}^{K} P\left(Y(x_i) \in \mathcal{V}^{(i)} \middle| X=i, U=u, Z=z\right) \cdot P\left(X=i \middle| U=u, Z=z\right)\right)\right. \\
&\qquad \left.\cdot P(U=u \mid Z=z)\right) \\
&\quad= \sum_{u}\left(\left(\sum_{i=1}^{K} P\left(Y(x_i) \in \mathcal{V}^{(i)} \middle| U=u\right) \cdot P(X=i \mid U=u, Z=z)\right) \cdot P(U=u)\right) \\
&\quad\geq \sum_{u}\left(\left(\sum_{i=1}^{K} P\left(Y(x_1) \in \mathcal{V}^{(1)}, \ldots, Y(x_i) \in \mathcal{V}^{(i)}, \ldots, Y(x_K) \in \mathcal{V}^{(K)} \middle| U=u\right)\right.\right. \\
&\qquad \left.\left.\cdot P(X=i \mid U=u, Z=z)\right) \cdot P(U=u)\right) \\
&\quad= \sum_{u}\left(P\left(Y(x_1) \in \mathcal{V}^{(1)}, \ldots, Y(x_K) \in \mathcal{V}^{(K)} \middle| U=u\right) \cdot P(U=u)\right. \\
&\qquad \left.\cdot\left(\sum_{i=1}^{K} P(X=i \mid U=u, Z=z)\right)\right) \\
&\quad= \sum_{u}\left(P\left(Y(x_1) \in \mathcal{V}^{(1)}, \ldots, Y(x_K) \in \mathcal{V}^{(K)} \middle| U=u\right) \cdot P(U=u)\right) \\
&\quad= P\left(Y(x_1) \in \mathcal{V}^{(1)}, \ldots, Y(x_K) \in \mathcal{V}^{(K)}\right)
\end{aligned}
$$

The first and second equations are by algebra and consistency; the third follows since $U \perp\!\!\!\perp Z$ and $Y(x_i) \perp\!\!\!\perp X, Z \mid U$ by Lemma 3; the fourth is a probability inequality; the last three equalities are algebra.

□

(iv) Proof of sufficiency for the randomization model

Claim: $\mathcal{T} \subseteq \phi(\mathcal{M}_1)$

We need to show that given a pair of distributions

$$\mathcal{P} = \left(\tilde{P}(Y(x_1), \ldots, Y(x_K)), \tilde{P}(X, Y \mid Z)\right) \in \mathcal{T},$$

then there exists a joint distribution $P(Z, X(z_1), \ldots, X(z_Q), Y(x_1), Y(x_K))$ in $\mathcal{M}_1$ such that

$$\phi(P(Z, X, Y(x_1), \ldots, Y(x_K))) = \mathcal{P}.$$

The proof strategy breaks this problem down by considering each Z arm in turn. Specifically, in Lemma 4 we show that if for each z, $P(X, Y|Z = z)$ is compatible with $P(Y(x_1), \ldots, Y(x_K))$ in that there exists a compatible joint distribution

$$P(X(z), Y(x_1), \ldots, Y(x_K))$$

then there exists a single joint distribution $P(Z, X(z_1), X(z_Q), Y(x_1), \ldots, Y(x_K))$ over all of the $X(z)$ and $Y(x)$ potential outcomes and the observed distribution $P(X, Y|Z)$.

We first state the following lemma.

Lemma 4. *Given a set of Q distributions $P_q(X(z_q), Y(x_1), \ldots, Y(x_K))$, for $q \in \{1, \ldots, Q\}$ that agree on the common marginals so that we have $P_q(Y(x_1), \ldots Y(x_K)) = P_{q'}(Y(x_1), \ldots Y(x_K))$ for all $q, q' \in \{1, \ldots, Q\}$, then there exists a single joint distribution:*

$$P(X(z_1), \ldots, X(z_Q), Y(x_1), \ldots, Y(x_K))$$

that agrees with each of these Q marginals so that for all q,

$$P_q(X(z_q), Y(x_1), \ldots, Y(x_K)) = P(X(z_q), Y(x_1), \ldots, Y(x_K)).$$

Proof. We may form a joint distribution

$$P^*(X(z_1),\ldots,X(z_Q),Y(x_1),\ldots,Y(x_K)) = \frac{\Pi_{q=1}^{Q} P_q(X(z_q),Y(x_1),\ldots,Y(x_K))}{P_1(Y(x_1),\ldots,Y(x_K))^{Q-1}}$$

The resulting distribution P^* agrees with each P_k on the $(X(z_q),Y(x_1),\ldots,Y(x_K))$ margin.

Though not important for our argument we note that P^* enforces the joint conditional independence of the $X(z)$ counterfactuals given $Y(x_1),\ldots,Y(x_K)$:

$$X(z_1) \perp\!\!\!\perp X(z_2) \perp\!\!\!\perp \cdots \perp\!\!\!\perp X(z_Q) \mid \{Y(x_1),\ldots,Y(x_K)\}.$$

□

Under model $\mathcal{M}_1$ and by Proposition 1 equation (2.19), we have

$$\begin{aligned} P(X(z_q)=i,Y(x_i)=j) &= P(X(z_q)=i,Y(x_i)=j \mid Z=q) \\ &= P(X=i,Y=j \mid Z=q). \end{aligned} \tag{2.21}$$

This Lemma implies that we can consider each level of $Z=q\in\{1,\ldots,Q\}$ separately: if we can construct Q marginal distributions $(X(z_q),Y(x_1),\ldots,Y(x_K))$ which each obey (2.21) and agree on the $(Y(x_1),\ldots,Y(x_K))$ margin, then we can form a single joint distribution. Hence, it remains to show given a pair $\left(\tilde{P}(Y(x_1),\ldots,Y(x_K)),\tilde{P}(X,Y\mid Z=q)\right)$, the satisfaction of inequalities in the form of (2.10) are sufficient to ensure that there exists a joint distribution $P(X(z_q),Y(x_1),\ldots,Y(x_K))$ such that

$$\begin{aligned} P\left(Y(x_1)=y^1,\ldots,Y(x_K)=y^K\right) &= \tilde{P}\left(Y(x_1)=y^1,\ldots,Y(x_K)=y^K\right), \text{ and} \\ P(X(z_q)=i,Y(x_i)=j) &= \tilde{P}(X=i,Y=j\mid Z=q). \end{aligned}$$

As mentioned in Section 2.2, Strassen's Theorem characterizes the conditions which a

joint distribution $P(X(z_q), Y(x_1), \ldots, Y(x_K))$ exists. Hence, by Strassens theorem, it suffices to show for all $\mathcal{V}^{(1)}, \ldots, \mathcal{V}^{(k)} \subset \{1, \ldots, M\}$, the constraints obtained are (2.10).

We have for every $z \in \{1, \ldots Q\}$

$$\begin{aligned} & P\left(Y(x_1) \in \mathcal{V}^{(1)}, Y(x_2) \in \mathcal{V}^{(2)}, \ldots, Y(x_K) \in \mathcal{V}^{(K)}\right) \\ \leq & \sum_{i=1}^{K} P\left(X(z) = i, Y(x_i) \in \mathcal{V}^{(i)}\right) \\ = & \sum_{i=1}^{K} P\left(X = i, Y \in \mathcal{V}^{(i)} \middle| Z = z\right), \end{aligned}$$

where the first inequality is by Strassen's theorem, and the second line is by consistency and assumption A1-1 and A2-1.

The bounds above are sufficient to characterize the joint counterfactual probability distribution based on Strassen's Theorem.

Note that it is essential to have at least one k such that $\mathcal{V}^{(k)}$ is a strict subset of $\{1, \ldots, M\}$ to obtain a non-trivial inequality. Otherwise, the right-hand side of (2.10),

$$\sum_{i=1}^{K} P\left(X = i, Y(x_i) \in \mathcal{V}^{(i)} \middle| Z = z\right) = 1,$$

if $\mathcal{V}^{(1)} = \cdots = \mathcal{V}^{(K)} = \{1, \ldots, M\}$.

□

Chapter 3

NON-REDUNDANT CHARACTERIZATION SET: FACETS OF THE POLYTOPE

3.1 Introduction

Inequalities in the form of (2.10) define half-spaces that the joint counterfactual probability distribution, $P(Y(x_1), \ldots, Y(x_K))$, lives in. Thus, the finite polytope defining the joint counterfactual probability distribution is the bounded intersection of half-spaces implied by inequalities, one for each facet; See Theorem 2.15 (Representation theorem for polytopes) in Ziegler [1995]. Inequalities that do not imply facets of the polytope are thus redundant and can be jointly implied by others in the set.

We give an analogy of how an inequality can be necessary and sufficient, but redundant, given other inequalities in the set; see Figure 3.1. Imagine there is an inequality defining the half-space below the blue plane, and another inequality defining the half-space left of the orange plane. A third inequality defines the half-space that is below the green plane, which touches the intersection line of the blue and orange planes. The inequality defining the green plane is necessary since its violation also implies outside of the region defined by the blue and orange half-spaces. However, it is redundant since it could be jointly implied by the blue and orange half-spaces.

3.2 Main Results

The following theorem characterizes the non-redundant inequalities, i.e. facets of the polytope defining $P(Y(x_1), \ldots, P(Y(x_K))$

Theorem 4. *The set of inequalities (2.10) that satisfy either*

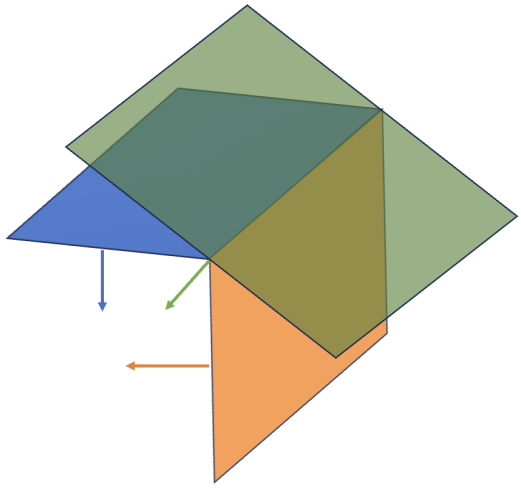


Figure 3.1: A plot showing the inequality corresponding to the green plane is implied by the inequalities corresponding to the blue and orange planes jointly.

1. *for at least two values* k *and* k^**, we have* $\mathcal{V}^{(k)} \neq \{1, \ldots, M\}$ *and* $\mathcal{V}^{(k^*)} \neq \{1, \ldots, M\}$*;*

   or

2. *There exists a* k *and* m *such that* $\mathcal{V}^{(k)} = \{1, \ldots, M\} \setminus \{m\}$ *and for* $k^* \neq k$*, we have* $\mathcal{V}^{(k^*)} = \{1, \ldots, M\}$*. Equivalently, (2.10) then becomes* $P'(Y(x_k) \neq m) \leq 1 - P(X = k, Y = m \mid Z = z)$*.*

are jointly non-redundant.

Equivalently, the inequalities that are redundant take the form $P'(Y(x_k) \notin \{m_1, \ldots, m_J\}) \leq 1 - \sum_{j=1}^{J} P(X = k, Y = j \mid Z = z)$ *where* $J \geq 2$*. There are* $Q(K(2^M - M - 2))$ *of those jointly redundant inequalities. The non-redundant inequalities corresponds to facets of the polytope defining the joint counterfactual probability distribution* $P'(Y(x_1), \ldots, Y(x_K))$*.*

Remark 3. *In the case of* $M = 2$*, the set of inequalities (2.10) are all jointly non-redundant. This is because condition 2 in Theorem 4 is always satisfied since we know there is at least one* k *such that* $\mathcal{V}^{(k)} \neq \{1, 2\}$ *and* $\mathcal{V}^{(k)} \neq \emptyset$*. Otherwise, we will have a trivial inequality.*

In the econometrics literature, Russell [2021] also characterized the set of non-redundant bounds obtaining sharp bounds on convex functionals of the joint counterfactual probability distribution using core determining class theory. We give a detailed discussion comparing the results in section 3.4.

In the next proposition, we quantify the number of necessary, sufficient, and non-redundant inequalities in the form of (2.10) that characterize a joint counterfactual distribution with $K, M, Q \geq 2$.

Proposition 2. *The number of non-trivial constraints in the form of (2.10) is* $Q((2^M-1)^K-1)$. *Thus, the number of necessary, sufficient, non-trivial, and non-redundant inequalities in the form of (2.10) that characterize a joint counterfactual distribution is* $Q((2^M-1)^K - K(2^M - M - 2) - 1)$.

The example below gives a redundant inequality and an illustration of how they are implied by other inequalities.

Example 2. Suppose $K = 2, M = 3$, and $\mathcal{V}_3^{(1)} = \{1, 2, 3\}$ and $\mathcal{V}_3^{(2)} = \{2\}$. We also have $\mathcal{V}_3 = \mathcal{V}_3^{(1)} \times \mathcal{V}_3^{(2)} = \{(1, 2), (2, 2), (3, 2)\}$. By Theorem 2, we have an inequality

$$P'(Y(x_2) = 2) \leq 1 - P(X = 2, Y = 1 \mid Z = z) - P(X = 2, Y = 3 \mid Z = z). \tag{3.1}$$

Checking the conditions given in Theorem 4, we see that neither condition is satisfied, so this inequality is redundant.

Note that equation (2.13) can be rewritten as

$$P'(Y(x_2) = 3) \geq P(X = 2, Y = 3 \mid Z = z), \tag{3.2}$$

and equation (2.14) can be rewritten as

$$P'(Y(x_2) = 1) \geq P(X = 2, Y = 1 \mid Z = z), \tag{3.3}$$

while equation (3.1) can be rewritten as

$$P'(Y(x_2)=1)+P'(Y(x_2)=3)\geq P(X=2,Y=1\mid Z=z)+P(X=2,Y=3\mid Z=z). \tag{3.4}$$

It is clear to see that equation (3.4) is jointly implied by equations (3.2) and (3.3). Hence, inequality (3.1) is jointly implied by inequalities (2.13) and (2.14), and thus it is redundant. □

Repeating the same steps above $\forall \mathcal{V}^{(1)},\ldots,\mathcal{V}^{(K)}\subseteq\{1,\ldots,M\}$ where for at least one k such that $\mathcal{V}^k$ is a strict subset of $\{1,\ldots,M\}$ and $Z=z\in\{1,\ldots,M\}$, we can obtain a set of necessary, sufficient, and non-redundant inequalities which characterize $P'(Y(x_1),Y(x_2),Y(x_3))$ when $K=2$ and $M=3$. Applying Proposition 2, the number of such inequalities is $333Q$ where $|Z|=Q$.

For simplicity of notation, we use P to denote all probability distributions hereafter.

3.3 Proofs

3.3.1 Proof of Theorem 4

First, we restate Proposition 4 with conditions that characterize redundant inequalities, which is equivalent to Theorem 4.

Proposition 3. *For the inequalities in the form of (2.10), they are redundant if and only if :*

1. *there exist a single value k such that $\mathcal{V}^{(k)}\neq\{1,\ldots,M\}$. For all other $k^*\neq k$, we have $\mathcal{V}^{(k)}=\{1,\ldots,M\}$; and if condition 1 holds,*
2. $\left|\mathcal{V}^{(k)}\right|<M-1$.

The redundant inequalities are implied solely by the set of non-redundant inequalities.

It is sufficient to then prove Proposition 3. We let $A \subseteq \mathcal{A}_z$ and $\mathcal{N}_{\mathfrak{C}}(A) \subseteq \mathcal{B}$ be the neighborhood of A. We use $\overline{A}$ to denote the complement of A in $\mathcal{A}_z$, and $\overline{\mathcal{N}_{\mathfrak{C}}(A)}$ for the complement of $\mathcal{N}_{\mathfrak{C}}(A)$ in $\mathcal{B}$. Similarly, let $\mathcal{V} = \mathcal{V}^{(1)} \times \cdots \times \mathcal{V}^{(K)} \subseteq \mathcal{B}$, we use $\overline{\mathcal{N}_{\mathfrak{C}}(\mathcal{V})}$ to denote the complement of $\mathcal{N}_{\mathfrak{C}}(\mathcal{V})$ in $\mathcal{A}_z$.

Suppose we have a distribution P on $\mathcal{A}_z \times \mathcal{B}$ such that $P(\mathfrak{C}) = 1$. Note that given $\mathcal{V} \subseteq \mathcal{B}$, we have an inequality $P(\mathcal{V}) \leq P(\mathcal{N}_{\mathfrak{C}}(\mathcal{V}))$ by Strassen's theorem which is in the form of (2.10). Let $A = \overline{\mathcal{N}_{\mathfrak{C}}(\mathcal{V})}$, and we have $\mathcal{N}_{\mathfrak{C}}(A) = \overline{\mathcal{V}}$. Then, by Strassen's theorem, we have an inequality

$$
\begin{aligned}
&P(A) \leq P(\mathcal{N}_{\mathfrak{C}}(A)) \Leftrightarrow 1 - P(A) \geq 1 - P(\mathcal{N}_{\mathfrak{C}}(A)) \\
&\Leftrightarrow 1 - P(\overline{\mathcal{N}_{\mathfrak{C}}(\mathcal{V})}) \geq 1 - P(\overline{\mathcal{V}}) \Leftrightarrow P(\mathcal{V}) \leq P(\mathcal{N}_{\mathfrak{C}}(\mathcal{V})).
\end{aligned}
$$

Because of this equivalence relation between the two, in the proof below, we will consider the inequalities $P(A) \leq P(\mathcal{N}_{\mathfrak{C}}(A))$ where $A = \overline{\mathcal{N}_{\mathfrak{C}}(\mathcal{V})} \subseteq \mathcal{A}_z$, $\forall \mathcal{V} \subseteq \mathcal{B}$, for the simplicity of the proof construction.

We begin by considering the extreme probability distributions where there is only one counterfactual type $\big(Y(x_1) = y^1, Y(x_2) = y^2, \ldots, Y(x_K) = y^K\big) \in \mathcal{B}$ occurring with probability 1 in the population, and we only observe one of its coherent observed event $(X(z) = x, Y(x) = y) \in \mathcal{A}_z$ with probability 1. Extreme distributions of this type are in bijection with edges in the bi-partite graph. For example, when $K = 2$ and $M = 3$, the extreme probability distributions consist $P(Y(x_1) = 1, Y(x_2) = 2) = 1$ & $P(X(z) = 1, Y(x_1) = 1) = 1, P(Y(x_1) = 1, Y(x_2) = 1) = 1$ & $P(X(z) = 2, Y(x_2) = 1) = 1, P(Y(x_1) = 1, Y(x_2) = 2) = 1$ & $P(X(z) = 2, Y(x_2) = 2) = 1$, etc. Then, given a set of inequalities $\mathcal{I} = \{P(A) \leq P(\mathcal{N}_{\mathfrak{C}}(A)),\ \text{for } A \subseteq \mathcal{A}_z\}$. The extreme points of joint distributions on $\mathfrak{C}$ assign probability 1 to a single point in $\mathfrak{C}$. For a given inequality in $\mathcal{I}$, what corresponds to $P((A, B)) = 1$ is either $P(A) = P(\mathcal{N}_{\mathfrak{C}}(A)) = 1$ or $P(A) = P(\mathcal{N}_{\mathfrak{C}}(A)) = 0$. Each extreme point $(X(z), Y(x_1), \ldots, Y(x_K)$ with $P\left((X(z), Y(x_1), \ldots, Y(x_K)) = \big(x, y^1, \ldots, y^K\big)\right) = 1$ correspond to an edge in $g(\mathcal{A}_z)$. Using a graph $g(A)$ to represent a collection of extreme

distributions when $P(A) = P(\mathcal{N}_{\mathfrak{C}}(A))$, we only include edges connecting A and $\mathcal{N}_{\mathfrak{C}}(A)$ ($P(A) = P(\mathcal{N}_{\mathfrak{C}}(A)) = 1$) and connecting $\overline{A}$ and $\overline{\mathcal{N}_{\mathfrak{C}}(A)}$ ($P(A) = P(\mathcal{N}_{\mathfrak{C}}(A)) = 0$). We denote such a graph by $g(A)$; Figure 3.2 shows various $g(A)$ under different choices of A. An inequality is redundant if and only if the extreme points making it an equality are a subset of the extreme points of another inequality. Graphically, this means there exists another set A' such that $g(A)$ is a subgraph of $g(A')$. Therefore, to show inequalities $P(A) \leq P(\mathcal{N}_{\mathfrak{C}}(A))$ are not redundant, it is equivalent to showing $g(A)$ is not a subgraph of $g(A')$ for all $A' \neq A$.

Proof. We will prove Theorem 4 in four steps. Step I and II show the "only if" part of the statement; Step III shows the "if" part of the statement; Step IV proves the number of redundant constraints.

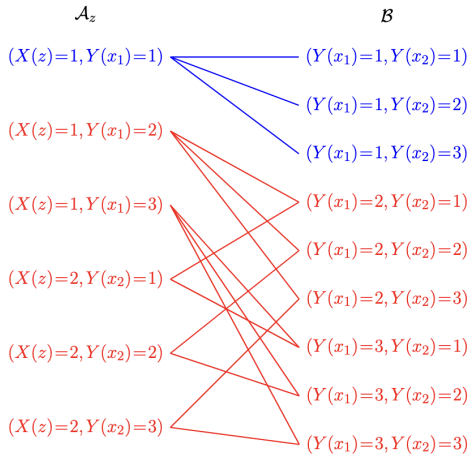
I Let $A \subseteq \mathcal{A}_z$ s.t $|A| = 1$. In this case, condition (1) holds but condition (2) is violated. We will then show the inequalities induced by A and $\mathcal{N}_{\mathfrak{C}}(A)$ are not redundant. Specifically, there is no $A' \neq A$ such that $g(A)$ is a subgraph of $g(A')$; e.g. Figure 3.2 (a).

II Let $A \subseteq \mathcal{A}_z$ s.t $|A| > 1$, and A contains the events at more than 1 x-level. In this case, condition (1) is violated and, We will show the inequalities induced by A and $\mathcal{N}_{\mathfrak{C}}(A)$ are not redundant. Specifically, there is no $A' \neq A$ such that $g(A)$ is a subgraph of $g(A')$; e.g. Figure 3.2 (b).

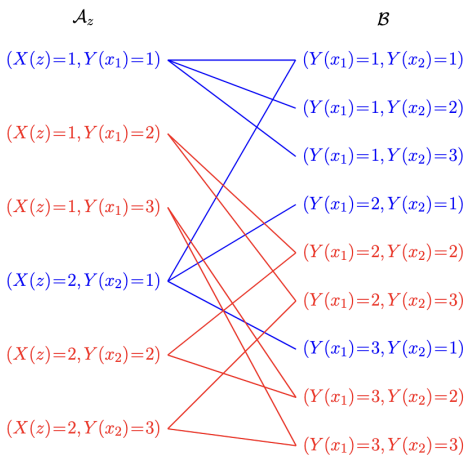
III Let $A \subseteq \mathcal{A}_z$ s.t $|A| > 1$, and A only contains the events at only 1 x-level. In this case, both condition (1) and (2) are satisfied. We will show the inequalities induced by A and $\mathcal{N}_{\mathfrak{C}}(A)$ are redundant. Specifically, for A' such that $A' \subset A$ and $|A'| = 1$, we have $g(A)$ is a subgraph of $g(A')$; e.g. Figure 3.2 (c).

IV We will show the number of redundant constraints in the form of (2.10) is $Q(K(2^M - M - 2)$.

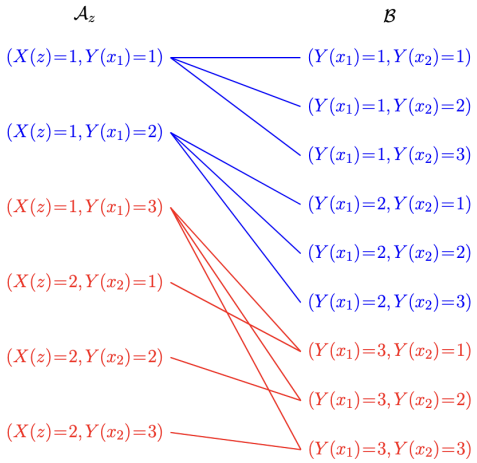
Throughout the proof, we only consider sets A and $\mathcal{N}_{\mathfrak{C}}(A)$ which induce non-trivial inequalities, which are non-empty sets $A \neq \mathcal{A}_z$ and $\mathcal{N}_{\mathfrak{C}}(A) \neq \mathcal{B}$.



(a) Example of Case I where $|A| = 1$. $A = \{(X(z) = 1, Y(x_1) = 1)\}$.



(b) Example of Case II where $|A| > 1$, and A contains the events at more than 1 x-level. $A = \{(X(z) = 1, Y(x_1) = 1), (X(z) = 2, Y(x_2) = 1)\}$.



(c) Example of Case III where $|A| > 1$, and A contains the events at only 1 x-level. $A = \{(X(z) = 1, Y(x_1) = 1), (X(z) = 1, Y(x_2) = 2)\}$.

Figure 3.2: $g(A)$ under various A. We use blue nodes to denote events in A and $\mathcal{N}_{\mathfrak{C}}(A)$ and red nodes to denote events in $\overline{A}$ and $\overline{\mathcal{N}_{\mathfrak{C}}(A)}$. Similarly, we use blue lines to connect $A \leftrightarrow \mathcal{N}_{\mathfrak{C}}(A)$ and red lines to connect $\overline{A} \leftrightarrow \overline{\mathcal{N}_{\mathfrak{C}}(A)}$. Note that Figure (3.2c) is a subgraph of Figure (3.2a).

Proof of I:

Firstly, we remind the readers that $g(A)$ contains edges connecting $A \leftrightarrow \mathcal{N}_{\mathfrak{C}}(A)$ and $\overline{A} \leftrightarrow \overline{\mathcal{N}_{\mathfrak{C}}(A)}$. Comparing $g(A)$ to $g(\mathcal{A}_{\ddagger})$, the edges missing in $g(A)$ are those connecting $\overline{A} \leftrightarrow \mathcal{N}_{\mathfrak{C}}(A)$ since $A \not\leftrightarrow \overline{\mathcal{N}_{\mathfrak{C}}(A)}$ by definition.

Let $|A| = 1$. Note that the inequality implied by A corresponds to $\mathcal{V} = \overline{\mathcal{N}_{\mathfrak{C}}(A)}$ and $\mathcal{N}_{\mathfrak{C}}(\mathcal{V}) = \overline{A}$, so there exists a single k such that $\mathcal{V}^{(K)} \neq \{1, \dots, M\}$ but $\left|\mathcal{V}^{(k)}\right| = M - 1$. This satisfies condition (1) but violates condition (2). Without loss of generality, suppose let $A = \{(X(z) = 1, Y(x_1) = 1)\}$. Then we know

$$\mathcal{N}_{\mathfrak{C}}(A) = \left\{ \left(Y(x_1) = 1, Y(x_2) = y^2, \dots, Y(x_K) = y^K\right) : y^2, \dots, y^K \in \{1, \dots, M\} \right\}.$$

Suppose for a contradiction that there exists a non-empty set A' that gives a non-trivial inequality such that $A' \neq \mathcal{A}_z$ and $\mathcal{N}_{\mathfrak{C}}(A') \neq \mathcal{B}$ where $g(A')$ is a supergraph of $g(A)$. We know there exists at least one edge in $g(A')$ connecting $(X(z) = i, Y(x_i) = y^i) \leftrightarrow (\dots, Y(x_i) = y^i, \dots)$ where $(X(z) = i, Y(x_i) = y^i) \in \overline{A}$ and $(\dots, Y(x_i) = y^i, \dots) \in \mathcal{N}_{\mathfrak{C}}(A)$. Since events $(X(z) = i, Y(x_i) = y^i) \in \overline{A}$ with $i = 1$ and $y^i \neq 1$ are not compatible with events in $\mathcal{N}_{\mathfrak{C}}(A)$, we know $i \neq 1$. Specifically, without loss of generality, we let the edge to be $(X(z) = 2, Y(x_2) = y^*) \leftrightarrow \left(Y(x_1) = 1, Y(x_2) = y^*, \dots, Y(x_K) = y^K\right)$ in $g(A')$. Note that here we have $\left(Y(x_1) = 1, Y(x_2) = y^*, \dots, Y(x_K) = y^K\right) \in \mathcal{N}_{\mathfrak{C}}(A)$ since $\mathcal{N}_{\mathfrak{C}}(A)$ includes all nodes with $Y(x_1) = 1$. Since $g(A')$ is a supergraph of $g(A)$, we know $g(A')$ includes all edges in $g(A)$. The edge $(X(z) = 2, Y(x_2) = 1) \leftrightarrow \left(Y(x_1) = 1, Y(x_2) = y^*, \dots, Y(x_K) = y^K\right)$ in $g(A')$ either connects I. $A' \leftrightarrow \mathcal{N}_{\mathfrak{C}}(A')$ or II. $\overline{A'} \leftrightarrow \overline{\mathcal{N}_{\mathfrak{C}}(A')}$. We show that in both cases, we will have trivial inequalities which is a contradiction.

We know, in $g(A)$, the event $(X(z) = 2, Y(x_2) = y^*)$ connects to events $\left\{\left(Y(x_1) = y^1, Y(x_2) = y^*, \dots, Y(x_K) = y^K\right) : y^1 \neq 1 \text{ and } y^3, \dots, y^K \in \{1, \dots, M\}\right\}$, which are not neighbors of A, since these are edges connecting $\overline{A} \leftrightarrow \overline{\mathcal{N}_{\mathfrak{C}}(A)}$. Therefore, we know edges $(X(z) = 2, Y(x_2) = y^*) \leftrightarrow \{\left(Y(x_1) = y^1, Y(x_2) = y^*, \dots, Y(x_K) = y^K\right) : y^1 \neq 1 \text{ and } y^3, \dots, y^K \in$

$\{1,\ldots,M\}\}$ are in $g(A)$ and hence $g(A')$. In addition, we know edges $(X(z)=2, Y(x_2)=1) \leftrightarrow \{\big(Y(x_1)=1, Y(x_2)=y^*,\ldots,Y(x_K)=y^K\big) : y^3,\ldots,y^K \in \{1,\ldots,M\}\}$ are in $g(A')$ by construction. Note that sets

$$
\begin{aligned}
&\big\{\big(Y(x_1)=y^1, Y(x_2)=y^*,\ldots,Y(x_K)=y^K\big) : y^1 \neq 1 \text{ and } y^3,\ldots,y^K \in \{1,\ldots,M\}\big\} \cup \\
&\big\{\big(Y(x_1)=1, Y(x_2)=y^*,\ldots,Y(x_K)=y^K\big) : y^3,\ldots,y^K \in \{1,\ldots,M\}\big\} \\
=&\big\{\big(Y(x_1)=1, Y(x_2)=y^*,\ldots,Y(x_K)=y^K\big) : y^1, y^3,\ldots,y^K \in \{1,\ldots,M\}\big\}.
\end{aligned}
$$

Hence, in $g(A')$, the event $(X(z)=2, Y(x_2)=y^*)$ connects to all the events in $B_1 = \big\{\big(Y(x_1)=y^1, Y(x_2)=y^*,\ldots,Y(x_K)=y^K\big) : y^1, y^3,\ldots,y^K \in \{1,\ldots,M\}\big\}$. Further, let $A_1 = \{(X(z)=1, Y(x_1)=y^1) : y^1 \in \{1,\ldots,M\}\}$. Since events in $A_1 \setminus A \equiv \{(X(z)=1, Y(x_1)\neq 1)\}$ are connected to $B_1 \setminus \{\big(Y(x_1)=1, Y(x_2)=y^*,\ldots,Y(X_K)=y^K\big) : y^3,\ldots,y^K \in \{1,\ldots,M\}\}$ while $\{\big(Y(x_1)=1, Y(x_2)=y^*,\ldots,Y(X_K)=y^K\big) : y^3,\ldots,y^K \in \{1,\ldots,M\}\}$ is connected to A, we know events in B_1 connects to all the events in A_1 in $g(A)$ and thus in $g(A')$ where we have $\mathcal{N}_{\mathfrak{C}}(A_1)=\mathcal{B}$.

Case I: If $(X(z)=2, Y(x_2)=1) \leftrightarrow \big(Y(x_1)=1, Y(x_2)=1,\ldots,Y(x_K)=y^K\big)$ is an edge connecting $A' \leftrightarrow \mathcal{N}_{\mathfrak{C}}(A')$, then we know $(X(z)=2, Y(x_2)=1) \in A'$, so we have $B_1 \subseteq \mathcal{N}_{\mathfrak{C}}(A')$ since these are connected in $g(A')$. Further, we have $A_1 \subseteq A'$ since events in A_1 are connected to B_1 in $g(A')$. Consequently, we have $\mathcal{N}_{\mathfrak{C}}(A')=\mathcal{B}$ since $\mathcal{N}_{\mathfrak{C}}(A_1)=\mathcal{B}$, which is a contradiction since the inequality would be trivial.

Case II: If $(X(z)=2, Y(x_2)=1) \leftrightarrow \big(Y(x_1)=1, Y(x_2)=1,\ldots,Y(x_K)=y^K\big)$ is an edge connecting $\overline{A'} \leftrightarrow \overline{\mathcal{N}_{\mathfrak{C}}(A')}$, then we know $(X(z)=2, Y(x_2)=1) \in \overline{A'}$, so we have $B_1 \subseteq \overline{\mathcal{N}_{\mathfrak{C}}(A')}$ since they are connected in $g(A')$. Further, we have $A_1 \subseteq \overline{A'}$ since events in A_1 are connected to B_1 in $g(A')$. Consequently, we will have $\overline{\mathcal{N}_{\mathfrak{C}}(A')}=\mathcal{B}$ and $A'=\emptyset$, which is again a contradiction since the inequality would be trivial.

Proof of II:

We first introduce the following lemmas.

Lemma 5. *If A is a non-trivial inequality, then $\overline{A}$ contains events with every x-level. Or in other words, for every x-level, there exists a y such that $(X(z) = x, Y(x) = y)$ is not in A.*

Proof. If A leads to a non-trivial inequality, then there is at least one type, $\big(Y(x_1) = y^1, \ldots, Y(x_K) = y^K\big)$, in $\mathcal{B}$ that is not a neighbor of A. The set of y values for each x-level in $\big(Y(x_1) = y^1, \ldots, Y(x_K) = y^K\big)$ satisfies the claim. Otherwise, there exists and x-level such that $\{(X(z) = x, Y(x) = y) : y \in \{1, \ldots, M\}\} \subseteq A$, and we have $\mathcal{N}_{\mathfrak{C}}(A) = \mathcal{B}$, which leads to a trivial inequality. □

Lemma 6. *Suppose $A \neq \mathcal{A}$ and $\mathcal{N}_{\mathfrak{C}}(A) \neq \mathcal{B}$.*

1. *There exists a path in $g(A)$ connecting any event in $\overline{A}$ to any event $\overline{\mathcal{N}_{\mathfrak{C}}(A)}$.*

2. *If $|A| > 1$ and A contains events with more than 1 x-level, then there exists a path in $g(A)$ connecting any event in A to any event $\mathcal{N}_{\mathfrak{C}}(A)$.*

Proof. The key thing in the Lemma and in the proof is that the sets of interest in $\mathcal{A}$, namely A and $\overline{A}$, contain more than 1 x-level, which automatically holds for $\overline{A}$ that doesn't induce trivial inequalities; See Lemma 5.

Denote $A = \{(X(z) = \alpha, Y(\alpha) = \beta) : (\alpha, \beta) \in \{(i_1, j_1), \ldots, (i_n, j_n)\}\}$ where $n = |A|$. A key observation is that in $g(A)$ all events $(X(z) = \alpha^*, Y(\alpha) = \beta^*)$ in $\overline{A}$ connect to all events in $\overline{\mathcal{N}_{\mathfrak{C}}(A)} \cap \{Y(\alpha^*) = \beta^*, \ldots\}$. Similarly, in $g(A)$, all events $(X(z) = \alpha', Y(\alpha) = \beta')$ in A are adjacent to $\mathcal{N}_{\mathfrak{C}}(A) \cap \{Y(\alpha') = \beta'\}$. Further note that for events $(X(z) = \alpha^*, Y(\alpha^*) = \beta^*) \in \overline{A}$ with $\alpha \in [i_1, \ldots, i_n]$ but $(\alpha^*, \beta^*) \notin A$, it is possible that all events in $\overline{\mathcal{N}_{\mathfrak{C}}(A)}$ has $Y(\alpha^*) = \beta^*$ and the event $(X(z) = \alpha^*, Y(\alpha^*) = \beta^*)$ connects to all events in $\overline{\mathcal{N}_{\mathfrak{C}}(A)}$.

First, we prove Lemma 6.1 that there exists a path connecting any event in $\overline{A}$ to any event $\overline{\mathcal{N}_{\mathfrak{C}}(A)}$. Denote an event $(X(z) = \alpha_1, Y(\alpha_1) = \beta_1) \in \overline{A}$. We know in $g(A)$, we have $(X(z) = \alpha_1, Y(\alpha_1) = \beta_1) \leftrightarrow \overline{\mathcal{N}_{\mathfrak{C}}(A)} \cap \{Y(\alpha_1) = \beta_1\}$. Therefore, it remains to show $(X(z) = \alpha_1, Y(\alpha_1) = \beta_1) \leftrightarrow \overline{\mathcal{N}_{\mathfrak{C}}(A)} \cap \{Y(\alpha_1) \neq \beta_1\}$. For an arbitrary event $(Y(\alpha_1) = \gamma, Y(\alpha_2) = \beta_2, \ldots, Y(\alpha_K) = \beta_K) \in \overline{\mathcal{N}_{\mathfrak{C}}(a)} \cap \{Y(\alpha_1) \neq \beta_1\}$ where $\gamma \neq \beta_1$, we know events

$(X(z) = \alpha_1, Y(\alpha_1) = \gamma), (X(z) = \alpha_2, Y(\alpha_2) = \beta_2), \ldots, (X(z) = \alpha_K, Y(\alpha_K) = \beta_K) \in \overline{A}$. Hence, we have a line $(X(z) = \alpha_2, Y(\alpha_2) = \beta_2) \leftrightarrow (Y(\alpha_1) = \gamma, Y(\alpha_2) = \beta_2, \ldots, Y(\alpha_K) = \beta_K)$ in $g(A)$ since it is a line connecting $\overline{A} \leftrightarrow \overline{\mathcal{N}_{\mathfrak{C}}(A)}$. Further, since $(X(z) = \alpha_1, Y(\alpha_1) = \beta_1) \in \overline{A}$ and $(X(z) = \alpha_2, Y(\alpha_2) = \beta_2), \ldots, (X(z) = \alpha_K, Y(\alpha_K) = \beta_K) \in \overline{A}$, we have $(Y(\alpha_1) = \beta_1, Y(\alpha_2) = \beta_2, \ldots, Y(\alpha_K) = \beta_K) \in \overline{\mathcal{N}_{\mathfrak{C}}(A)}$, so we have a line $(X(z) = \alpha_2, Y(\alpha_2) = \beta_2) \leftrightarrow (Y(\alpha_1) = \beta_1, Y(\alpha_2) = \beta_2, \ldots, Y(\alpha_K) = \beta_K)$ in $g(A)$. Hence, in $g(A)$, we have $(Y(\alpha_1) = \gamma, Y(\alpha_2) = \beta_2, \ldots, Y(\alpha_K) = \beta_K) \leftrightarrow (X(z) = \alpha_2, Y(\alpha_2) = \beta_2) \leftrightarrow (Y(\alpha_1) = \beta_1, Y(\alpha_2) = \beta_2, \ldots, Y(\alpha_K) = \beta_K) \leftrightarrow (X(z) = \alpha_1, Y(\alpha_1) = \beta_1)$. Since $(Y(\alpha_1) = \gamma, Y(\alpha_2) = \beta_2, \ldots, Y(\alpha_K) = \beta_K)$ is arbitrary, we know $(X(z) = \alpha_1, Y(\alpha_1) = \beta_1) \leftrightarrow \overline{\mathcal{N}_{\mathfrak{C}}(A)} \cap \{Y(\alpha_1) \neq \beta_1\}$

Then, we prove Lemma 6.2 that if $|A| > 1$ and A contains events with more than 1 x-level, then there exists a path connecting any event in A to any event $\mathcal{N}_{\mathfrak{C}}(A)$. Since any event in $\mathcal{N}_{\mathfrak{C}}(A)$ is connected to at least one event in A, it suffices to show all events in A are connected. Consider events $(X(z) = \alpha_1, Y(\alpha_1) = \beta_1)$ and $(X(z) = \alpha_2, Y(\alpha_1) = \beta_2)$ in A. If $\alpha_1 \neq \alpha_2$, then we have $(X(z) = \alpha_1, Y(\alpha_1) = \beta_1) \leftrightarrow (Y(\alpha_1) = \beta_1, Y(\alpha_2) = \beta_2, \ldots)$, since $(Y(\alpha_1) = \beta_1, Y(\alpha_2) = \beta_2, \ldots) \in \mathcal{N}_{\mathfrak{C}}((X(z) = \alpha_1, Y(\alpha_1) = \beta_1)) \cap \mathcal{N}_{\mathfrak{C}}((X(z) = \alpha_2, Y(\alpha_2) = \beta_2))$. If $\alpha_1 = \alpha_2$, then we know there exists an event $(X(z) = \alpha_3, Y(\alpha_3) = \beta_3) \in A$ where $\alpha_3 \neq \alpha_1$ and $\alpha_3 \neq \alpha_2$ since we know A contains events with more than 2 x-levels. Since $\alpha_3 \neq \alpha_1$ and $\alpha_3 \neq \alpha_2$, we know $(X(z) = \alpha_3, Y(\alpha_3) = \beta_3)$ is connected with both $(X(z) = \alpha_1, Y(\alpha_1) = \beta_1)$ and $(X(z) = \alpha_2, Y(\alpha_2) = \beta_2)$ by similar arguments above. Hence, the events $(X(z) = \alpha_1, Y(\alpha_1) = \beta_1)$ and $(X(z) = \alpha_2, Y(\alpha_2) = \beta_2)$ are connected as well.

Note that if there exists a path between two events in $\mathcal{A}$ and $\mathcal{B}$, then they are in A and $\mathcal{N}_{\mathfrak{C}}(A)$ respectively or in $\overline{A}$ and $\overline{\mathcal{N}_{\mathfrak{C}}(A)}$ respectively as stated in Lemma 6.

□

Lemma 7. *If there exists a path between two events in $\mathcal{A}$ and $\mathcal{B}$, then they are in A and $\mathcal{N}_{\mathfrak{C}}(A)$ respectively, or in $\overline{A}$ and $\overline{\mathcal{N}_{\mathfrak{C}}(A)}$ respectively by definition.*

Suppose $|A| > 1$, and A contains events with more than 1 x-level and leads to a non-

trivial inequality. Since $\mathcal{V} = \overline{\mathcal{N}_{\mathfrak{C}}(A)}$ and $\mathcal{N}_{\mathfrak{C}}(\mathcal{V}) = \overline{A}$, there exists at least two k, k^* such that $\mathcal{V}^{(k)} \neq \{1, \ldots, M\}$ and $\mathcal{V}^{(k^*)} \neq \{1, \ldots, M\}$, which satisfies condition (1) in Theorem 4. This also means at least two x-levels are not marginalized in $\overline{A}$ since they have events in A. We want to show $g(A)$ is not a subgraph of $g(A')$ for all $A' \neq A$, where $A' \subset \mathcal{A}_z$ and is non-empty.

Case ①: Let $A' \not\subset A$. It is sufficient to show $g(A)$ contains edges $\overline{A'} \leftrightarrow \mathcal{N}_{\mathfrak{C}}(A')$ which are not in $g(A')$. We will show that if there exists A' such that $g(A')$ is a supergraph of $g(A)$, then the inequality induced by A' is trivial which is a contradiction. Since $A' \not\subset A$, there exists $a' \in A'$ such that $a' \in \overline{A}$. Denote $a' = \{(X(z) = i, Y(x_i) = y^i)\}$, and $B = \mathcal{N}_{\mathfrak{C}}(a') = \{\big(Y(x_1) = y^1, \ldots, Y(x_i) = y^i, \ldots, Y(x_K) = y^K\big) : y^1, \ldots, y^{i-1}, y^{i+1}, \ldots, y^K \in \{1, \ldots, M\}\}$.

We partition B to B_1, B_2 such that $B_1 \equiv \mathcal{N}_{\mathfrak{C}}(A) \cap B$ and $B_2 \equiv B \setminus \mathcal{N}_{\mathfrak{C}}(A) \subseteq \overline{\mathcal{N}_{\mathfrak{C}}(A)}$. Since $|A| > 1$ and A contains more than 1 x-level, we have $B_1, B_2 \neq \emptyset$. Firstly, B_1 is non-empty because A contains events with more than two x-level. Therefore, there is an event in A at a different level in an event in A to that in a'. Hence, there exists a common neighbor in B_1. Regarding B_2, since we suppose A does not lead to a trivial inequality, so by Lemma 5, there exists a type in $\mathcal{B}$ that is not a neighbor of A. We then consider switching the value at $Y(x_i)$ to be y_i to establish $B_2 \neq 0$. By Lemma 6(2) applied to A and $\mathcal{N}_{\mathfrak{C}}(A)$, we know there is a path connecting any $b_1 \in B_1 \subseteq B$ to all $a \in A$ in $g(A)$. Since by assumption $g(A')$ is a supergraph of $g(A)$ and $B_1 = \mathcal{N}_{\mathfrak{C}}(A) \cap \mathcal{N}_{\mathfrak{C}}(a') \subseteq \mathcal{N}_{\mathfrak{C}}(A) \cap \mathcal{N}_{\mathfrak{C}}(A')$, so by Lemma 7, we have $A \subseteq A'$, and thus $\mathcal{N}_{\mathfrak{C}}(A) \subseteq \mathcal{N}_{\mathfrak{C}}(A')$. By construction, there are edges connecting $a' \leftrightarrow B_2$ are in $g(A')$ which are edges connecting $\overline{A} \leftrightarrow \overline{\mathcal{N}_{\mathfrak{C}}(A)}$ in $g(A')$. Since $a' \in A'$ we have $B_2 \subseteq \mathcal{N}_{\mathfrak{C}}(A')$. Then, by Lemma 6(1), we know for any event in $B_2 \subseteq \overline{\mathcal{N}_{\mathfrak{C}}(A)}$, there exists a path in $g(A)$, and thus also in $g(A')$, that connects to $\overline{a}$, for all $\overline{a} \in \overline{A}$. Note that since $g(A')$ only has edges $A' \leftrightarrow \mathcal{N}_{\mathfrak{C}}(A')$ and $\overline{A'} \leftrightarrow \overline{\mathcal{N}_{\mathfrak{C}}(A')}$ and $B_2 \subseteq \mathcal{N}_{\mathfrak{C}}(A')$, we have $\overline{A} \subseteq A'$. Thus, since $A \subseteq A'$, we have $\overline{A} \cup A \subseteq A'$, so $A' = \mathcal{A}_{\ddagger}$ which leads to a trivial inequality.

Case ②: Let $A' \subset A$. Since $A' \leftrightarrow \overline{\mathcal{N}_{\mathfrak{C}}(A')}$ is not possible, we need to show $g(A)$ contains edges $\overline{A'} \leftrightarrow \mathcal{N}_{\mathfrak{C}}(A')$ which are by construction not in $g(A')$. Since $A' \subset A$, there exists b such that $b \in A$ but $b \notin A'$, so $b \in A \cap \overline{A'}$. If A' doesn't contain events with all

levels of X present in A, then there exists $b \in A \cap \overline{A'}$ with event $(X(z) = x', Y(x') = y')$ where there is an event with $X(z) = x'$ in A but no event with $X(z) = x'$ is present in A'. Define x_1 to be a level of X occurring in A'. Let $(X(z) = x_1, Y(x_1) = y^1)$ be an event in A'. Therefore, there exists an event $(Y(x_1) = y^1, \dots, Y(x') = y', \dots)$ in $\mathcal{N}_{\mathfrak{C}}(b) \subset \mathcal{N}_{\mathfrak{C}}(A)$, so we also have $(Y(x_1) = y^1, \dots, Y(x') = y', \dots) \in \mathcal{N}_{\mathfrak{C}}(A')$. Hence, we have an edge in $g(A)$, $(X(z) = x', Y(x') = y') \leftrightarrow (Y(x_1) = y^1, \dots, Y(x') = y', \dots)$, that connects $\overline{A'}$ and $\mathcal{N}_{\mathfrak{C}}(A')$ which is not in $g(A')$. If A' contains events with all levels of X present in A, then without loss of generality there exists $b \in A \cap \overline{A'}$ with $(X(z) = x_1, Y(x_1) = y^2)$. Let x' be a level different from x_1 and let y' be a level such that $(X(z) = x', Y(x') = y') \in A'$. Therefore, we have an event $(Y(x_1) = y^2, Y(x') = y', \dots) \in \mathcal{N}_{\mathfrak{C}}(b) \subset \mathcal{N}_{\mathfrak{C}}(A)$. Since $(X(z) = x', Y(x') = y') \in A'$, we also have $(Y(x_1) = y^2, Y(x') = y', \dots) \in \mathcal{N}_{\mathfrak{C}}(A')$. Hence, we have an edge in $g(A)$, $(X(z) = x_1, Y(x_1) = y^2) \leftrightarrow (Y(x_1) = y^2, Y(x') = y', \dots)$ that connects $\overline{A'}$ and $\mathcal{N}_{\mathfrak{C}}(A')$ which is thus not in $g(A')$.

Proof of III:

Let $|A| = r > 1$, and A only contains events that have the same level of $X(z) = x$. This means $|C \equiv \Phi(\mathcal{V}, \{1, \dots, M\})| = K - 1$, and $|\operatorname{Uniq}(\mathcal{V}^{(\{1,\dots K\} \setminus C(\mathcal{V}))})| < M - 1$, which satisfies both condition (1) and (2). Without loss of generality, assume $A = \{(X(z) = x_1, Y(x_1) = y^1), (X(z) = x_1, Y(x_1) = y^2), (X(z) = x_1, Y(x_1) = y^3), \dots (X(z) = x_1, Y(x_1) = y^r\} = \{a_1, a_2, a_3, \dots a_r\}$. We know $\mathcal{N}_{\mathfrak{C}}(a_i) \cap \mathcal{N}_{\mathfrak{C}}(a_j) = \emptyset$ for $i \neq j$. Let $A' = \{a_1\}$. It is sufficient to show that $g(A)$ is a subgraph of $g(A')$, which means all edges in $g(A)$ are in $g(A')$. Again, we remind the readers that $g(A)$ contains edges connecting A and $\mathcal{N}_{\mathfrak{C}}(A)$ as well as edges connecting $\overline{A}$ and $\overline{\mathcal{N}_{\mathfrak{C}}(A)}$. First, consider edges connecting A and $\mathcal{N}_{\mathfrak{C}}(A)$. We have $a_1 \leftrightarrow \mathcal{N}_{\mathfrak{C}}(a_1)$ in $g(A')$ since they are edges connecting A' and $\mathcal{N}_{\mathfrak{C}}(A')$. We also have edges $a_2 \leftrightarrow \mathcal{N}_{\mathfrak{C}}(a_2)$ in $g(A')$ since $a_2 \in \overline{A'}$, $\mathcal{N}_{\mathfrak{C}}(a_2) \subset \overline{\mathcal{N}_{\mathfrak{C}}(A')}$ since $\mathcal{N}_{\mathfrak{C}}(a_1) \cap \mathcal{N}_{\mathfrak{C}}(a_2) = \emptyset$, and the same argument can be repeated for edges $a_3 \leftrightarrow \mathcal{N}_{\mathfrak{C}}(a_3)$, etc. Therefore, all edges connecting $A \leftrightarrow \mathcal{N}_{\mathfrak{C}}(A)$ are in $g(A')$. Then, consider edges connecting $\overline{A}$ and $\overline{\mathcal{N}_{\mathfrak{C}}(A)}$. Since $\overline{A} \subset \overline{A'}$ and $\mathcal{N}_{\mathfrak{C}}(a_i) \cap \mathcal{N}_{\mathfrak{C}}(a_j) = \emptyset$ for $i \neq j$ which means $\overline{\mathcal{N}_{\mathfrak{C}}(A)} \subset \overline{\mathcal{N}_{\mathfrak{C}}(A')}$, the edges connecting $\overline{A} \leftrightarrow \overline{\mathcal{N}_{\mathfrak{C}}(A)}$ are also edges

connecting $\overline{A'} \leftrightarrow \overline{\mathcal{N}_{\mathfrak{C}}(A')}$ and are thus in $g(A')$. Therefore, all edges in $g(A)$ are in $g(A')$, and $g(A)$ is a subgraph of $g(A')$.

Proof of IV:

By steps (1)-(3), we know the redundant constraints are induced by sets $A \subset \mathcal{A}_z$ such that $|A| > 1$ and A only contain events with the same x-level. Therefore, there are a total of $K * \sum_{i=2}^{M-1} \binom{M}{i} = K(2^M - M - 2)$ distinct A and, thus, $K(2^M - M - 2)$ of constraints being redundant.

□

3.3.2 Proof of Proposition 2

Proof. As stated in Section 2.3, the number of inequalities in the form of (2.10) is $Q((2^M - 1)^K - 1)$: here $2^M - 1$ counts the non-empty subsets of $\{1, \ldots, M\}$; the second '-1' arises from the requirement that at least one $\mathcal{V}^{(k)}$ be a strict subset (otherwise the inequality becomes trivial, since both sides are 1).

This can also be obtained through a counting procedure of the unique upper bounds in inequalities (2.10) since there is a one-to-one correspondence between the left-hand side and right-hand side of (2.10). It is equivalent to counting the distinct $M \times K$ matrices comprised of binary entries that obey two constraints:

1. in each column, there is at least one entry with 1,

2. we can't have a matrix with all 1 since that results in a trivial constraint.

Note that such matrices also aid the computational construction of our bound.

Thus, the total number of non-trivial inequalities in the form of (2.10) is $Q((2^M - 1)^K - 1)$ combining all arms by Lemma 4, and the total number of non-redundant inequalities is $Q((2^M - 1)^K - K(2^M - M - 2) - 1)$.

□

3.4 Comparison to results in Russell

Russell [2021] presented sharp bounds on any continuous functionals of the joint counterfactual distribution under the joint independence assumption (V2) that $Z \perp Y(x_1), \ldots, Y(x_K)$. His work is an extension of Beresteanu et al. [2012] in which Russels used results in Luo and Wang [2017] to further eliminate the redundant inequalities implied by Artstein's theorem and obtain constraints in the exact core determining class defining the joint counterfactual distribution. The inequalities in the exact core determining class bounds the counterfactual probabilities in the form of

$$\begin{cases} P(Y(x_{s_1}) = y_1, Y(x_{s_2}) = y_2, \ldots, Y(x_{s_K}) \in \mathcal{M}),\ \forall \mathcal{M} \subseteq \{1, \ldots, M\}, & \text{when } K > 2 \text{ or } M \leq K \\ P(Y(x_i) = y_i, Y(x_j) \in \mathcal{M}',\ \forall \mathcal{M}' \subset \{1, \ldots, M\}, & \text{when } K = 2 \text{ and } K < M \end{cases} \tag{3.5}$$

where $(s_1, \ldots, s_K)$ is any permutation of $(1, \ldots, K)$. In summary, they bound the counterfactual probabilities of events with $K-1$ counterfactual outcomes in common. Note that when $K = 2$ and $M = 2$, the inequalities in (3.5) are the same as our results in (2.10) and results in Richardson and Robins [2014]. However, they deviate from our results when $K \neq 2$ or $M \neq 3$, and the bounds in (3.5) are a strict subset of the non-redundant constraints given by (2.10) and Theorem 4. Below, we use $K = 2, M = 3$, and $Q = 2$ as an illustration.

In our results, the non-redundant inequalities defining the joint counterfactual probability distributions include:

(Set 1) $P(Y(x_i) = y_1^{x=i}) + P(Y(x_i) = y_2^{x=i}) \leq 1 - P(X = i, Y = y_3^{x=i} \mid Z = z)$, $y_1^{x=1} \neq y_2^{x=2} \neq y_3^{x=3}$;

(Set 2) $P(Y(x_1) = y_1^{x=1}, Y(x_2) = y_1^{x=2}) + P(Y(x_1) = y_1^{x=1}, Y(x_2) = y_2^{x=2}) + P(Y(x_1) = y_2^{x=1}, Y(x_2) = y_1^{x=2}) + P(Y(x_1) = y_2^{x=1}, Y(x_2) = y_2^{x=2}) \leq 1 - P(X = 1, Y = y_3^{x=1} \mid Z = z) - P(X = 2, Y = y_3^{x=2} \mid Z = z)$, $y_1^{x=1} \neq y_2^{x=1} \neq y_3^{x=1}$ and $y_1^{x=2} \neq y_2^{x=2} \neq y_3^{x=2}$;

(Set 3) $P(Y(x_i) = y^i, Y(x_j) = y^j) + P(Y(x_i) = y^i, Y(x_j) = \tilde{y}^{x=j}) \leq P(X = i, Y = y^i \mid Z = z) + P(X = j, Y = y^j \mid Z = z) + P(X = j, Y = \tilde{y}^{x=j} \mid Z = z), \quad y^j \neq \tilde{y}^{x=j};$

(Set 4) $P(Y(x_1) = y^1, Y(x_2) = y^2) \leq P(X = 1, Y = y^1 \mid Z = z) + P(X = 2, Y = y^2 \mid Z = z).$

However, only Set 3 and 4 are in the exact core determining class given by Russell [2021] since the left-hand side of the inequalities in Set 3 and 4 are probabilities of counterfactual events with exactly $K - 1 = 1$ outcome in common while Set 1 and 2 violate this condition. In total, there are $42 * 2 = 84$ non-redundant inequalities by our results, and $27 * 2 = 54$ inequalities in the exact core determining class by Russell [2021]. Similarly, when $K = 2, M = 4$, and $Q = 2$, inequalities in the exact core determining class further include inequalities of the form

$$\begin{aligned} &P(Y(x_i) = y^i, Y(x_j) = y_1^{x=j}) + P(Y(x_i) = y^i, Y(x_j) = y_2^{x=j}) + P(Y(x_i) = y^i, Y(x_j) = y_3^{x=j}) \\ \leq &P(X = j, Y = y_1^{x=j} \mid Z = z) + P(X = j, Y = y_2^{x=j} \mid Z = z) + P(X = j, Y = y_3^{x=j} \mid Z = z) \\ &+ P(X = i, Y = y^i \mid Z = z), \qquad y_1^{x=j} \neq y_2^{x=j} \neq y_3^{x=j}, \end{aligned}$$

which are a total of $96 * 2 = 192$ inequalities. These numbers match with results presented in Table 1 in Russell [2021].

Now, we argue that the exact core determining class (1) does not correctly define the joint counterfactual distribution such that it may fail to detect observed probabilities that violate the IV model and (2) can fail to provide a sharp bound on the functionals of the joint counterfactual distribution. We provide examples when $K = 2, M = 3$, and $Q = 2$.

To illustrate the first point, we consider the observed probabilities below.

Using constraints in the exact core determining class, we are able to obtain a convex polytope of the joint counterfactual probability distribution given observed probabilities in

	$P(X=1,Y=1 \mid Z)$	$P(X=1,Y=2 \mid Z)$	$P(X=1,Y=3 \mid Z)$	$P(X=2,Y=1 \mid Z)$	$P(X=2,Y=2 \mid Z)$	$P(X=2,Y=3 \mid Z)$
$Z=1$	0.43	0.05	0.07	0.10	0.20	0.15
$Z=2$	0.01	0.36	0.40	0.18	0.03	0.02

Table 3.1: A synthetic data with $K = 2$, $M = 3$, and $Q = 2$ which violates the IV model

Table 3.1. However, Set 1 includes inequalities

$$P(Y(x_1) = 1) + P(Y(x_1) = 2) \leq 0.60 \tag{3.6}$$

$$P(Y(x_1) = 1) + P(Y(x_1) = 3) \leq 0.64 \tag{3.7}$$

$$P(Y(x_1) = 2) + P(Y(x_1) = 3) \leq 0.57 \tag{3.8}$$

Since $P(Y(x_1) = 1) + P(Y(x_1) = 2) + P(Y(x_1) = 3) = 1$, we can see that (3.6) implies $P(Y(x_1) = 3) \geq 0.40$ and (3.7) implies $P(Y(x_1) = 2) \geq 0.36$, which contradicts (3.8). Thus, the observed probabilities in Table 3.1 violate the IV model while the exact determining core fails to reject the IV model.

To illustrate the second point, we consider the observed probabilities below which obey the IV model.

We compare bounds on various functionals of the joint counterfactual probability distribution using our results vs. the exact core determining class below.

We can see that using our results, we can obtain sharper bounds on all functionals of the joint counterfactual probability distribution presented in Table 3.3 above given observed probabilities in 3.2. Specifically, we bound the pairwise ATE $P(Y(x_1) = 1) - P(Y(x_1) = 3)$ away from 0 while the bound using the exact core determining class includes 0.

	$P(X=1,Y=1 \mid Z)$	$P(X=1,Y=2 \mid Z)$	$P(X=1,Y=3 \mid Z)$	$P(X=2,Y=1 \mid Z)$	$P(X=2,Y=2 \mid Z)$	$P(X=2,Y=3 \mid Z)$
$Z=1$	0.12	0.21	0.30	0.15	0.08	0.14
$Z=2$	0.08	0.44	0.14	0.25	0.03	0.06

Table 3.2: A synthetic data with $K = 2$, $M = 3$, and $Q = 2$

	$P(Y(x_1)=2, Y(x_2)=1)$	$P(Y(x_2)=1)$	$P(Y(x_1)=1)+P(Y(x_1)=2)$	$P(Y(x_1)=1)-P(Y(x_1)=3)$
(2.10) + Theorem 4	[0.01, 0.36]	[0.26, 0.78]	[0.56, 0.70]	[-0.32, -0.04]
Exact Core Determining Class	[0, 0.36]	[0.17, 0.805]	[0.45, 1.00]	[-0.55, 0.44]

Table 3.3: Bounds on the functionals of the joint counterfactual probability distribution using (2.10) and Theorem 4 vs. the exact core determining class

Chapter 4

STATISTICAL INFERENCE ON PARTIAL IDENTIFICATION BOUNDS

4.1 Introduction

We discuss how statistical inference and a confidence region of the partial identification bound on any convex functional of the joint counterfactual distribution can be achieved using a convex optimization algorithm. We illustrate our methods in Sections 2.3 and 4.2 with real data from the Minneapolis Domestic Violence Experiment, where the instrument and treatment both take three levels. In particular, we show that restricting the analysis to only a subset of exposure, outcome, or instrument levels can lead to inefficient or invalid results.

4.2 Algorithm

Given an observed distribution, the inequalities defining the counterfactual probability distributions in Theorem 2 can be used to obtain partial identification bounds on any convex functions of the joint counterfactual probabilities, such as the ATE, with the help of existing linear programming software. In this section, we show how a confidence region could also be constructed using the existing method Guo and Richardson [2021] recently proposed. which are finite-sample tail bounds of the likelihood ratio test (LRT) under multinomial samplings. It essentially further constrains the probability distribution of $P(X, Y \mid Z)$ given the observed probabilities from data via Kullback-Leibler (KL) divergence $KL(P(X, Y \mid Z)||\hat{P}(X, Y \mid Z))$ where $\hat{P}(X, Y \mid Z)$ is the empirical distribution of $P(X, Y \mid Z)$.

Firstly, we provide an algorithm proposed by Guo and Richardson [2021] to obtain a confidence region of any convex functions of the joint counterfactual probabilities including

the ATE.

Algorithm 1 Convex Program Formulation

Require: Functions f; Matrices $H_{obz.z}$, H_j; Vector c; Constant t_α; Empirical probabilities $\hat{p}_{obs.z} \equiv \hat{P}(X, Y \mid Z = z) \quad \forall z = 1, \ldots, Q$;

1: **Variables:**

$$p_{obs.z} \equiv P(X, Y \mid Z = z) \in [0,1]^{K \times M}, \forall z = 1, \ldots, Q$$
$$p_j \equiv P(Y(x_1), \ldots, Y(X_K)) \in [0,1]^{M^K}$$

2: Solve the following convex program:

$$\begin{aligned} \min_{p_j} \quad & f(p_j); \quad \max_{p_j} \quad f(p_j) \\ \text{s.t.} \quad & H_{obs.z}^\top p_{obs.z} + H_j^\top p_j \leq c, \forall z = 1, \ldots, Q, \\ & KL(p_{obs.Z} || \hat{p}_{obs.Z}) \leq t_\alpha, \\ & \sum p_j = 1; \quad \sum p_{obs.z} = 1, \forall z = 1, \ldots, Q, \\ & p_j \geq 0; \quad p_{obs.z} \geq 0, \forall z = 1, \ldots, Q. \end{aligned}$$

3: **return** $f(p_j^*)$

Then, we give a description of the algorithm in words below. In order to define the convex polytope in which the joint counterfactual probability distribution lives, we can either use the intersection of a finite collection of closed half-spaces (H-representation matrix) or the convex hull of a finite collection of points and directions (V-representation matrix). Note that the H-representation matrix would be equivalent to the set of inequalities given by Theorem 2 and 4.

1. Obtain the empirical probabilities $\hat{P}(X, Y \mid Z)$ from the data.

2. Construct a V-representation matrix and use 'RCDD' package to obtain the corresponding H-representation matrix. This is equivalent to obtaining the H-representation matrix using Theorem 2 and 4.

   - The first M^K columns of the V-representation matrix are for the joint counterfactual probabilities $P\left(Y(x_1) = y^1, \ldots, Y(x_K) = y^K\right)$ followed by the MK columns

for conditional observed probabilities $P(X = x, Y = y \mid Z = z)$.

- Each row represents a possible extreme distribution consisting of one joint counterfactual probability distribution $P(Y(x_1), \ldots, Y(x_K))$ and one of its coherent conditional observed probabilities $P(X, Y \mid Z = z)$. 'Extreme' means we only allow one possible joint counterfactual probability and one of its coherent observed probabilities in the distribution with their probabilities to be one and the others being zero. For example, we have that $P(Y(x_1) = 1, \ldots, Y(x_K) = 1) = 1$ and $P(X = 1, Y = 1 \mid Z = z) = 1$ for all $Z = z$ is one of the possible extreme distributions and $P(Y(x_1) = 1, \ldots, Y(x_K) = 1) = 1$ and $P(X = K, Y = 1 \mid Z = z) = 1$ is another one. We complete the matrix with all possible extreme distributions.
- We finish the V-representation matrix by adding a column of zero and a column of one as the first two columns of the matrix. The resulting V-representation matrix defines a convex hull of a polytope using a finite collection of points and directions. The convex polytope described is the set of convex combinations of these points, and the coefficients are (1) nonnegative if the first column of the V-representation matrix is zero and (2) sum to one over each row if the second column is one.

3. Then, we have the following constraints which defines a convex program:

   (a) Using the H-representation matrix and the observed probabilities $P(X, Y \mid Z = z)$, we can get a set of bounds on the joint counterfactual probabilities $P\left(Y(x_1) = y^1, \ldots, Y(x_K) = y^K\right)$ for each instrument arm $Z = z$.

   (b) We also consider the bound $KL(P(X, Y \mid Z) || \hat{P}(X, Y \mid Z)) \leq t_\alpha$ where t_α is a finite-sample critical value which can be obtained using the 'Chernoff' package in R.

   (c) $\sum_{y^1=1}^{M} \cdots \sum_{y^K=1}^{M} P\left(Y(x_1) = y^1, \ldots, Y(x_K) = y^K\right) = 1$ and $\sum_{i=i}^{K} \sum_{j=1}^{M} P(X = i, Y = j) = 1$.

(d) $P\left(Y(x_1) = y^1, \ldots, Y(x_K) = y^K\right) \geq 0$ and $P(X = i, Y = j) \geq 0$ for all $i \in \{1, \ldots, M\}$ and $j, y^1, \ldots, y^K \in \{1, \ldots, M\}$.

4. Using the convex programming package 'CVXR' in R, we can obtain a confidence region for any convex function of the joint counterfactual probability distribution, e.g. a pair-wise average treatment effect $P(Y(x_i) = y) - P(Y(x_j) = y)$ with $i \neq j$, $i, j \in \{1, \ldots, K\}$, and $y \in \{1, \ldots, M\}$.

Note that without inequalities 3(b), the procedure above is how partial identification bounds on parameters of interest could be obtained by plugging in the empirical probabilities $\hat{P}(X, Y \mid Z))$ for $P(X, Y \mid Z))$.

In the next Section, we apply the above procedure to the real data example.

4.3 Real Data Analysis

We now revisit the Minneapolis Domestic Violence Experiment discussed in Section 1.1. We discuss and compare the partial identification bounds and confidence regions obtained using all data or partial data which might be inefficient and/or biased.

All researchers below are interested in the following pairwise average treatment effects: (1) Advice ($X = Adv$) vs. Arrest ($X = Arr$); (2) Separate ($X = Sep$) vs. Arrest ($X = Arr$); (3) Separate ($X = Sep$) vs. Advice ($X = Adv$).

Researcher 1: They are always hesitant to ignore any data available. They read our paper and used all the data for all three pairwise ATEs.

Researcher 2: They know all three Z arms are independent because of randomization. Therefore, they decided to ignore a random Z arm. Researcher 3: They were not aware of our paper but knew the partial identification bound for binary exposure, X. They decided to remove participants who took the treatment that was not of interest for a given pairwise ATE. For example, when estimating the ATE comparing the Arrest response vs. the Advice response, they ignored treatment $X = Sep$ which are the police officers who decided to separate the suspect from the scene.

Researcher 4: They thought if they were only interested in the pairwise ATEs, both the Z and the X levels related to the third treatment could be ignored. They believe the problem is then reduced to IV model with binary Z, X, and Y, which is well-studied and computationally straightforward. Therefore, when estimating the ATE comparing the Arrest response vs. the Advice response, they ignored data with $Z = Sep$ and $X = Sep$, which are the police officers who were either assigned to separate the suspect from the scene regardless of their final decision or police officers who decided to separate the suspect from the scene regardless of their original random assignment received. All results are summarized in Table 4.1.

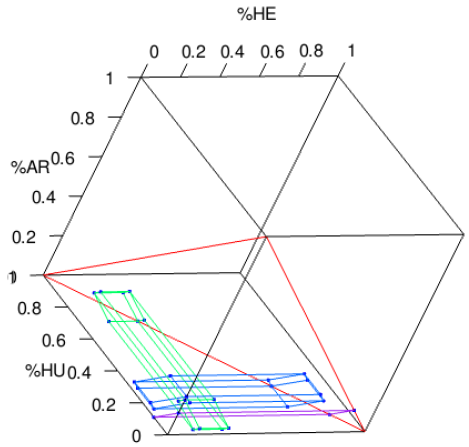
Comparing the results between Researcher 2 and Researcher 1, we can see that an additional Z arm could be helpful. Researcher 1 obtained no wider plug-in bounds with the help of the third Z arm. Furthermore, the confidence regions from Researcher 1 control the 5% family-wise error rate (FWER) across all three pair-wise ATEs, while the confidence regions from Researcher 2 only control the 5% FWER for a given pair-wise ATE. This is also the reason why some of the confidence regions from Research 2 are narrower than those from Research 1.

Deleting X groups, like what Researcher 3 did, gives the wrong estimand and could be dangerous. They ignored the population taking the treatment that they were not interested in. They may end up with data that violates the instrumental variable model, as shown in the table for Researcher 3. We obtained p-values assessing the null hypothesis of not rejecting the IV model, which are 0.55, 0.625, and 0.586 for data with treatment groups Advice and Arrest, Separate and Arrest, and Separate and Advice respectively. We also plotted the polytopes for the joint counterfactual probability distribution; see Figure 4.1. We can see that two out of three polytopes have empty intersections, which explains why the data is outside of the IV model. However, as the distances between the polytopes that don't intersect are small, we have non-significant p-values.

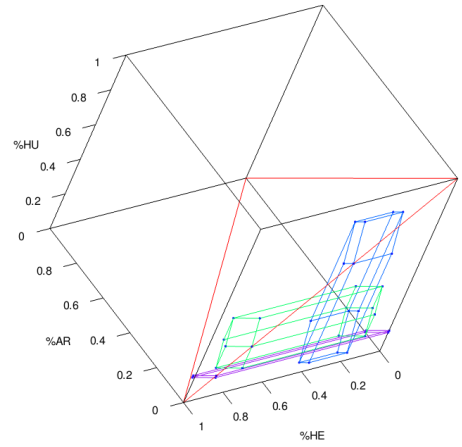
Finally, Researcher 4 restricted the problem to a binary IV model, which is not valid for similar reasons to Researcher 3.

Table 4.1: Results for the Minneapolis Domestic Violence Experiment obtained by different researchers

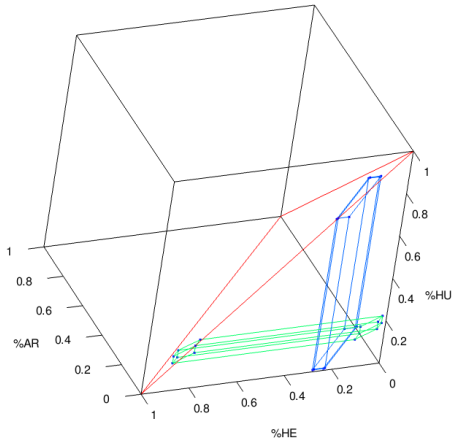
		Advice vs. Arrest		Separate vs. Arrest		Separate vs. Advice	
		Plug-in	**Confidence Region (95%)**	**Plug-in**	**Confidence Region (95%)**	**Plug-in**	**Confidence Region (95%)**
Researcher 1	**All data**	(0.019, 0.252)	(-0.374, 0.633)	(0.057, 0.343)	(-0.346, 0.702)	(-0.184, 0.312)	(-0.583, 0.683)
Researcher 2	**Delete Z=Arrest**	(-0.675, 0.317)	(-0.885, 0.621)	(-0.637, 0.407)	(-0.858, 0.691)	(-0.184, 0.312)	(-0.533, 0.639)
	Delete Z=Advice	(-0.111, 0.856)	(-0.407, 0.981)	(0.057, 0.343)	(-0.285, 0.660)	(-0.788, 0.442)	(-0.953, 0.721)
	Delete Z=Separate	(0.019, 0.252)	(-0.314, 0.586)	(-0.092, 0.864)	(-0.394, 0.985)	(-0.403, 0.866)	(-0.628, 0.973)
Researcher 3	**Delete X=Separate, X=Advice, or X=Arrest**	NA	(-0.283, 0.526)	NA	(-0.264, 0.625)	NA	(-0.355, 0.457)
Researcher 4	**Binary IV model: Delete** $X = Sep$ **and** $Z = Sep$, $X = Adv$ **and** $Z = Adv$, **or** $X = Sep$ **and** $Z = Sep$	(0.037, 0.214)	(-0.241, 0.506)	(0.066, 0.317)	(-0.222, 0.598)	(-0.003, 0.121)	(-0.337, 0.446)



(a) Polytopes defining the joint counterfactual probability distribution given data with treatment groups Arrest and Advice.



(b) Polytopes defining the joint counterfactual probability distribution given data with treatment groups Arrest and Separate.



(c) Polytopes defining the joint counterfactual probability distribution given data with treatment groups Separate and Advice.

Figure 4.1: Visualization of polytopes defining the joint counterfactual probability distribution with binary treatment and outcome, and instrument taking three levels. The purple polytope represents the Arrest arm; the green polytope represents the Advice arm; the blue polytope represents the Separate arm.

4.4 Example: algorithm on binary IV

When we have a binary X, Y and Z, we construct a V-representation matrix in the following way:

1. The columns represent the joint counterfactual probabilities of Y and the observed probabilities of X, Y given $Z = z$. Under the binary IV model, we have the columns being $P(Y(x_1) = 1, Y(x_2) = 1), P(Y(x_1) = 1, Y(x_2) = 2), P(Y(x_1) = 2, Y(x_2) = 1), P(Y(x_1) = 2, Y(x_2) = 2), P(X = 1, Y = 1 \mid Z = z), P(X = 1, Y = 2 \mid Z = z), P(X = 2, Y = 1 \mid Z = z), P(X = 2, Y = 2 \mid Z = z)$.

2. We consider the extreme distributions consisting of one type defined by the joint counterfactual probability distributions $P(Y(x_1), Y(x_2))$ and one of its coherent observed probabilities $P(X, Y \mid Z = z)$. For example, we have $P(Y(x_1) = 1, Y(x_2) = 1) = 1$ and $P(X = 1, Y = 1 \mid Z = z) = 1$. We complete the matrix with all possible extreme distributions.

3. We add a column of zero and a column of one as the first two columns of the matrix

The V-representation matrix for the binary IV model is below in table 4.2.

Table 4.2: V-representation matrix for binary IV model.

0	1	1	0	0	0	1	0	0	0
0	1	0	1	0	0	0	1	0	0
0	1	0	0	1	0	1	0	0	0
0	1	0	0	0	1	0	1	0	0
0	1	1	0	0	0	0	0	1	0
0	1	0	1	0	0	0	0	1	0
0	1	0	0	1	0	0	0	0	1
0	1	0	0	0	1	0	0	0	1

Using the 'RCDD' package in R, we are able to obtain the so-called H-representation given the V-representation above. The inequalities given in the H-representation matrix relate the

joint counterfactual probability distributions to the observed probabilities conditional on $Z = z$, and the H-representation matrix is the same for all $Z = z$. They also characterize half-spaces in the joint counterfactual probability distribution, and the intersection of the half-spaces defines a polytope where the joint counterfactual probability distributions live in given the observed probabilities in arm $Z = z$. Note that the inequalities outputted by 'RCDD' are exactly in the form of (2.10) for each instrument arm $Z = z$, and there are $(2^M - 1)^K - 1 - K(2^M - M - 2) = 8$ inequalities since we have $K = M = 2$.

We use the convex programming package 'CVXR' in R to obtain the confidence region for the average treatment effect $P(Y(x_2) = 2) - P(Y(x_1) = 2)$. There are a total of 4 sets of constraints:

1. the inequalities given by the H-representation matrix under arm $Z = 1$ and $Z = 2$;

2. $KL(P(X, Y \mid Z)||\hat{P}(X, Y \mid Z)) \leq t_\alpha$ where t_α is a finite-sample critical value which could be obtained using the 'Chernoff' package in R;

3. $\sum_{i=1}^{2}\sum_{j=1}^{2} P(Y(x_1) = i, Y(x_2) = j) = 1$ and $\sum_{i=1}^{2}\sum_{j=1}^{2} P(X = i, Y = j) = 1$;

4. $P(Y(x_1) = i, Y(x_2) = j) \geq 0$ and $P(X = i, Y = j) \geq 0$ for all $i, j \in \{1, 2\}$.

Chapter 5

FALSIFICATION TESTS OF THE IV MODEL

5.1 Introduction

Pearl [1995] gives inequalities that are necessary and sufficient for $P(X, Y \mid Z)$ to be compatible with the binary IV model. There are other works that explored falsification tests for the assumptions especially the independence condition beyond the binary IV model. Bonet [2001] used an optimization program to check that the IV inequalities proposed by Pearl [1995] are necessary but not sufficient in general. They are shown to be sufficient when the instrument, exposure, and outcome are all binary but not when the instrument takes 3 levels while the exposure and outcome are binary. Kédagni and Mourifié [2020] proposed a generalized set of IV inequalities for the joint independence assumption and showed they are necessary and sufficient to detect all violations of the joint independence assumption and exclusion restriction condition with binary outcome. Moreover, Richardson and Robins [2014] gives a necessary and sufficient characterization of the joint counterfactual distribution when the instrument is discrete and takes finite states but X and Y are binary. This can also be used as a falsification test by checking whether the resulting set is empty for the IV model with binary exposure and treatment and discrete instrument. Similarly, since the partial identification bounds we derived in Chapter 2 are also necessary and sufficient, these bounds can also serve as a falsification test for the IV model with discrete instrument, exposure, and outcome.

5.2 Simulation Results

5.2.1 Proportion of Multinomial Distributions that are not in the categorical IV Model

Since the bounds in the form of (2.10) are sharp to characterize the categorical IV model, violations of the bounds (i.e. the lower bound being greater than the upper bound) are equivalent to falsification of the IV model. We simulated the observed probability distribution with binary Y in each instrument arm $Z = z$ from a uniform Dirichlet distribution, $Dirich(\underbrace{1, \ldots, 1}_{2K})$, with a sample size of 500. The proportion of the observed distribution not compatible with the categorical IV model when Y is binary is presented below. The number of simulations is 200.

5.2.2 Reasoning Behind the Scene: Helly's Theorem

One question people might ask is if I have an instrumental variable Z that takes a lot of states, i.e. Q is large, do I need to check if there is any non-empty intersection of the polytope that the counterfactual distribution lives in for all possible combinations of Z to make sure my observed probability distribution is compatible with the IV model? The answer is no by Helly's Theorem. The statement of Helly's Theorem is below.

Helly's Theorem: Let $C_1, ..., C_n$ be a finite collection of convex subsets of $\mathbf{R}^d$, with $n \geq d + 1$. If the intersection of every $d + 1$ of these sets is nonempty, then the whole collection has a nonempty intersection.

In our problem, we have $d = M^K - 1$, and this implies that when we have $Q > d+1$, the forms of constraints on the marginal counterfactual distributions $P(Y(x_1)), \ldots, P(Y(x_K))$ should be the same as $Q = d + 1$, although the number of constraints should still be larger. This is also why we see a turning point at $Q = M^K$ in Figure 5.1 above. Also, suppose there is a non-empty intersection of the polytope the counterfactual distribution lives in for any $d + 1$ of Z arms. In that case, the observed probability distribution is compatible with the

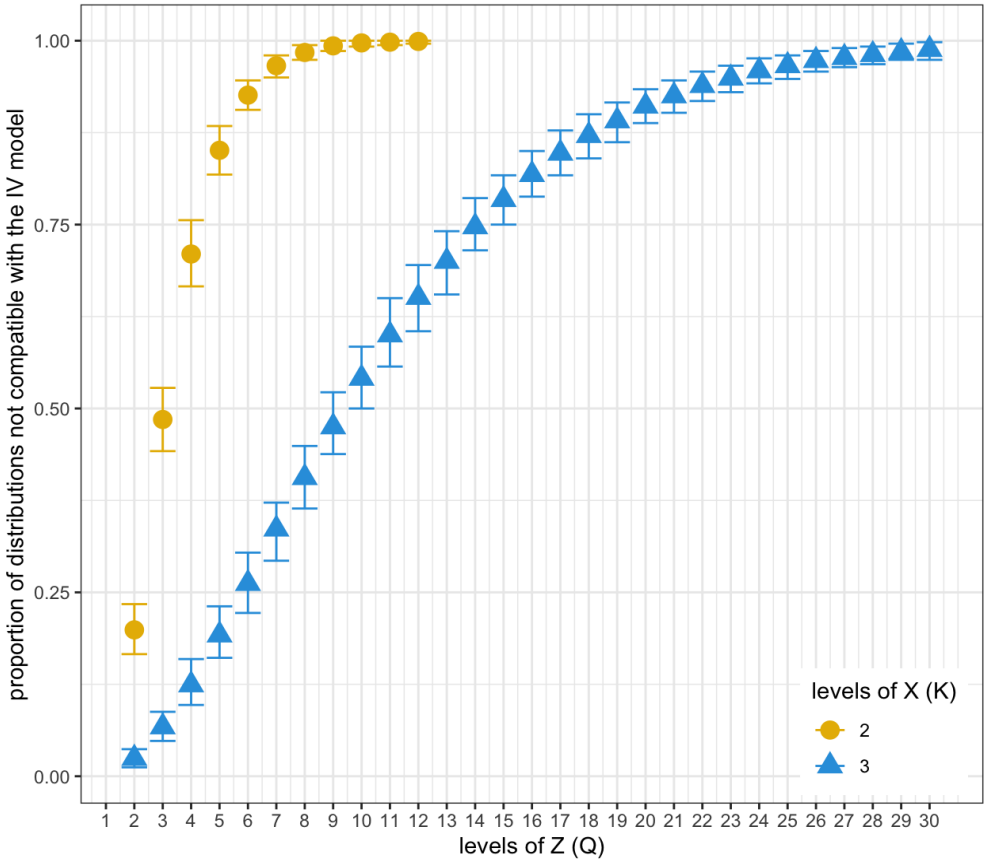


Figure 5.1: The proportion of empirical distributions that are not compatible with the IV model. The observed distribution is simulated under uniform Dirichlet distribution $Dirich(1, \ldots, 1)$. Y is binary.

IV model regardless of how large Q is. When we have $K = M = 2$, it is sufficient to check any 4 of Z arms, and when we have $K = 3$ and $M = 2$, we need to check any 8 of Z arms.

Chapter 6

VARIATION (IN)DEPENDENCE OF THE COUNTERFACTUAL MARGINAL DISTRIBUTION

6.1 Introduction

Richardson and Robins [2024] showed that when X, Y are binary, $P(Y(x_1))$ and $P(Y(x_2))$ are variation independent as shown in Figure 6.1. This makes it easy to compute the bounds on $ATE = P(Y(x_2) = 2) - P(Y(x_1) = 2)$: the lower bound on ATE is the lower bound of $P(Y(x_2) = 2)$ subtracting the upper bound of $P(Y(x_1) = 2)$ while the upper bound on ATE is the upper bound of $P(Y(x_2) = 2)$ subtracting the lower bound of $P(Y(x_1) = 2)$. However, we claim that when Y is binary and X takes more than 2 levels, $P(Y(x_1)), \ldots, P(Y(x_K))$ are NOT variation independent of each other, regardless of the number of levels of Z. This also means in the bounds relating marginals of counterfactual probabilities $\{P(Y(x_i)) : i = 1, \ldots, K\}$ and observed conditional probabilities $\{P(X = x, Y = y|Z = z) : x = 1, \ldots, K, y = 1, 2, z = 1, \ldots Q\}$, there are bounds involving more than one marginal counterfactual probability at a time. In this case, knowing one $P(Y(x_i))$ will give us additional information on the others. Hence, the bound on the ATE obtained by calculating the difference of the upper and lower bounds of the marginal probabilities will not be sharp anymore.

This further emphasizes the importance of our Theorem 2 being formulated on the joint counterfactual probability distribution instead of on the marginal counterfactual probability distribution – obtaining sharp bounds on the marginal counterfactual probabilities is not always sufficient for sharp bounds on the ATE. Hence, in practice, direct calculation of the bounds on all pairwise ATEs is needed. Bounds on $P(Y(x_2) = 2) - P(Y(x_1) = 2)$ and $P(Y(x_3) = 2) - P(Y(x_1) = 2)$ may not imply bounds on $P(Y(x_3) = 2) - P(Y(x_2) = 2)$).

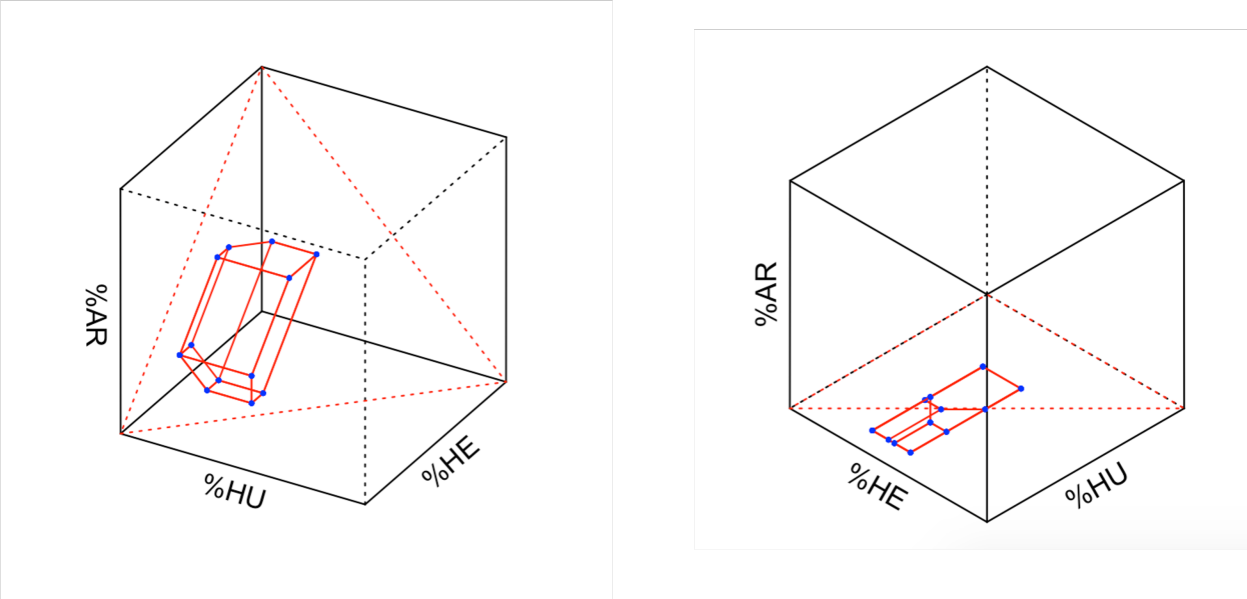


Figure 6.1: Variation independence of $P(Y(x_1))$ and $P(Y(x_2))$ when $K = M = 2$

6.2 Main Results and Proofs

We provide a theorem on the variational independence property on the marginal counterfactual probability distribution $P(Y(x_1)), \ldots, P(Y(x_K))$ in the categorical IV model when the outcome is binary.

Theorem 5. *When $M = 2$, the marginal counterfactual probabilities $P(Y(x_1)), \ldots, P(Y(x_K))$ are variational independent if and only if $K = 2$, regardless of the levels of Z.*

Proof. The "if" part has been shown in Richardson and Robins [2024] that $P(Y(x_1))$ and $P(Y(x_2))$ are variational independent when $K = M = 2$ through Fourier-Motzkin elimination. The idea of the proof is the following: they first rewrote the inequalities in terms of the joint counterfactual probabilities $P(Y(x_1), Y(x_2))$ to the marginal probabilities $P(Y(x_1))$ and $P(Y(x_2))$. Then, they showed all inequalities involving both $P(Y(x_1))$ and $P(Y(x_2))$ can be implied by inequalities only involving a single marginal counterfactual probability. Thus, this showed that $P(Y(x_1))$ and $P(Y(x_2))$ are variational independent.

The "only if" part can be shown similarly by providing a necessary and non-redundant inequality that involves more than one marginal counterfactual probability at the same time. We only need to show the base case with $K = 3, M = 2, Q = 2$. For other cases with $K > 3$, or $Q > 2$ with $M = 2$, the base case can be seen as special cases with probabilities of X or Z taking the additional level being zero, and thus can be easily proved by induction. We now provide inequalities involving more than two marginal counterfactual probabilities under the base case and show they are necessary. They are further non-redundant as shown through 'RCDD' calculation and synthetic examples.

Claim: when $K = 3, M = 2, Q = 2$, $P(Y(x_1)), P(Y(x_2))$ and $P(Y(x_3))$ are not variation independent.

Given Theorem 2 and 4, we know the inequalities of the form below are necessary and non-redundant when $K = 3, M = 2$, and $Q = 2$ where $i, i' \in \{1, 2, 3\}$, $j, j', \tilde{j} \in \{1, 2\}$, and $i \neq i'$.

$$P(Y(x_1) = j, Y(x_2) = j', Y(x_3) = \tilde{j}) \leq 1 - p_{1(3-j).z} - p_{2(3-j').z} - p_{3(3-\tilde{j}).z} \tag{6.1}$$

Then, we have

$$\begin{aligned}
&P(Y(x_{i'}) = j') - P(Y(x_i) = 3 - j) \\
&\quad \leq P(Y(x_{i'}) = j') - P(Y(x_i) = 3 - j, Y(x_{i'}) = j') \\
&\quad = P(Y(x_i) = j, Y(x_{i'}) = j') \\
&\quad = P(Y(x_i) = j, Y(x_{i'}) = j', Y(x_{\tilde{i}}) = 1) + P(Y(x_i) = j, Y(x_{i'}) = j', Y(x_{\tilde{i}}) = 2) \\
&\quad \leq 2 - P(X = i, Y = 3 - j \mid Z = z) - P(X = i, Y = 3 - j \mid Z = z') \\
&\quad - P(X = i', Y = 3 - j' \mid Z = z) - P(X = i', Y = 3 - y' \mid Z = z') \\
&\quad - P(X = \widetilde{i}, Y = 3 - \widetilde{j} \mid Z = z) - P(X = \widetilde{i}, Y = \widetilde{i} \mid Z = z')
\end{aligned}$$

The first three steps are by algebra, and the fourth inequality is by inequality 6.1. □

Below, we give the full sets of inequalities characterizing the marginal counterfactual

probability distribution under different combinations of K and Q with $M = 2$.

6.2.1 Base case: Binary Z, X taking three levels, binary Y

We denote $p_{xy.z} = P(X = x, Y = y \mid Z = z)$. From Theorem 2, we can write out all inequalities characterizing the joint distribution $P(Y(x_1), Y(x_2), Y(x_3))$. They can be categorized into the following three sets with $i, i' \in \{1, 2, 3\}$, $j, j', \tilde{j} \in \{1, 2\}$, and $i \neq i'$.

$$P(Y(x_i) = j) \leq 1 - p_{i(3-j).z} \tag{6.2}$$

$$P(Y(x_i) = j, Y(x_{i'}) = j') \leq 1 - p_{i(3-j).z} - p_{i'(3-j').z} \tag{6.3}$$

$$P(Y(x_1) = j, Y(x_2) = j', Y(x_3) = \tilde{j}) \leq 1 - p_{1(3-j).z} - p_{2(3-j').z} - p_{3(3-\tilde{j}).z} \tag{6.4}$$

Notice that bounds in the form of (6.2) is a bound on $P(Y(x_i) = j)$ which is a summation of 4 joint probabilities $P(Y(x_1) = j, Y(x_2) = j', Y(x_3) = \tilde{j})$. For example, $P(Y(x_1) = 1) = P(Y(x_1) = 1, Y(x_2) = 1, Y(x_3) = 1) + P(Y(x_1) = 1, Y(x_2) = 1, Y(x_3) = 2) + P(Y(x_1) = 1, Y(x_2) = 2, Y(x_3) = 1) + P(Y(x_1) = 1, Y(x_2) = 2, Y(x_3) = 2)$.
Bounds in the form of (6.3) is a bound on $P(Y(x_i) = j, Y(x_{i'}) = j')$, which is a summation of 2 joint probabilities $P(Y(x_1) = j, Y(x_2) = j', Y(x_3) = \tilde{j})$. For example, $P(Y(x_1) = 1, Y(x_2) = 1) = P(Y(x_1) = 1, Y(x_2) = 1, Y(x_3) = 1) + P(Y(x_1) = 1, Y(x_2) = 1, Y(x_3) = 2)$
Bounds in the form of (6.4) is a bound directly on the joint probability $P(Y(x_1) = j, Y(x_2) = j', Y(x_3) = \tilde{j})$.

If we want to use bounds in the form of (6.2)-(6.3) to construct bounds on the marginal counterfactual probabilities $P(Y(x_i) = j)$, then we have the following.

1. From set 1, we will have $6 \times 2 = 12$ bounds since $Q = 2$.

2. From set 2, we can add the inequalities pairwise - e.g. $P(Y(x_1) = 1) = P(Y(x_1) = 1, Y(x_2) = 0) + P(Y(x_1) = 1, Y(x_2) = 1) \leq p_{11.z} + p_{20.z} + p_{30.z} + p_{31.z} + p_{11.z'} + p_{21.z} + p_{30.z'} + p_{31.z'}$ or $P(Y(x_1) = 1) = P(Y(x_1) = 1, Y(x_3) = 0) + P(Y(x_1) = 1, Y(x_3) =$

1) $\leq p_{11.z} + p_{20.z} + p_{21.z} + p_{30.z} + p_{11.z'} + p_{20.z'} + p_{21.z'} + p_{31.'}$. Notice that the right-hand side involves both levels of Z. Otherwise, it will be trivial since they can be implied by inequalities in set 1. Then, we can get $12 * 2 = 24$ bounds.

3. From set 3, we can first add the inequalities pairwise to get bounds on $P(Y(x_i) = j, Y(x_{i'}) = j')$. Then, we have $P(Y(x_i) = j, Y(x_{i'}) = j') = P(Y(x_{i'}) = j') - P(Y(x_i) = 3 - j, Y(x_{i'}) = j') \geq P(Y(x_{i'}) = j') - P(Y(x_i) = 3 - j)$. Therefore, we will have $12 * 2 = 24$ bounds. These bounds involve two $P(Y(x_i) = y_i)$ and thus is the reason why we don't have variation independence anymore. Specifically, they are in the form

$$\begin{aligned} P(Y(x_{i'}) = j') - P(Y(x_i) = 3 - j) \leq & 2 - P(X = i, Y = 3 - j \mid Z = z) \\ & - P(X = i', Y = 3 - j' \mid Z = z) - P(X = \widetilde{i}, Y = 3 - \widetilde{j} \mid Z = z) \\ & - P(X = i, Y = 3 - j \mid Z = z') - P(X = i', Y = 3 - y' \mid Z = z') \\ & - P(X = \widetilde{i}, Y = \widetilde{i} \mid Z = z') \end{aligned}$$

Adding them together, we will have 60 bounds, which is the same as what we obtain from the RCDD code.

6.2.2 Z, X taking three levels, binary Y

Using Theorem 2, we will get bounds in the same form as (6.2)-(6.4) with $z = 1, 2, 3$. Similarly with binary Z, from set 1, we have $6 \times 3 = 18$ bounds. From set 2, we have $12 \times \binom{3}{2} \times 2 = 72$ bounds. From set 3, we also have $12 \times \binom{3}{2} \times 2 = 72$ bounds.

Then since now we have 3 levels of Z, we have a little bit more 'freedom'. Firstly, after obtaining bounds on $P(Y(x_i) = j, Y(x_{i'}) = j')$ using set 3, we could combine it with bounds in set 2 to get another set of bounds on $P(Y(x_i) = j)$. In this way, we have $6 \times 2 \times 2 \times \binom{3}{2} \times 2 = 144$ bounds.

Additionally, we can take two inequalities from set 2, $P(Y(x_i) = j, Y(x_{i'} = j'))$ and $P(Y(x_i) = j, Y(x_{\tilde{i}} = \tilde{j}))$ where $i' \neq \tilde{i}$, which agree on the potential outcome $Y(x_i) = j$,

and one corresponding inequality from set 3, $P(Y(x_i) = 3 - j, Y(x_{i'} = j'), Y(x_{\tilde{i}} = \tilde{j}))$. For example, $P(Y(x_2) = 1, Y(x_3) = 1)$ and $P(Y(x_1) = 1, Y(x_2) = 1)$ from set 2 is corresponding with $P(Y(x_1) = 1, Y(x_2) = 2, Y(x_3) = 1)$ from set 3. Note that for each pair of inequalities from set 2 described above, there is only one corresponding inequality from set 3. Adding them together, we can get

$$
\begin{aligned}
&P(Y(x_2) = 0, Y(x_3) = 0) + P(Y(x_1) = 0, Y(x_2) = 0) + P(Y(x_1) = 0, Y(x_2) = 1, Y(x_3) = 0) \\
&= P(Y(x_2) = 0) - P(Y(x_2) = 0, Y(x_3) = 1) + P(Y(x_1) = 0) - P(Y(x_1) = 0, Y(x_2) = 1) \\
&+ P(Y(x_1) = 0, Y(x_2) = 1) - P(Y(x_1) = 0, Y(x_2) = 1, Y(x_3) = 1) \\
&= P(Y(x_2) = 0) - P(Y(x_2) = 0, Y(x_3) = 1) + P(Y(x_1) = 0) \\
&- P(Y(x_1) = 0, Y(x_2) = 1, Y(x_3) = 1) \\
&\geq P(Y(x_2) = 0) - P(Y(x_2) = 0, Y(x_3) = 1) + P(Y(x_1) = 0) - P(Y(x_2) = 1, Y(x_3) = 1) \\
&= P(Y(x_2) = 0) + P(Y(x_1) = 0) - P(Y(x_3) = 1)
\end{aligned}
$$

These bounds will involve all three margins of the counterfactual probabilities. The right-hand sides of the bounds on $P(Y(x_i) = j, Y(x_{i'} = j'))$, $P(Y(x_i) = j, Y(x_{\tilde{i}} = \tilde{j}))$, and $P(Y(x_i) = 3 - j, Y(x_{i'} = j'), Y(x_{\tilde{i}} = \tilde{j}))$ need to involve all three levels of Z. Hence, there are $\binom{3}{2} \times 2 \times 2 \times 2 \times 2 \times 3 \times 2 = 288$ bounds.

Notice that the set of bounds we obtain with binary Z is always a subset of bounds we have when Z takes more levels. Since $P(Y(x_i) = y_i)$ are not variation independent with each other with $i \in \{1, 2, 3\}$, they are variation dependent regardless of the number of levels in Z.

6.2.3 X with more levels

We used computing software to check the variation dependence for the following combinations of Q, M, K which are presented in Table 6.1.

Q	K	M
2, 3, 4, 5	3	2
2	4	2
2	5	2
2	6	2

Table 6.1: Variation dependence property under the following combinations of Q (levels of Z) and K (levels of X), with $M = 2$ (levels of Y)

6.3 Simulation Results

6.3.1 Consequences of Variation Dependence Property

The variation dependence property is a blessing and a curse at the same time. On the one hand, it makes our computation on the bounds of the ATE more complicated in that it is no longer simply a subtraction of the upper and lower bounds of each marginal counterfactual probability $P(Y(x_i))$. On the other hand, we might be able to get tighter bounds because of the bounds that involve multiple marginal counterfactual probabilities.

Take the situation with $K = 3, M = 2, Q = 2$ as an example, in summary, we have bounds in the following sets

1. $P(Y(x_i) = y) \leq 1 - P(X = i, Y = 1 - y \mid Z = z)$

2. $P(Y(x_i) = y) = P(Y(x_i) = y, Y(x_j) = 0) + P(Y(x_i) = y, Y(x_j) = 0) \leq (1 - P(X = i, Y = 1 - y \mid Z = z) - P(X = j, Y = 1 \mid Z = z)) + (1 - P(X = j, Y = 1 - y \mid Z = z') - P(X = j, Y = 0 \mid Z = z'))$ where $i \neq j$

3. $P(Y(x_i) = y) - P(Y(x_j) = y') \leq P(X = k, Y = 0 \mid Z = z) + P(X = i, Y = y \mid Z = z) + P(X = j, Y = 1 - y' \mid Z = z) + P(X = k, Y = 1 \mid Z = z') + P(X = i, Y = y \mid Z = z') + P(X = j, Y = 1 - y' \mid Z = z')$ where $i \neq j \neq k$.

Denote the right-hand side of bounds in set 3 as $\tau_{ij,yy'}$ for the simplicity of notation. Consider the treatment effect $P(Y(x_2) = 2) = P(Y(x_1) = 2)$, we can get upper and lower

bounds on $P(Y(x_1) = 2)$ and $P(Y(x_2) = 2)$ separately using bounds in sets 1 and 2 above. Then, it is not hard to see if there exists a $\tau_{21,22}$ such that $\tau_{21,22} < \max_{z,z'} P(Y(x_2) = 2) - \min_{z,z'} P(Y(x_1) = 2)$, it will give us a tighter bound on $P(Y(x_2) = 2) - P(Y(x_1) = 2)$; see Figure 6.2.

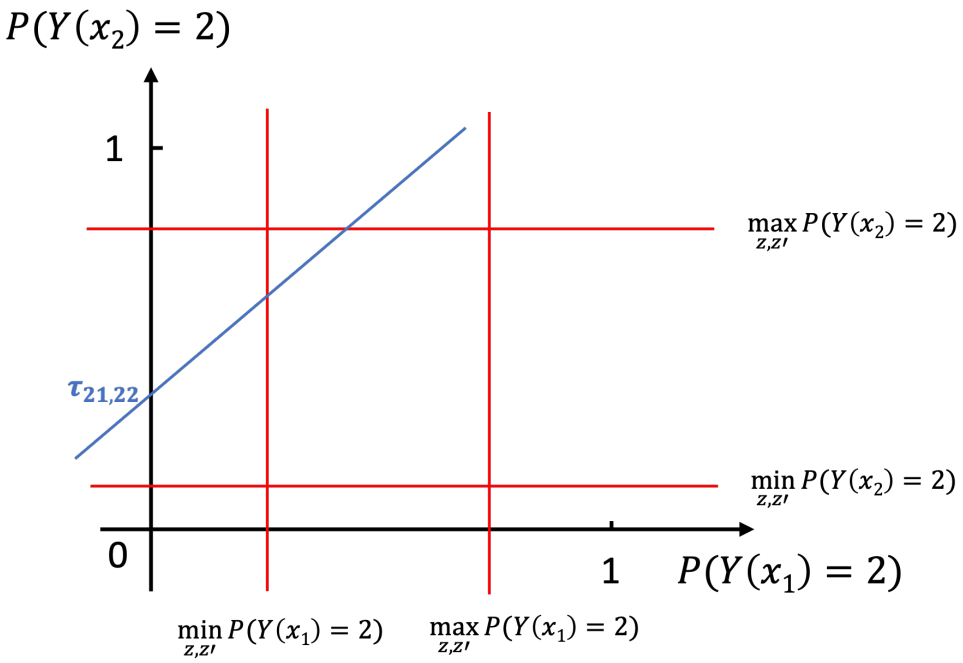


Figure 6.2: Example on how variation dependence property can help us with getting a tighter bound on the average causal effect.

6.3.2 Synthetic example

We simulate the observed probability distribution $P(X, Y \mid Z = z)$ using a uniform Dirichlet distribution, $Dirich(\underbrace{1, \ldots, 1}_{MK})$, for each instrument arm $Z = z$ with $K = Q = 3, M = 2$. Then, using Theorem 2 and a linear programming software, we calculate the partial identification bound on the pair-wise ATE $P(Y(x_2) = 2) - P(Y(x_1) = 2)$ as well as on the marginal counterfactual probabilities $P(Y(x_2) = 2)$ and $P(Y(x_1) = 2)$. We compare the upper and lower bounds on $P(Y(x_2) = 2) - P(Y(x_1) = 2)$ vs. $\max(P(Y(x_2) = 2)) - \min(P(Y(x_1) = 2))$

and $\min(P(Y(x_2)=2)) - \max(P(Y(x_1)=2))$. The number of simulations is $n = 2000$.

Out of the 1877 simulation runs, which are compatible with the categorical IV model, we have 142 (7.57%) of them that the bounds on the pair-wise ATE directly obtained using Theorem 2 is sharper than those obtained using differences of the upper and lower bounds of $P(Y(x_2)=2)$ and $P(Y(x_1)=2)$. We give a specific example below.

Example 3. Suppose we have observed probabilities, $P(X, Y \mid Z = z)$, as in Table 6.2.

	$P(X=1,Y=1 \mid Z)$	$P(X=1,Y=2 \mid Z)$	$P(X=2,Y=1 \mid Z)$	$P(X=2,Y=2 \mid Z)$	$P(X=3,Y=1 \mid Z)$	$P(X=3,Y=2 \mid Z)$
$Z=1$	0.126	0.198	0.100	0.027	0.540	0.009
$Z=2$	0.060	0.148	0.126	0.233	0.022	0.411
$Z=3$	0.119	0.065	0.178	0.116	0.461	0.061

Table 6.2: A synthetic data with exposure and instrument taking 3 levels, and a binary outcome

Estimand	$P(Y(x_1)=2)$	$P(Y(x_2)=2)$	$P(Y(x_2)=2) - P(Y(x_1)=2)$	$(\min(P(Y(x_2)=2)) - \max(P(Y(x_1)=2)), \max(P(Y(x_2)=2)) - \min(P(Y(x_1)=2)))$
Partial Identification Bound	(0.297, 0.863)	(0.233, 0.822)	(-0.600, 0.477)	(-0.630, 0.525)

Table 6.3: Upper and lower bounds on various estimands using Theorem 2

Bound on the fifth column is calculated by the difference of the upper and lower bounds of $P(Y(x_1)=2)$ and $P(Y(x_2)=2)$, i.e. $(0.233 - 0.863, 0.822 - 0.297)$. Comparing it to the fourth column in Table 6.3, we can see that both the upper and the lower bound on the pair-wise ATE, $P(Y(x_2)=2) - P(Y(x_1)=2)$, are not sharp.

This example is a direct implication of the variation dependence properties of $P(Y(x_1))$, $P(Y(x_2))$ and $P(Y(x_3))$ when $K = 3$ and $M = Q = 2$. We can see that obtaining the upper of lower bounds on the difference of $P(Y(x_1) = 2)$ and $P(Y(x_2) = 2)$ using the difference of the upper and lower bounds on $P(Y(x_1) = 2)$ and $P(Y(x_2) = 2)$ can result in wider bounds. This provides clear practical guidance for researchers that they should obtain partial identification bounds on the pair-wise ATE treating it as the estimand of interest, instead of analyzing $P(Y(x_1) = 2)$ and $P(Y(x_2) = 2)$ separately. This illustrates the

importance of our results in Chapter 2 that is developed on the joint counterfactual distribution $P(Y(x_1), \ldots, Y(x_K))$, which is a more fundamental distribution than the marginal counterfactual distribution $P(Y(x_i))$.

Chapter 7

CONCLUSION

7.1 Discussion

Our generalization of the bounds characterizing the IV model with categorical instrument, treatment, and outcome can also find intriguing applications in the field of quantum mechanics, particularly in the analysis of Bell's inequality. In causal inference, our work amongst others provides bounds on the causal effects when dealing with unmeasured confounding. Analogously, in quantum mechanics, Bell's inequalities offer a way to test the presence of local hidden variables and the extent of quantum entanglement, revealing non-classical correlations in the physical world. The essential argument of a Bell test involves a rejection of any 'classical' influences on any observed correlation using a causal analysis, and thus, the causes of physical phenomena could only be explained by non-classical physical natures. Pearl [1995] pointed out the similarities between Bell's inequality in quantum mechanics [Cushing and McMullin, 1989, Suppes, 1988] and the IV inequalities that both of them address a set of observed correlations that cannot be accounted for by assuming hidden common causes. Furthermore, in situations where a direct causal link between the correlated variables X and Y is allowed, the IV inequalities are actually a variant form of Bell's inequality. We believe this work can enhance our understanding of non-classical causality and correlations in quantum mechanics and could shed new light on the quantification of causal effects within quantum networks, potentially providing a fresh perspective on the nature of quantum entanglement and informing experimental designs aimed at testing foundational principles of quantum theory. We hope the synergy between these two frameworks may help bridge classical causal inference methods and quantum information science, offering a more nuanced characterization of quantum correlations that go beyond Bell's classical limits.

7.2 Future Directions

In summary, we consider the instrumental variable model with categorical Y taking M states, categorical X taking K states, and categorical Z taking Q states. We assume there is no direct effect of Z on Y so that $Y(x, z) = Y(x)$ and discussed variations of independence conditions. We first provide a simple characterization of the set of joint distributions of the potential outcomes $P(Y(x_1), \ldots, Y(x_K))$ compatible with a given observed probability distribution $P(X, Y|Z)$. Our bounds are necessary, sufficient, and non-redundant. Partial identification bounds on any linear function of the joint counterfactual probabilities could also be obtained using our results. Results in Richardson and Robins [2014] and the Balke-Pearl bounds [Balke and Pearl, 1997] are special cases of our work with binary X and Y. Using the existing algorithm proposed by Guo and Richardson [2021], we give a general procedure of how confidence regions on any convex functions of the joint counterfactual probabilities could be obtained. By applying the method to real data on a financial incentive program for smoking cessation, we show how results can be impacted by how researchers handle and analyze the data. Specifically, data on additional instrument arms could potentially improve the efficiency of the results and give tighter bounds. Any deletion of the treatment groups to restrict the data to a subset of the population will change the estimand of interest (restricting the parameter of interest, e.g. ATE, to the sub-population which might not be well-defined) and may lead to invalid/wrong inference.

We will leave closed-form characterization on the IV model with continuous exposure, outcome, and/or instrument and on the marginal counterfactual probabilities $P(Y(x_1)), \ldots,$ $P(Y(x_K))$ to future work although bounds on them could be obtained numerically using our results. Further, the use of the monotonicity assumption to obtain point identification of the ATE should also be explored in the categorical IV model. Finally, the algorithm proposed by Guo and Richardson [2021] could be conservative. Other statistical inference methods could be developed using our characterization of the categorical IV model.

BIBLIOGRAPHY

F. Richard Guo and Thomas S. Richardson. Chernoff-type concentration of empirical probabilities in relative entropy. *IEEE Transactions on Information Theory*, 67:549–558, 2021. URL https://api.semanticscholar.org/CorpusID:213004900. i, 2, 51, 74

Lawrence W. Sherman and Richard A. Berk. The minneapolis domestic violence experiment. Technical report, National Policing Institute, Washington, DC, 1984. URL https://www.policinginstitute.org/publication/the-minneapolis-domestic-violence-experiment/. Report. 2

Joshua D. Angrist. Instrumental variables methods in experimental criminological research: what, why and how. *Journal of Experimental Criminology*, 2:23–44, 2006. doi: 10.1007/s11292-005-5126-x. URL https://doi.org/10.1007/s11292-005-5126-x. 2, 5

Sonja A. Swanson, James M. Robins, Matthew Miller, and Miguel A. Hernán. Selecting on Treatment: A Pervasive Form of Bias in Instrumental Variable Analyses. *American Journal of Epidemiology*, 181(3):191–197, 01 2015. ISSN 0002-9262. doi: 10.1093/aje/kwu284. URL https://doi.org/10.1093/aje/kwu284. 5

Thomas M. Russell. Sharp bounds on functionals of the joint distribution in the analysis of treatment effects. *Journal of Business & Economic Statistics*, 39(2):532–546, 2021. doi: 10.1080/07350015.2019.1684300. URL https://doi.org/10.1080/07350015.2019.1684300. 6, 33, 45, 46

Guido W. Imbens and Joshua D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/2951620. 7

Joshua D. Angrist and Guido W. Imbens. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90(430):431–442, 1995. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2291054. 7

Joshua Angrist, Guido Imbens, and Donald Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996. doi: 10.1080/01621459.1996.10476902. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1996.10476902. 7

James M Robins, Lee Sechrest, H Freeman, and A Mulley. Health service research methodology: a focus on aids. *Washington, DC: US Public Health Service, National Center for Health Services Research*, pages 113–59, 1989. 7

Charles F Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323, 1990. 7

Alexander Balke and Judea Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997. doi: 10.1080/01621459.1997.10474074. URL https://doi.org/10.1080/01621459.1997.10474074. 7, 74

Thomas S. Richardson and James M. Robins. Ace bounds; sems with equilibrium conditions. *Statistical Science*, 29(3):363–366, 2014. ISSN 08834237, 21688745. URL http://www.jstor.org/stable/43288513. 7, 8, 11, 12, 22, 45, 60, 74

Sonja A. Swanson, Miguel A. Hernán, Matthew Miller, James M. Robins, and Thomas S. Richardson. Partial identification of the average treatment effect using instrumental variables: Review of methods for binary instruments, treatments, and outcomes. *Journal of the American Statistical Association*, 113(522):933–947, 2018. doi: 10.1080/01621459.2018.

1434530. URL https://doi.org/10.1080/01621459.2018.1434530. PMID: 31537952. 8, 10

Blai Bonet. Instrumentality tests revisited. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, UAI '01, page 48–55, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558608001. URL https://arxiv.org/abs/1301.2258. 8, 60

Judea Pearl. On the testability of causal models with latent and instrumental variables. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, page 435–443, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1558603859. 8, 60, 73

Désiré Kédagni and Ismael Mourifié. Generalized instrumental inequalities: testing the instrumental variable independence assumption. *Biometrika*, 107(3):661–675, 02 2020. ISSN 0006-3444. doi: 10.1093/biomet/asaa003. URL https://doi.org/10.1093/biomet/asaa003. 8, 60

Jing Cheng and Dylan S. Small. Bounds on Causal Effects in Three-Arm Trials With Non-Compliance. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(5):815–836, 10 2006. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2006.00568.x. URL https://doi.org/10.1111/j.1467-9868.2006.00568.x. 8

Arie Beresteanu, Ilya Molchanov, and Francesca Molinari. Partial identification using random set theory. *Journal of Econometrics*, 166(1):17–32, 2012. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2011.06.003. URL https://www.sciencedirect.com/science/article/pii/S0304407611001163. 8, 45

Keisuke Hirano, Guido W Imbens, Donald B Rubin, and Xiao-Hua Zhou. Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, 1(1):69–88, Mar 2000. doi: 10.1093/biostatistics/1.1.69. 10

Toru Kitagawa. The identification region of the potential outcome distributions under instrument independence. *Journal of Econometrics*, 225(2):231–253, 2021. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2021.03.006. URL https://www.sciencedirect.com/science/article/pii/S0304407621000968. Themed Issue: Treatment Effect 1. 11

A Philip Dawid. Causal inference using influence diagrams: the problem of partial compliance. In *Highly Structured Stochastic Systems*. Oxford University Press, 05 2003. ISBN 9780198510550. doi: 10.1093/oso/9780198510550.003.0005. URL https://doi.org/10.1093/oso/9780198510550.003.0005. 11

V. Strassen. The Existence of Probability Measures with Given Marginals. *The Annals of Mathematical Statistics*, 36(2):423 – 439, 1965. doi: 10.1214/aoms/1177700153. URL https://doi.org/10.1214/aoms/1177700153. 16

Shmuel Friedland, Jingtong Ge, and Lihong Zhi. Quantum strassen's theorem. *arXiv: Functional Analysis*, 2019. URL https://api.semanticscholar.org/CorpusID:155099804. 16

Twan Koperberg. Couplings and matchings: Combinatorial notes on strassen's theorem. *Statistics & Probability Letters*, 209:110089, 2024. ISSN 0167-7152. doi: https://doi.org/10.1016/j.spl.2024.110089. URL https://www.sciencedirect.com/science/article/pii/S0167715224000580. 16, 17

Thomas S. Richardson and James M. Robins. Assumptions and bounds in the instrumental variable model, 2024. URL https://arxiv.org/abs/2401.13758. 20, 24, 63, 64

Günter M. Ziegler. *Lectures on polytopes*. Springer-Verlag, New York, 1995. URL http://www.worldcat.org/search?qt=worldcat_org_all&q=9780387943657. 31

Ye Luo and Hai Wang. Core determining class and inequality selection. *American Economic*

Review, 107(5):274–77, May 2017. doi: 10.1257/aer.p20171041. URL https://www.aeaweb.org/articles?id=10.1257/aer.p20171041. 45

James T. Cushing and Ernan McMullin, editors. *Philoophical Consequences of Quantum Theory*. University of Notre Dame Press, 1989. 73

Patrick Suppes. *Probabilistic Causality in Space and Time*, pages 135–151. Springer Netherlands, Dordrecht, 1988. ISBN 978-94-009-2863-3. doi: 10.1007/978-94-009-2863-3_9. URL https://doi.org/10.1007/978-94-009-2863-3_9. 73

I Mourifié, M Henry, and R Meango. Sharp bounds for the Roy model. *Unpublished manuscript*, 2015.

Scott D. Halpern, Benjamin French, Dylan S. Small, Kathryn Saulsgiver, Michael O. Harhay, Janet Audrain-McGovern, George Loewenstein, Troyen A. Brennan, David A. Asch, and Kevin G. Volpp. Randomized trial of four financial-incentive programs for smoking cessation. *New England Journal of Medicine*, 372(22):2108–2117, 2015. doi: 10.1056/NEJMoa1414293. URL https://www.nejm.org/doi/full/10.1056/NEJMoa1414293.

Appendix A

CODE

A.1 Code for Table 4.1: data analysis of the Minneapolis Domestic Violence Data

We used the following code to compute the plug-in bounds and confidence regions in Table 4.1.

```
library(multChernoff)
library(CVXR)
library(rcdd)

crim_all<-matrix(c(81, 10, 0, 0, 1, 0,
                   15, 3, 69, 15, 3, 3,
                   21, 5, 4, 1, 62, 20), nrow = 3, byrow = TRUE)

crim_ac<-matrix(c(56, 7, 0, 0, 1, 0,
                  15, 3, 38, 7, 2, 2,
                  18, 4, 1, 1, 31, 9), nrow = 3, byrow = TRUE)

crim_nac<-matrix(c(25, 3, 0, 0, 0, 0,
                   0, 0, 31, 8, 1, 1,
                   3, 1, 3, 0, 31, 11), nrow = 3, byrow = TRUE)
crim_ac+crim_nac==crim_all
```

Researcher 1 – use all data to compute the bound for pair-wise ATEs.

```

##### Plug-in bound

get.naive.3x.3z.2y<-function(data){
  n.z <- apply(data, 1, sum)
  p.empirical <- data / n.z
  # make objective -----
  p1 <- Variable(8) # this is on the joint of counterfactual

  # additional bound from Strassen's theorem using RCDD
```

```
p <- 3 # number of levels for x

# matrix to feed into rcdd
vmat <- matrix(nrow=(p*(2^p)),ncol=((2^p)+p*2))

# # counterfactual matrix:

factr.list <- list()
for(i in 1:p){
  factr.list[[i]] <- c(0,1)
}
factr.list[[p+1]] <- (1:p)


cntr.fact.names<-as.matrix(expand.grid(factr.list),nrow=p*(2^p))

cntr.facts <- matrix(0,nrow=(p*(2^p)),ncol=((2^p)+1))

cntr.facts[,1:(2^p)] <- matrix(rep(t(diag(1,2^p)), p), ncol = (2^p),
   byrow = TRUE)
cntr.facts[,(2^p)+1] <- rep((1:p),rep(2^p,p))

obs.dist <- matrix(0,nrow=(p*(2^p)),ncol=p*2)
for(i in 1:(p*(2^p))){
  xval <- cntr.fact.names[i,p+1]
  yval <- cntr.fact.names[i,xval]
  obs.dist[i,2*(xval-1)+yval+1] <- 1
}

vmat[1:(p*(2^p)),1:(2^p)] <- cntr.facts[1:(p*(2^p)),1:(2^p)]
vmat[1:(p*(2^p)),(((2^p)+1):((2^p)+2*p))] <- obs.dist

vmat <- cbind(rep(0,(p*(2^p))), rep(1,(p*(2^p))),vmat)

colnames(vmat)<- c("eq","inter","Y0=0,Y1=0,Y2=0", "Y0=1,Y1=0,Y2=0",
   "Y0=0,Y1=1,Y2=0", "Y0=1,Y1=1,Y2=0",
                   "Y0=0,Y1=0,Y2=1", "Y0=1,Y1=0,Y2=1",
                      "Y0=0,Y1=1,Y2=1", "Y0=1,Y1=1,Y2=1",
                   "X=0,Y=0","X=0,Y=1", "X=1,Y=0","X=1,Y=1",
                      "X=2,Y=0","X=2,Y=1")

hmat <- scdd(vmat,representation="V")$output
colnames(hmat)<- c("eq","inter","Y0=0,Y1=0,Y2=0", "Y0=1,Y1=0,Y2=0",
```

```
      "Y0=0,Y1=1,Y2=0", "Y0=1,Y1=1,Y2=0",
50                      "Y0=0,Y1=0,Y2=1", "Y0=1,Y1=0,Y2=1",
                           "Y0=0,Y1=1,Y2=1", "Y0=1,Y1=1,Y2=1",
51                      "X=0,Y=0","X=0,Y=1", "X=1,Y=0","X=1,Y=1",
                           "X=2,Y=0","X=2,Y=1")
52
53  hmat.ineq <- hmat[hmat[,1]==0 & apply(abs(hmat[,3:10]), 1, sum)!=0 &
       apply(abs(hmat[,11:16]), 1, sum)!=0,]
54
55  p2<-as.matrix(p.empirical)
56  constr1.lp <- hmat.ineq[,3:10] %*% p1 + hmat.ineq[,11:16] %*%
       matrix(p2[1,], ncol=1) >= -hmat.ineq[,2]
57  constr2.lp <- hmat.ineq[,3:10] %*% p1 + hmat.ineq[,11:16] %*%
       matrix(p2[2,], ncol=1) >= -hmat.ineq[,2]
58  constr3.lp <- hmat.ineq[,3:10] %*% p1 + hmat.ineq[,11:16] %*%
       matrix(p2[3,], ncol=1) >= -hmat.ineq[,2]
59  constr <- list(constr1.lp, constr2.lp,constr3.lp, p1 >= 0,
       sum_entries(p1, axis=2) == 1)
60
61  # objective function
62  obj <- p1[3] + p1[7] - p1[2] - p1[6] #arm 1vs.0
63  obj <- p1[5] + p1[7] - p1[2] - p1[4] #arm 2vs.0
64  obj <- p1[5] + p1[6] - p1[3] - p1[4] #arm 2vs.1
65
66  # ACE lower bound ----
67  prob <- Problem(Minimize(obj), constr)
68  result <- solve(prob)
69
70  #print(result$status)
71  ACE.lb <- result$value
72  # p.lb <- result$getValue(p)
73
74  # ACE upper bound ----
75  prob <- Problem(Maximize(obj), constr)
76  result <- solve(prob, solver="ECOS")
77  #print(result$status)
78  ACE.ub <- result$value
79  # p.ub <- result$getValue(p)
80  return (c(ACE.lb, ACE.ub))
81
82 }
83
84 round(get.naive.3x.3z.2y(crim_all),3)
```

```
round(get.naive.3x.3z.2y(crim_ac),3)
round(get.naive.3x.3z.2y(crim_nac),3)


###### Confidence Region

get.cr.3x.3z.2y<-function(data, alpha){
  n.z <- apply(data, 1, sum)
  p.empirical <- data / n.z

  # get critical value ----
  t.alpha <- criticalValue(c(6,6,6), n.z, p=alpha, verbose = TRUE)
  # cat(sprintf("critical value = %f\n", t.alpha))


  # make objective -----
  p1 <- Variable(8) # this is on the joint of counterfactual
  p2 <- Variable(3,6)

  # first bound on KL divergence
  KL <- sum(sum_entries(p.empirical * (log(p.empirical) - log(p2)),
     axis=1) * 2 * n.z)

  # additional bound from Strassen's theorem using RCDD
  p <- 3 # number of levels for x

  # matrix to feed into rcdd
  vmat <- matrix(nrow=(p*(2^p)),ncol=((2^p)+p*2))

  # # counterfactual matrix:

  factr.list <- list()
  for(i in 1:p){
    factr.list[[i]] <- c(0,1)
  }
  factr.list[[p+1]] <- (1:p)


  cntr.fact.names<-as.matrix(expand.grid(factr.list),nrow=p*(2^p))

  cntr.facts <- matrix(0,nrow=(p*(2^p)),ncol=((2^p)+1))

  cntr.facts[,1:(2^p)] <- matrix(rep(t(diag(1,2^p)), p), ncol = (2^p),
```

```
    byrow = TRUE)
cntr.facts[,(2^p)+1] <- rep((1:p),rep(2^p,p))

obs.dist <- matrix(0,nrow=(p*(2^p)),ncol=p*2)
for(i in 1:(p*(2^p))){
  xval <- cntr.fact.names[i,p+1]
  yval <- cntr.fact.names[i,xval]
  obs.dist[i,2*(xval-1)+yval+1] <- 1
}

vmat[1:(p*(2^p)),1:(2^p)] <- cntr.facts[1:(p*(2^p)),1:(2^p)]
vmat[1:(p*(2^p)),(((2^p)+1):((2^p)+2*p))] <- obs.dist

vmat <- cbind(rep(0,(p*(2^p))), rep(1,(p*(2^p))),vmat)

colnames(vmat)<- c("eq","inter","Y0=0,Y1=0,Y2=0", "Y0=1,Y1=0,Y2=0",
    "Y0=0,Y1=1,Y2=0", "Y0=1,Y1=1,Y2=0",
                   "Y0=0,Y1=0,Y2=1", "Y0=1,Y1=0,Y2=1",
                      "Y0=0,Y1=1,Y2=1", "Y0=1,Y1=1,Y2=1",
                   "X=0,Y=0","X=0,Y=1", "X=1,Y=0","X=1,Y=1",
                      "X=2,Y=0","X=2,Y=1")

hmat <- scdd(vmat,representation="V")$output
colnames(hmat)<- c("eq","inter","Y0=0,Y1=0,Y2=0", "Y0=1,Y1=0,Y2=0",
    "Y0=0,Y1=1,Y2=0", "Y0=1,Y1=1,Y2=0",
                   "Y0=0,Y1=0,Y2=1", "Y0=1,Y1=0,Y2=1",
                      "Y0=0,Y1=1,Y2=1", "Y0=1,Y1=1,Y2=1",
                   "X=0,Y=0","X=0,Y=1", "X=1,Y=0","X=1,Y=1",
                      "X=2,Y=0","X=2,Y=1")

hmat.ineq <- hmat[hmat[,1]==0 & apply(abs(hmat[,3:10]), 1, sum)!=0 &
    apply(abs(hmat[,11:16]), 1, sum)!=0,]

constr1 <- hmat.ineq[,3:10] %*% p1 + hmat.ineq[,11:16] %*% t(p2[1,])
    >= -hmat.ineq[,2]
constr2 <- hmat.ineq[,3:10] %*% p1 + hmat.ineq[,11:16] %*% t(p2[2,])
    >= -hmat.ineq[,2]
constr3 <- hmat.ineq[,3:10] %*% p1 + hmat.ineq[,11:16] %*% t(p2[3,])
    >= -hmat.ineq[,2]
constr4 <- KL <= t.alpha
constr <- list(constr1, constr2, constr3, constr4, p1 >= 0, p2 >=0,
    sum_entries(p1, axis=2) == 1, sum_entries(p2, axis=1) == 1)

```

```
  # Not sure what the order exactly is... Try chronological order below
  # ACE lower bound ----
  obj <- p1[3] + p1[7] - p1[2] - p1[6] #arm 1vs.0
  obj <- p1[5] + p1[7] - p1[2] - p1[4] #arm 2vs.0
  obj <- p1[5] + p1[6] - p1[3] - p1[4] #arm 2vs.1
  prob <- Problem(Minimize(obj), constr)
  result <- solve(prob)

  #print(result$status)
  ACE.lb <- result$value
  # p.lb <- result$getValue(p)

  # ACE upper bound ----
  prob <- Problem(Maximize(obj), constr)
  result <- solve(prob, solver="ECOS")
  #print(result$status)
  ACE.ub <- result$value
  # p.ub <- result$getValue(p)
  return (c(ACE.lb, ACE.ub))
}



round(get.cr.3x.3z.2y(crim_all, 0.05),3)
round(get.cr.3x.3z.2y(crim_ac, 0.05),3)
round(get.cr.3x.3z.2y(crim_nac, 0.05),3)
```

Researcher 2 – use data with 1 less Z arm to compute the bound for pair-wise ATEs

```

#### get plug-in bounds

get.naive.3x.2z.2y<-function(data, alpha){
  n.z <- apply(data, 1, sum)
  p.empirical <- data / n.z
  # make objective -----
  p1 <- Variable(8) # this is on the joint of counterfactual

  # additional bound from Strassen's theorem using RCDD
  p <- 3 # number of levels for x

  # matrix to feed into rcdd
  vmat <- matrix(nrow=(p*(2^p)),ncol=((2^p)+p*2))

```

```
16  # # counterfactual matrix:
17
18  factr.list <- list()
19  for(i in 1:p){
20    factr.list[[i]] <- c(0,1)
21  }
22  factr.list[[p+1]] <- (1:p)
23
24
25  cntr.fact.names<-as.matrix(expand.grid(factr.list),nrow=p*(2^p))
26
27  cntr.facts <- matrix(0,nrow=(p*(2^p)),ncol=((2^p)+1))
28
29  cntr.facts[,1:(2^p)] <- matrix(rep(t(diag(1,2^p)), p), ncol = (2^p),
       byrow = TRUE)
30  cntr.facts[,(2^p)+1] <- rep((1:p),rep(2^p,p))
31
32  obs.dist <- matrix(0,nrow=(p*(2^p)),ncol=p*2)
33  for(i in 1:(p*(2^p))){
34    xval <- cntr.fact.names[i,p+1]
35    yval <- cntr.fact.names[i,xval]
36    obs.dist[i,2*(xval-1)+yval+1] <- 1
37  }
38
39  vmat[1:(p*(2^p)),1:(2^p)] <- cntr.facts[1:(p*(2^p)),1:(2^p)]
40  vmat[1:(p*(2^p)),(((2^p)+1):((2^p)+2*p))] <- obs.dist
41
42  vmat <- cbind(rep(0,(p*(2^p))), rep(1,(p*(2^p))),vmat)
43
44  colnames(vmat)<- c("eq","inter","Y0=0,Y1=0,Y2=0", "Y0=1,Y1=0,Y2=0",
       "Y0=0,Y1=1,Y2=0", "Y0=1,Y1=1,Y2=0",
45                     "Y0=0,Y1=0,Y2=1", "Y0=1,Y1=0,Y2=1",
                          "Y0=0,Y1=1,Y2=1", "Y0=1,Y1=1,Y2=1",
46                     "X=0,Y=0","X=0,Y=1", "X=1,Y=0","X=1,Y=1",
                          "X=2,Y=0","X=2,Y=1")
47
48  hmat <- scdd(vmat,representation="V")$output
49  colnames(hmat)<- c("eq","inter","Y0=0,Y1=0,Y2=0", "Y0=1,Y1=0,Y2=0",
       "Y0=0,Y1=1,Y2=0", "Y0=1,Y1=1,Y2=0",
50                     "Y0=0,Y1=0,Y2=1", "Y0=1,Y1=0,Y2=1",
                          "Y0=0,Y1=1,Y2=1", "Y0=1,Y1=1,Y2=1",
51                     "X=0,Y=0","X=0,Y=1", "X=1,Y=0","X=1,Y=1",
                          "X=2,Y=0","X=2,Y=1")
```

```
52
53   hmat.ineq <- hmat[hmat[,1]==0 & apply(abs(hmat[,3:10]), 1, sum)!=0 &
        apply(abs(hmat[,11:16]), 1, sum)!=0,]
54
55   p2<-as.matrix(p.empirical)
56   constr1.lp <- hmat.ineq[,3:10] %*% p1 + hmat.ineq[,11:16] %*%
        matrix(p2[1,], ncol=1) >= -hmat.ineq[,2]
57   constr2.lp <- hmat.ineq[,3:10] %*% p1 + hmat.ineq[,11:16] %*%
        matrix(p2[2,], ncol=1) >= -hmat.ineq[,2]
58   #  constr3.lp <- hmat.ineq[,3:10] %*% p1 + hmat.ineq[,11:16] %*%
        matrix(p2[3,], ncol=1) >= -hmat.ineq[,2]
59   constr <- list(constr1.lp, constr2.lp, p1 >= 0, sum_entries(p1,
        axis=2) == 1)
60
61   # Objective
62
63   obj <- p1[3] + p1[7] - p1[2] - p1[6] #arm 1vs.0
64   obj <- p1[5] + p1[7] - p1[2] - p1[4] #arm 2vs.0
65   obj <- p1[5] + p1[6] - p1[3] - p1[4] #arm 2vs.1
66
67
68   # ACE lower bound ----
69   prob <- Problem(Minimize(obj), constr)
70   result <- solve(prob)
71
72   #print(result$status)
73   ACE.lb <- result$value
74   # p.lb <- result$getValue(p)
75
76   # ACE upper bound ----
77   prob <- Problem(Maximize(obj), constr)
78   result <- solve(prob, solver="ECOS")
79   #print(result$status)
80   ACE.ub <- result$value
81   # p.ub <- result$getValue(p)
82   return (c(ACE.lb, ACE.ub))
83 }
84
85
86 round(get.naive.3x.2z.2y(crim_all[-1,]),3)
87 round(get.naive.3x.2z.2y(crim_all[-2,]),3)
88 round(get.naive.3x.2z.2y(crim_all[-3,]),3)
89
```

```
round(get.naive.3x.2z.2y(crim_ac[-1,]),3)
round(get.naive.3x.2z.2y(crim_ac[-2,]),3)
round(get.naive.3x.2z.2y(crim_ac[-3,]),3)

round(get.naive.3x.2z.2y(crim_nac[-1,]),3)
round(get.naive.3x.2z.2y(crim_nac[-2,]),3)
round(get.naive.3x.2z.2y(crim_nac[-3,]),3)



### Confidence region

get.cr.3x.2z.2y<-function(data, alpha){
  n.z <- apply(data, 1, sum)
  p.empirical <- data / n.z

  # get critical value ----
  alpha <- 0.05
  t.alpha <- criticalValue(rep(6, 2), n.z, p=alpha, verbose = TRUE)
  # cat(sprintf("critical value = %f\n", t.alpha))


  # make objective -----
  p1 <- Variable(8) # this is on the joint of counterfactual
  p2 <- Variable(2,6)

  # first bound on KL divergence
  KL <- sum(sum_entries(p.empirical * (log(p.empirical) - log(p2)),
     axis=1) * 2 * n.z)

  # additional bound from Strassen's theorem using RCDD
  p <- 3 # number of levels for x

  # matrix to feed into rcdd
  vmat <- matrix(nrow=(p*(2^p)),ncol=((2^p)+p*2))

  # # counterfactual matrix:

  factr.list <- list()
  for(i in 1:p){
    factr.list[[i]] <- c(0,1)
  }
```

```
132 factr.list[[p+1]] <- (1:p)
133
134
135 cntr.fact.names<-as.matrix(expand.grid(factr.list),nrow=p*(2^p))
136
137 cntr.facts <- matrix(0,nrow=(p*(2^p)),ncol=((2^p)+1))
138
139 cntr.facts[,1:(2^p)] <- matrix(rep(t(diag(1,2^p)), p), ncol = (2^p),
       byrow = TRUE)
140 cntr.facts[,(2^p)+1] <- rep((1:p),rep(2^p,p))
141
142 obs.dist <- matrix(0,nrow=(p*(2^p)),ncol=p*2)
143 for(i in 1:(p*(2^p))){
144   xval <- cntr.fact.names[i,p+1]
145   yval <- cntr.fact.names[i,xval]
146   obs.dist[i,2*(xval-1)+yval+1] <- 1
147 }
148
149 vmat[1:(p*(2^p)),1:(2^p)] <- cntr.facts[1:(p*(2^p)),1:(2^p)]
150 vmat[1:(p*(2^p)),(((2^p)+1):((2^p)+2*p))] <- obs.dist
151
152 vmat <- cbind(rep(0,(p*(2^p))), rep(1,(p*(2^p))),vmat)
153
154 colnames(vmat)<- c("eq","inter","Y0=0,Y1=0,Y2=0", "Y0=1,Y1=0,Y2=0",
       "Y0=0,Y1=1,Y2=0", "Y0=1,Y1=1,Y2=0",
155                    "Y0=0,Y1=0,Y2=1", "Y0=1,Y1=0,Y2=1",
                          "Y0=0,Y1=1,Y2=1", "Y0=1,Y1=1,Y2=1",
156                    "X=0,Y=0","X=0,Y=1", "X=1,Y=0","X=1,Y=1",
                          "X=2,Y=0","X=2,Y=1")
157
158 hmat <- scdd(vmat,representation="V")$output
159 colnames(hmat)<- c("eq","inter","Y0=0,Y1=0,Y2=0", "Y0=1,Y1=0,Y2=0",
       "Y0=0,Y1=1,Y2=0", "Y0=1,Y1=1,Y2=0",
160                    "Y0=0,Y1=0,Y2=1", "Y0=1,Y1=0,Y2=1",
                          "Y0=0,Y1=1,Y2=1", "Y0=1,Y1=1,Y2=1",
161                    "X=0,Y=0","X=0,Y=1", "X=1,Y=0","X=1,Y=1",
                          "X=2,Y=0","X=2,Y=1")
162
163 hmat.ineq <- hmat[hmat[,1]==0 & apply(abs(hmat[,3:10]), 1, sum)!=0 &
       apply(abs(hmat[,11:16]), 1, sum)!=0,]
164
165 constr1 <- hmat.ineq[,3:10] %*% p1 + hmat.ineq[,11:16] %*% t(p2[1,])
       >= -hmat.ineq[,2]
```

```
  constr2 <- hmat.ineq[,3:10] %*% p1 + hmat.ineq[,11:16] %*% t(p2[2,])
     >= -hmat.ineq[,2]
  constr4 <- KL <= t.alpha
  constr <- list(constr1, constr2, constr4, p1 >= 0, p2 >=0,
     sum_entries(p1, axis=2) == 1, sum_entries(p2, axis=1) == 1)

  # Objective function
  #obj <- p1[3] + p1[7] - p1[2] - p1[6] #arm 1vs.0
  #obj <- p1[5] + p1[7] - p1[2] - p1[4] #arm 2vs.0
  obj <- p1[5] + p1[6] - p1[3] - p1[4] #arm 2vs.1

  # ACE lower bound ----
  prob <- Problem(Minimize(obj), constr)
  result <- solve(prob)

  #print(result$status)
  ACE.lb <- result$value
  # p.lb <- result$getValue(p)

  # ACE upper bound ----
  prob <- Problem(Maximize(obj), constr)
  result <- solve(prob, solver="ECOS")
  #print(result$status)
  ACE.ub <- result$value
  # p.ub <- result$getValue(p)
  return (c(ACE.lb, ACE.ub))
}



round(get.cr.3x.2z.2y(crim_all[-1,], 0.05),3)
round(get.cr.3x.2z.2y(crim_all[-2,], 0.05),3)
round(get.cr.3x.2z.2y(crim_all[-3,], 0.05),3)

round(get.cr.3x.2z.2y(crim_ac[-1,], 0.05),3)
round(get.cr.3x.2z.2y(crim_ac[-2,], 0.05),3)
round(get.cr.3x.2z.2y(crim_ac[-3,], 0.05),3)

round(get.cr.3x.2z.2y(crim_nac[-1,], 0.05),3)
round(get.cr.3x.2z.2y(crim_nac[-2,], 0.05),3)
round(get.cr.3x.2z.2y(crim_nac[-3,], 0.05),3)
```

Researcher 3 – use data with 1 less X arm to compute the bound for the pair-wise ATEs

```
#### get plug-in bounds

get.naive.2x.3z.2y<-function(data){
  n.z <- apply(data, 1, sum)
  p.empirical <- data / n.z
  # make objective -----
  p1 <- Variable(4) # this is on the joint of counterfactual

  # additional bound from Strassen's theorem using RCDD
  p <- 2 # number of levels for x

  # matrix to feed into rcdd
  vmat <- matrix(nrow=(p*(2^p)),ncol=((2^p)+p*2))

  # # counterfactual matrix:

  factr.list <- list()
  for(i in 1:p){
    factr.list[[i]] <- c(0,1)
  }
  factr.list[[p+1]] <- (1:p)


  cntr.fact.names<-as.matrix(expand.grid(factr.list),nrow=p*(2^p))

  cntr.facts <- matrix(0,nrow=(p*(2^p)),ncol=((2^p)+1))

  cntr.facts[,1:(2^p)] <- matrix(rep(t(diag(1,2^p)), p), ncol = (2^p),
     byrow = TRUE)
  cntr.facts[,(2^p)+1] <- rep((1:p),rep(2^p,p))

  obs.dist <- matrix(0,nrow=(p*(2^p)),ncol=p*2)
  for(i in 1:(p*(2^p))){
    xval <- cntr.fact.names[i,p+1]
    yval <- cntr.fact.names[i,xval]
    obs.dist[i,2*(xval-1)+yval+1] <- 1
  }

  vmat[1:(p*(2^p)),1:(2^p)] <- cntr.facts[1:(p*(2^p)),1:(2^p)]
  vmat[1:(p*(2^p)),(((2^p)+1):((2^p)+2*p))] <- obs.dist

  vmat <- cbind(rep(0,(p*(2^p))), rep(1,(p*(2^p))),vmat)

```

```
43   hmat <- scdd(vmat,representation="V")$output
44   colnames(hmat)<- c("eq","inter","Y0=0,Y1=0", "Y0=1,Y1=0",
        "Y0=0,Y1=1", "Y0=1,Y1=1",
45                      "Y=0,X=0","Y=1,X=0", "Y=0,X=1","Y=1,X=1")
46
47   hmat.ineq <- hmat[hmat[,1]==0 & apply(hmat[,3:6], 1, sum)!=0 &
        apply(hmat[,7:10], 1, sum)!=0,]
48
49   p2<-as.matrix(p.empirical)
50
51   constr1.lp <- hmat.ineq[,3:6] %*% p1 + hmat.ineq[,7:10] %*%
        matrix(p2[1,], ncol=1) >= -hmat.ineq[,2]
52   constr2.lp <- hmat.ineq[,3:6] %*% p1 + hmat.ineq[,7:10] %*%
        matrix(p2[2,], ncol=1) >= -hmat.ineq[,2]
53   constr3.lp <- hmat.ineq[,3:6] %*% p1 + hmat.ineq[,7:10] %*%
        matrix(p2[3,], ncol=1) >= -hmat.ineq[,2]
54   constr.lp <- list(constr1.lp, constr2.lp, constr3.lp, p1 >= 0,
        sum_entries(p1, axis=2) == 1)   # ACE lower bound ----
55   obj <- p1[3] - p1[2]
56   prob <- Problem(Minimize(obj), constr.lp)
57   result <- solve(prob)
58   #print(result$status)
59   ACE.lb <- result$value
60   # p.lb <- result$getValue(p)
61
62   # ACE upper bound ----
63   prob <- Problem(Maximize(obj), constr.lp)
64   result <- solve(prob)
65   #print(result$status)
66   ACE.ub <- result$value
67   # p.ub <- result$getValue(p)
68   return(c(ACE.lb, ACE.ub))
69
70 }
71
72
73 round(get.naive.2x.3z.2y(crim_all[,-c(5,6)]),3)
74 round(get.naive.2x.3z.2y(crim_all[,-c(3,4)]),3)
75 round(get.naive.2x.3z.2y(crim_all[,-c(1,2)]),3)
76
77 round(get.naive.2x.3z.2y(crim_ac[,-c(5,6)]),3)
78 round(get.naive.2x.3z.2y(crim_ac[,-c(3,4)]),3)
79 round(get.naive.2x.3z.2y(crim_ac[,-c(1,2)]),3)
```

```
80
81 round(get.naive.2x.3z.2y(crim_nac[,-c(5,6)]),3)
82 round(get.naive.2x.3z.2y(crim_nac[,-c(3,4)]),3)
83 round(get.naive.2x.3z.2y(crim_nac[,-c(1,2)]),3)
84
85
86
87
88 ### Confidence region
89
90 get.cr.2x.3z.2y<-function(data, alpha){
91   n.z <- apply(data, 1, sum)
92   p.empirical <- data / n.z
93
94   # get critical value ----
95   t.alpha <- criticalValue(rep(4, 3), n.z, p=alpha, verbose = TRUE)
96   #  cat(sprintf("critical value = %f\n", t.alpha))
97
98   # make objective -----
99   p1 <- Variable(4) # this is on the joint of counterfactual
100  p2 <- Variable(3,4)
101
102  # first bound on KL divergence
103  KL <- sum(sum_entries(p.empirical * (log(p.empirical) - log(p2)),
       axis=1) * 2 * n.z)
104
105  # additional bound from Strassen's theorem using RCDD
106  p <- 2 # number of levels for x
107
108  # matrix to feed into rcdd
109  vmat <- matrix(nrow=(p*(2^p)),ncol=((2^p)+p*2))
110
111  # # counterfactual matrix:
112
113  factr.list <- list()
114  for(i in 1:p){
115    factr.list[[i]] <- c(0,1)
116  }
117  factr.list[[p+1]] <- (1:p)
118
119
120  cntr.fact.names<-as.matrix(expand.grid(factr.list),nrow=p*(2^p))
121
```

```
122 cntr.facts <- matrix(0,nrow=(p*(2^p)),ncol=((2^p)+1))
123
124 cntr.facts[,1:(2^p)] <- matrix(rep(t(diag(1,2^p)), p), ncol = (2^p),
       byrow = TRUE)
125 cntr.facts[,(2^p)+1] <- rep((1:p),rep(2^p,p))
126
127 obs.dist <- matrix(0,nrow=(p*(2^p)),ncol=p*2)
128 for(i in 1:(p*(2^p))){
129   xval <- cntr.fact.names[i,p+1]
130   yval <- cntr.fact.names[i,xval]
131   obs.dist[i,2*(xval-1)+yval+1] <- 1
132 }
133
134 vmat[1:(p*(2^p)),1:(2^p)] <- cntr.facts[1:(p*(2^p)),1:(2^p)]
135 vmat[1:(p*(2^p)),(((2^p)+1):((2^p)+2*p))] <- obs.dist
136
137 vmat <- cbind(rep(0,(p*(2^p))), rep(1,(p*(2^p))),vmat)
138
139 hmat <- scdd(vmat,representation="V")$output
140 colnames(hmat)<- c("eq","inter","Y0=0,Y1=0", "Y0=1,Y1=0",
       "Y0=0,Y1=1", "Y0=1,Y1=1",
141                    "Y=0,X=0","Y=1,X=0", "Y=0,X=1","Y=1,X=1")
142
143 hmat.ineq <- hmat[hmat[,1]==0 & apply(hmat[,3:6], 1, sum)!=0 &
       apply(hmat[,7:10], 1, sum)!=0,]
144
145 constr1 <- hmat.ineq[,3:6] %*% p1 + hmat.ineq[,7:10] %*% t(p2[1,])
       >= -hmat.ineq[,2]
146 constr2 <- hmat.ineq[,3:6] %*% p1 + hmat.ineq[,7:10] %*% t(p2[2,])
       >= -hmat.ineq[,2]
147 constr3 <- hmat.ineq[,3:6] %*% p1 + hmat.ineq[,7:10] %*% t(p2[3,])
       >= -hmat.ineq[,2]
148 constr4 <- KL <= t.alpha
149 constr <- list(constr1, constr2, constr3, constr4, p1 >= 0, p2 >=0,
       sum_entries(p1, axis=2) == 1, sum_entries(p2, axis=1) == 1)
150
151 # Not sure what the order exactly is... Try chronological order below
152 # ACE lower bound ----
153 #obj <- p1[3] - p1[2]
154 obj <- p2[1,1]
155 prob <- Problem(Minimize(obj), constr)
156 result <- solve(prob)
157 #print(result$status)
```

```
  ACE.lb <- result$value
  # p.lb <- result$getValue(p)

  # ACE upper bound ----
  prob <- Problem(Maximize(obj), constr)
  result <- solve(prob, solver="ECOS")
  #  print(result$status)
  ACE.ub <- result$value
  # p.ub <- result$getValue(p)

  return(c(ACE.lb, ACE.ub))

}


round(get.cr.2x.3z.2y(crim_all[,-c(5,6)], 0.1),3)
round(get.cr.2x.3z.2y(crim_all[,-c(3,4)], 0.05),3)
round(get.cr.2x.3z.2y(crim_all[,-c(1,2)], 0.05),3)

round(get.cr.2x.3z.2y(crim_ac[,-c(5,6)], 0.05),3)
round(get.cr.2x.3z.2y(crim_ac[,-c(3,4)], 0.05),3)
round(get.cr.2x.3z.2y(crim_ac[,-c(1,2)], 0.05),3)

round(get.cr.2x.3z.2y(crim_nac[,-c(5,6)], 0.05),3)
round(get.cr.2x.3z.2y(crim_nac[,-c(3,4)], 0.05),3)
round(get.cr.2x.3z.2y(crim_nac[,-c(1,2)], 0.05),3)
```

Researcher 4 – use data with 1 less X arm and 1 less Z arm to compute the bound for the pair-wise ATEs

```
#### get plug-in bounds

get.naive.2x.2z.2y<-function(data, alpha){
  n.z <- apply(data, 1, sum)
  p.empirical <- data / n.z
  # make objective -----
  p1 <- Variable(4) # this is on the marginal of counterfactual

  # additional bound from Strassen's theorem using RCDD
  p <- 2 # number of levels for x

  # matrix to feed into rcdd
  vmat <- matrix(nrow=(p*(2^p)),ncol=((2^p)+p*2))
```

```
14
15  # # counterfactual matrix:
16
17  factr.list <- list()
18  for(i in 1:p){
19    factr.list[[i]] <- c(0,1)
20  }
21  factr.list[[p+1]] <- (1:p)
22
23
24  cntr.fact.names<-as.matrix(expand.grid(factr.list),nrow=p*(2^p))
25
26  cntr.facts <- matrix(0,nrow=(p*(2^p)),ncol=((2^p)+1))
27
28  cntr.facts[,1:(2^p)] <- matrix(rep(t(diag(1,2^p)), p), ncol = (2^p),
       byrow = TRUE)
29  cntr.facts[,(2^p)+1] <- rep((1:p),rep(2^p,p))
30
31  obs.dist <- matrix(0,nrow=(p*(2^p)),ncol=p*2)
32  for(i in 1:(p*(2^p))){
33    xval <- cntr.fact.names[i,p+1]
34    yval <- cntr.fact.names[i,xval]
35    obs.dist[i,2*(xval-1)+yval+1] <- 1
36  }
37
38  vmat[1:(p*(2^p)),1:(2^p)] <- cntr.facts[1:(p*(2^p)),1:(2^p)]
39  vmat[1:(p*(2^p)),(((2^p)+1):((2^p)+2*p))] <- obs.dist
40
41  vmat <- cbind(rep(0,(p*(2^p))), rep(1,(p*(2^p))),vmat)
42
43  hmat <- scdd(vmat,representation="V")$output
44  colnames(hmat)<- c("eq","inter","Y0=0,Y1=0", "Y0=1,Y1=0",
       "Y0=0,Y1=1", "Y0=1,Y1=1",
45                     "Y=0,X=0","Y=1,X=0", "Y=0,X=1","Y=1,X=1")
46
47  hmat.ineq <- hmat[hmat[,1]==0 & apply(hmat[,3:6], 1, sum)!=0 &
       apply(hmat[,7:10], 1, sum)!=0,]
48
49  p2<-as.matrix(p.empirical)
50
51  constr1.lp <- hmat.ineq[,3:6] %*% p1 + hmat.ineq[,7:10] %*%
       matrix(p2[1,], ncol=1) >= -hmat.ineq[,2]
52  constr2.lp <- hmat.ineq[,3:6] %*% p1 + hmat.ineq[,7:10] %*%
```

```
      matrix(p2[2,], ncol=1) >= -hmat.ineq[,2]
  constr.lp <- list(constr1.lp, constr2.lp, p1 >= 0, sum_entries(p1,
      axis=2) == 1)  # ACE lower bound ----
  obj <- p1[3] - p1[2]
  prob <- Problem(Minimize(obj), constr.lp)
  result <- solve(prob)
  #print(result$status)
  ACE.lb <- result$value
  # p.lb <- result$getValue(p)

  # ACE upper bound ----
  prob <- Problem(Maximize(obj), constr.lp)
  result <- solve(prob)
  #print(result$status)
  ACE.ub <- result$value
  # p.ub <- result$getValue(p)



  return(c(ACE.lb, ACE.ub))

}


round(get.naive.2x.2z.2y(crim_all[-3,-c(5,6)]),3)
round(get.naive.2x.2z.2y(crim_all[-2,-c(3,4)]),3)
round(get.naive.2x.2z.2y(crim_all[-1,-c(1,2)]),3)

round(get.naive.2x.2z.2y(crim_ac[-3,-c(5,6)]),3)
round(get.naive.2x.2z.2y(crim_ac[-2,-c(3,4)]),3)
round(get.naive.2x.2z.2y(crim_ac[-1,-c(1,2)]),3)

round(get.naive.2x.2z.2y(crim_nac[-3,-c(5,6)]),3)
round(get.naive.2x.2z.2y(crim_nac[-2,-c(3,4)]),3)
round(get.naive.2x.2z.2y(crim_nac[-1,-c(1,2)]),3)



### Confidence region

get.cr.2x.2z.2y<-function(data, alpha){
  n.z <- apply(data, 1, sum)
```

```
p.empirical <- data / n.z

# get critical value ----
t.alpha <- criticalValue(rep(4, 2), n.z, p=alpha, verbose = TRUE)
#  cat(sprintf("critical value = %f\n", t.alpha))

# make objective -----
p1 <- Variable(4) # this is on the joint of counterfactual
p2 <- Variable(2,4)

# first bound on KL divergence
KL <- sum(sum_entries(p.empirical * (log(p.empirical) - log(p2)),
   axis=1) * 2 * n.z)

# additional bound from Strassen's theorem using RCDD
p <- 2 # number of levels for x

# matrix to feed into rcdd
vmat <- matrix(nrow=(p*(2^p)),ncol=((2^p)+p*2))

# # counterfactual matrix:

factr.list <- list()
for(i in 1:p){
  factr.list[[i]] <- c(0,1)
}
factr.list[[p+1]] <- (1:p)


cntr.fact.names<-as.matrix(expand.grid(factr.list),nrow=p*(2^p))

cntr.facts <- matrix(0,nrow=(p*(2^p)),ncol=((2^p)+1))

cntr.facts[,1:(2^p)] <- matrix(rep(t(diag(1,2^p)), p), ncol = (2^p),
   byrow = TRUE)
cntr.facts[,(2^p)+1] <- rep((1:p),rep(2^p,p))

obs.dist <- matrix(0,nrow=(p*(2^p)),ncol=p*2)
for(i in 1:(p*(2^p))){
  xval <- cntr.fact.names[i,p+1]
  yval <- cntr.fact.names[i,xval]
  obs.dist[i,2*(xval-1)+yval+1] <- 1
}
```

```
135
136 vmat[1:(p*(2^p)),1:(2^p)] <- cntr.facts[1:(p*(2^p)),1:(2^p)]
137 vmat[1:(p*(2^p)),(((2^p)+1):((2^p)+2*p))] <- obs.dist
138
139 vmat <- cbind(rep(0,(p*(2^p))), rep(1,(p*(2^p))),vmat)
140
141 hmat <- scdd(vmat,representation="V")$output
142 colnames(hmat)<- c("eq","inter","Y0=0,Y1=0", "Y0=1,Y1=0",
       "Y0=0,Y1=1", "Y0=1,Y1=1",
143                    "Y=0,X=0","Y=1,X=0", "Y=0,X=1","Y=1,X=1")
144
145 hmat.ineq <- hmat[hmat[,1]==0 & apply(hmat[,3:6], 1, sum)!=0 &
       apply(hmat[,7:10], 1, sum)!=0,]
146
147 constr1 <- hmat.ineq[,3:6] %*% p1 + hmat.ineq[,7:10] %*% t(p2[1,])
       >= -hmat.ineq[,2]
148 constr2 <- hmat.ineq[,3:6] %*% p1 + hmat.ineq[,7:10] %*% t(p2[2,])
       >= -hmat.ineq[,2]
149 constr3 <- KL <= t.alpha
150 constr <- list(constr1, constr2, constr3, p1 >= 0, p2 >=0,
       sum_entries(p1, axis=2) == 1, sum_entries(p2, axis=1) == 1)
151
152 # ACE lower bound ----
153 obj <- p1[3] - p1[2]
154 prob <- Problem(Minimize(obj), constr)
155 result <- solve(prob)
156 #print(result$status)
157 ACE.lb <- result$value
158 # p.lb <- result$getValue(p)
159
160 # ACE upper bound ----
161 obj <- p1[3] - p1[2]
162 prob <- Problem(Maximize(obj), constr)
163 result <- solve(prob, solver="ECOS")
164 #  print(result$status)
165 ACE.ub <- result$value
166 # p.ub <- result$getValue(p)
167
168
169 prob.pval <- Problem(Minimize(KL), list(
170   constr1, constr2, p1 >= 0, p2 >=0, sum_entries(p1, axis=2) == 1,
         sum_entries(p2, axis=1) == 1
171 ))
```

```
  result.pval <- solve(prob.pval)
  #print(result.pval$status)
  KL.ml <- max(result.pval$value, 0)
  #p.mle <- result.pval$getValue(p1)
  p.value <- tailProbBound(KL.ml, rep(4, 2), n.z, verbose = TRUE)

  return(c(ACE.lb, ACE.ub, p.value))

}

round(get.cr.2x.2z.2y(crim_all[-3,-c(5,6)], 0.05),3)
round(get.cr.2x.2z.2y(crim_all[-2,-c(3,4)], 0.05),3)
round(get.cr.2x.2z.2y(crim_all[-1,-c(1,2)], 0.05),3)

round(get.cr.2x.2z.2y(crim_ac[-3,-c(5,6)], 0.05),3)
round(get.cr.2x.2z.2y(crim_ac[-2,-c(3,4)], 0.05),3)
round(get.cr.2x.2z.2y(crim_ac[-1,-c(1,2)], 0.05),3)

round(get.cr.2x.2z.2y(crim_nac[-3,-c(5,6)], 0.05),3)
round(get.cr.2x.2z.2y(crim_nac[-2,-c(3,4)], 0.05),3)
round(get.cr.2x.2z.2y(crim_nac[-1,-c(1,2)], 0.05),3)
```

VITA

Yilin Song was born in Binzhou, Shandong, China, in June 1998. She received her B.A from St. Olaf College, Northfield, MN, in 2020 in Mathematics with a concentration in Statistics and Data Science. She received her Ph.D. in Biotatistics from University of Washington, Seattle in 2025 under the supervision of Dr. Thomas Richardson and Dr. Gary Chan.