UNIVERSITY OF CALIFORNIA SAN DIEGO

Three essays in robust causal inference

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

in

Economics

by

Pietro Emilio Spini

Committee in charge:

Professor Yixiao Sun, Co-Chair
Professor Kaspar Wuthrich, Co-Chair
Professor Jeffrey Clemens
Professor James Hamilton
Professor Margaret E. Roberts

2022

The dissertation of Pietro Emilio Spini is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

To all of us who searched deeply within,

and found resilience

EPIGRAPH

Persistence is half the way

—*Unknown*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

My journey through the Ph.D. has been a long exercise in persistence. Perhaps the biggest gift I received from this journey has been the opportunity for transformative moments. Many different emotions well-up as I reflect on this transformation. Among them, I discover my heart being lightened by many instances of gratefulness.

With thousands of miles separating us, the time spent with my parents has been limited in length, yet never in depth. Each time I saw them, my mother's and my father's unconditional love has replenished my soul. Perhaps unknowingly, their long-lasting love has nurtured the true root of my resilience. I am thankful to Sergio for accepting me and loving me for who I am, and teaching me compassion towards myself and others. I will forever be a better person because I met him.

The silver lining of San Diego's expensive rental market is that there is no shortage of housemates. Even then, I was very lucky to call some of them my best friends. Gregoire and I initially bonded over our passion for European culinary preparations. A few years after, our friendship is more fulfilling and flavorful than any recipe. A big thank for all other friends who shared these years with me in San Diego: Laura, Margherita and Handa, Lorenzo and Nalini, Sohini and Tom, Deepika and Shashank, Dario, Gerard and my housemates Ryan, Tim and Jason. I hope to join you again for a laughter, let in be in San Diego or wherever you are.

I am indebted to my advisors, Yixiao Sun and Kaspar Wuthrich for their generous time and support. Yixiao's restless dedication to meticulous scientific inquiry is nothing short of admirable. I am deeply thankful for his mentorship. Kaspar was incredibly available, especially during the height of the job market. His words of advise and encouragement have been of great help for the most critical moments of the PhD journey. I will do my best to pay this favor forward to my students.

I thank Jim Hamilton and Jeff Clemens for the many suggestions and support throughout these years. There are many other folks at UCSD whose acts of kindness will not be forgotten: Julian, Tyler, Melissa, Giacomo, Itzik, Xinwei and Ying among many others.

Finally, a deep acknowledgement goes to my friend Alyssa. She has been an incredible friend and a true inspiration. She taught me that the path to heal and grow may be challenging but it is absolutely worth it.

Chapter 1 contains material being prepared for submission for academic publication in Economics journals. The dissertation author is the sole author of this material.

Chapter 2 contains material being prepared for submission for academic publication in Economics journals. The dissertation author is the sole author of this material.

Chapter 3 contains material being prepared for submission for academic publication in Economics journals. It is joint work with Julian Martinez Iriarte. The dissertation author is a primary author of this material.

VITA

2010-2013	B. S. in Applied Economics Università Bocconi
2014-2016	M. S. in Applied Economics Cornell University
2016-2022	Ph. D. in Economics University of California San Diego
2022-	Lecturer (Assistant Professor) University of Bristol

ABSTRACT OF THE DISSERTATION

Three essays in robust causal inference

by

Pietro Emilio Spini

Doctor of Philosophy in Economics

University of California San Diego, 2022

Professor Yixiao Sun, Co-Chair
Professor Kaspar Wuthrich, Co-Chair

Economics research often addresses questions with an implicit or explicit policy goal. When such a goal involves an active intervention, such as the assignment of a particular treatment variable to participants, the analysis of its effects requires the tools of causal inference. In such settings, the opportunity to use experimental or observational data to tease out policy parameters of interest requires a combination of statistical and causal assumptions. In reduced form work, where an explicit economic theory is not laid out to allow identification of policy parameters from data, the investigation of the causal assumptions becomes a critical exercise for the credibility of the results. Many robustness exercises evaluate the effect that relaxing and/or modifying assumptions produces on the results of the study. The scope of these exercises is very broad, reflecting the need to tailor specific robustness exercises to whichever assumptions are most likely to be violated in a

given domain. This dissertation is a collection of three essays on robust causal inference that share a unifying theme: preserving the nonparametric nature of the robustness exercise. This aspect has both a theoretical and practical relevance. First, causal assumptions are usually nonparametric: robustness exercises that restrict to parametric cases might lead to misleading insights. Further, economics research has started to incorporate more flexible nonparametric and semi-parametric techniques which may call for robustness exercises that are readily applicable to these approaches.

Because robustness exercises are context specific, each of these essays addresses a separate aspect of it. Chapter 1 investigates how changes in the distribution of covariates may invalidate given experimental results, with implications for evidence based policy-making. It proposes an explicit metric of robustness that measures the distance of the closest distribution of covariates for which experimental results are violated. Chapter 2 analyses the practice of robustness checks as a way to validate a researcher's identification strategy. It details out the limitations of these exercises in detecting failure of identification and proposes a non-parametric robustness test that bypasses functional form assumptions. Finally, Chapter 3 focuses on the robustness of Marginal Treatment Effect identification when the instrumental variables fail to incentivize treatment for a subset of the population. It provides two alternative identification results which can be relevant in practice.

Chapter 1

Robustness, Heterogeneous Treatment Effects, and Covariate Shifts

1.1 Introduction

The guiding principle of evidence-based policy-making is to use experimental and (quasi)-experimental studies to guide the adoption of policies in various settings. This approach rests on the premise that the (quasi)-experimental findings are sufficiently robust and generalizable to hold beyond the setting of the (quasi)-experiment. In practice, this premise does not always hold: there are several examples of policies that, when adopted in non-experimental settings, under-performed their own experimental estimates Cartwright and Hardie [2012], Deaton [2010], Williams [2020]. In this paper, I argue that experimental estimates are insufficient to guide policy adoption and should be complemented by a measure of robustness that accounts for how policy recipients differ from the experimental ones. I develop a robustness metric, given by a scalar δ^*, that quantifies how much the

characteristics of the recipients would have to change in order to invalidate the (quasi)-experimental findings. My metric summarizes the *out-of-sample* uncertainty[1] that the policy-maker faces regarding the policy recipients' characteristics. As such, my metric complements traditional summaries of *in-sample* uncertainty, like the standard errors, which routinely accompany (quasi)-experimental estimates.

As a motivating example, consider a policy-maker who must decide whether to offer medical insurance coverage to low-income households. The policy-maker has access to the experimental estimates of Finkelstein et al. [2012] which suggest that a similar intervention led to higher health-care utilization and reduced financial strain in Oregon. The target population of insurance recipients could differ from the experimental one in Oregon along important dimensions. Our goal is to quantify how robust the experimental findings would be if relevant characteristics of the recipients are allowed to change. In this paper, I provide a solution to this problem by leveraging the policy effect heterogeneity in the experiment.

When policy effects are heterogeneous across sub-populations with different covariate values, (quasi)-experimental findings are generally not robust to changes in the distribution of the covariates. In such cases, even small changes in the distribution of the covariates could lead to significant aggregate changes in the policy effects. For example, in the Oregon experiment, subsidized health insurance could benefit sicker patients more than healthier patients. Then, the proportion of recipients with a given pre-existing health status, health habits, and/or co-morbidities may strongly influence the overall effect of the policy. Usually, these types of covariates are exclusively collected in the experimental context and not all of them are accessible in the new policy prior to implementation. As a result, the procedures proposed by Hsu et al. [2020] and Hartman [2020] that re-weight sup-population effects by the new environment's entire set of covariates are generally not feasible. Moreover, the

[1]Quantifying other sources of *out-of-sample* uncertainty has been a central theme in the recent econometric literature including Andrews et al. [2017] for moment conditions, Altonji et al. [2005], Cinelli and Hazlett [2020], Oster [2019] for confounding factors, and the break-down approaches in Horowitz and Manski [1995], Masten and Poirier [2020].

heterogeneity of policy effects across sub-populations with different covariates values can be hard to model. This is because while domain knowledge can help select covariates that are predictive of the heterogeneity of policy effects, it usually cannot pin down a specific functional form for this heterogeneity. Because this heterogeneity links covariate shifts to shifts in the magnitudes of the aggregate policy effects, a general approach to robustness must reflect the uncertainty regarding the heterogeneity's functional form.

My robustness metric avoids the need to specify a functional form for the policy effect heterogeneity, letting it instead be flexibly estimated through the (quasi)-experimental data. Many popular existing approaches to robustness, like Altonji et al. [2005], Oster [2019] and Cinelli and Hazlett [2020], take advantage of specific functional forms. When designing a robustness metric for distributional changes, relying on functional form assumptions carries important implications for what type of shifts the metric can detect. If the way we measure a shift does not match the way we model heterogeneity, the resulting measure of robustness may be misleading. Consider, for example, measuring the difference between an arbitrary covariate distribution and the (quasi)-experimental one by reporting the difference in their means. With an unrestricted form for the heterogeneity of policy effects, we can, in general, construct a mean-preserving shift of the covariates' distribution which invalidates the policy-maker's claim. For example, in the Oregon experiment, if higher income recipients have negative effects while lower-income recipients have positive effects, we could construct a mean-preserving spread of the income distribution that induces a negative effect overall. Since their means coincide, such a distribution will have a distance of zero from the experimental covariates. A metric that, in most cases, is equal to zero cannot be very informative for assessing the robustness of (quasi)-experimental findings. This example suggests that a robustness metric should be general enough to accommodate unknown forms of policy effect heterogeneity. My robustness metric allows for arbitrary forms of policy effects heterogeneity, avoiding the limitations of a parametric model. Despite

its generality, my metric is still easy to construct and interpret: a one-number summary of heterogeneity which only depends on (quasi)-experimental data.

Measuring robustness to covariate shifts requires choosing a distance between an arbitrary distribution of the covariates and the (quasi)-experimental one. In my approach, I adopt Kullback-Leibler divergence distance (KL distance). The KL distance is a popular choice for sensitivity analysis exercises, appearing recently in Christensen and Connault [2019] who apply it to models defined by moment inequalities and Ho [2020] who uses it in a Bayesian context. It has several advantages in our context. First, it is invariant to smooth invertible transformations of the covariates, hence independent of the covariates' units. Second, it provides a closed-form expression for the proposed global robustness measure, while other popular robustness approaches, like Broderick et al. [2020] rely on local approximations. Leveraging the closed-form solution, I cast estimation of my robustness metric as a GMM problem where the moment equation depends on two components. The first is the observed covariate distribution. The second is a functional parameter capturing the heterogeneity of policy effects, which can be flexibly estimated in the (quasi)-experimental data.

The heterogeneity of policy effects is often sparse: out of the rich set of covariates available in the (quasi)-experiment, just a few are needed to approximate it well. When covariate data is even moderately high-dimensional, it can be hard to select which covariates are important *ex-ante*. Machine-learning estimators, like lasso, random forest and boosting, can exploit the sparsity to automatically select the key covariates, reducing the need for *ad-hoc* procedures. Using machine-learning to estimate policy effect heterogeneity is appealing, but it may result in substantial bias in the estimated robustness metric δ^*, due to regularization and/or model selection. To accommodate machine-learning methods, I construct a de-biased GMM estimator: I derive the nonparametric influence function correction for the GMM parameters and leverage the theory in Chernozhukov et al. [2020]

to eliminate the first-order bias from first-step estimators. I show that my metric δ^* can be consistently estimated at $\sqrt{n}$-rate under mild conditions on the first-step estimators of the policy effect heterogeneity. Under these conditions the functional parameter that summarizes heterogeneity can be estimated through modern high-dimensional methods like lasso, random forest, boosting and neural nets.

I apply my robustness procedure to study the Oregon health insurance experiment, whose findings have profound implications for public health Sanger-Katz [2014]. I replicate results in Finkelstein et al. [2012] and compute the robustness measure for several outcomes capturing recipients' heath-care utilization and financial strain. As discussed in Finkelstein et al. [2012] and Finkelstein [2013], the Oregon lottery recipients are older, in worse health, and feature a higher proportion of white individuals compared to the national average. These features invite questions about the robustness of the Oregon experiment's findings and the possibility of using them for policy adoption in other states. The differences in magnitude and sign between the effects of Medicaid expansion in Oregon and Massachusetts have motivated an effort to reconcile the discrepancy by identifying different populations of beneficiaries in the two states Kowalski [2018]. My robustness exercise is complementary to Kowalski [2018]: I compute the smallest change in the distribution of the key covariates relative to the Oregon benchmark, that can eliminate the positive effect of the lottery on recipients' health-care utilization and financial strain outcome measures. I find that the increase in outpatients visits is the most robust outcome among the measures of health-care utilization and financial strain.

This paper is also related to a larger strand of the econometric and statistics literature on robustness and sensitivity analysis originally initiated by Tukey [1960] and Huber [1965]. Recently, there are many other important but distinct robustness approaches: geared towards external validity Meager [2019], Gechter [2015], robustness to dropping a percentage of the sample Broderick et al. [2020], by looking at sub-populations Jeong

and Namkoong [2020], or with respect to unobservable distributions like in Christensen and Connault [2019], Armstrong and Kolesár [2021], Bonhomme and Weidner [2018], and Antoine and Dovonon [2020]. My contribution complements this tool-set by giving the policy-maker an explicit measure of robustness to shifts in the covariate distributions. There are two reasons to focus on observable characteristics. First, observable characteristics are readily available to the policy-maker and are likely to be of first-level importance when assessing the robustness of (quasi)-experimental findings. Second, the resulting robustness metric is identified through the (quasi)-experimental data, limiting the need for bounding or partial identification approaches.

The paper is organized as follows: Section 1.2 introduces the basic setting and the notion of robustness to changes in the covariate distribution. Section 1.3 presents the main estimator and its asymptotic properties using the de-biased GMM theory recently developed in Chernozhukov et al. [2020]. Section 1.4 applies the proposed robustness metric to the Oregon health insurance experiment and reports empirical findings. Section 1.5 briefly concludes. In the Appendix, I provide all the proofs and discuss multiple extensions.

1.2 A robustness metric for covariate shifts

In this section, I use the potential outcome framework to explicitly link the heterogeneity of policy effects to the notion of robustness outlined in the introduction. The discussion focuses on the average treatment effect (ATE) as the main aggregate policy effect of interest. The policy-maker wants to assess the robustness of a claim on the magnitude (and/or sign) of the ATE, of the form $ATE \geq \tilde{\tau}$. The claim is true in the (quasi)-experiment but may no longer be true if covariates changes too much. The idea is to take advantage of the Conditional Average Treatment Effect (CATE), a functional parameter which links sub-population level treatment effects with the ATE. I use CATE to characterize, among

the distributions that invalidate the policy-maker's claim ($ATE \geq \tau$), the one that is closest to the distribution of covariates in the (quasi)-experiment. I label this distribution the *least favorable distribution* because, among the distributions that invalidate the policy-maker's claim it is the hardest to distinguish from the covariates in the (quasi)-experiment. To measure the distance between two covariate distributions I use the Kullback-Leibler divergence distance. The value of the KL distance between the *least favorable distribution* and the (quasi)-experimental covariates will be the proposed robustness metric δ^*. Any covariate distribution that is closer than δ^* from the (quasi)-experimental covariates will be guaranteed to satisfy the policy-maker's claim ($ATE \geq \tilde{\tau}$).

1.2.1 Notation and Set Up

The policy-maker observes an outcome of interest $Y \in \mathcal{Y}$, a set of covariate measurements $X \in \mathcal{X}$ and a treatment status $D \in \{0,1\}$. I consider two sets of covariates. The first set includes covariates which are exclusively collected in the (quasi)-experimental data and for which no counterpart exists in census data. For example, in the Oregon health insurance experiment, the recipients' health status and previous health history is available through survey data but such information may not be accessible through census variables in other settings (perhaps other states). The second set includes covariates for which a counterpart exists in the census data in other states, for example participants' race and age. To reflect the division of these two covariate types, X could be partitioned into two sets: $X = X_c \cup X_e$ denoting *census covariates* and *(quasi)-experiment specific covariates* respectively. All variables in X will be used to estimate the treatment effect heterogeneity in the (quasi)-experiment, which is the functional parameter needed to compute the robustness metric. The details are introduced in Section 1.2.3. If the policy-maker had access to observations on X_c in both the (quasi)-experiment and in the setting where the policy is to be adopted, my robustness metric can be modified to account for this additional

information. To lighten the notation, in the main text I consider $X = X_e$ and discuss how to include X_c in the Appendix.

Now I introduce the notation to discuss changes in the distribution of the covariates. I use F_X to denote the distribution of the covariates in the (quasi)-experiment and and P_X to denote its associated probability measure. The propensity score is defined as $\pi(x) = P_X(D = 1|X = x)$. Following the traditional potential outcome framework, I denote Y_d for $d = \{0, 1\}$, the potential outcomes under treated and control status when the distribution of the covariates follows F_X. For example, in the Oregon experiment, Y_1 may represent the financial strain of a recipient if they receive insurance coverage while Y_0 represents the financial strain of the same recipient if they do not receive insurance coverage. In principle the distribution of the potential outcomes depends on the distribution of the covariates. To reflect this, I use Y_d and Y'_d to denote the potential outcomes when the distribution of the covariates follows F_X and F'_X respectively. Finally, for any random variable W, $\mathcal{W}$ denotes its support.

The parameter of interest for the policy-maker is the $ATE := \mathbb{E}[Y_1 - Y_0]$. The Conditional Average Treatment Effect (CATE) defined by $\tau(x) := CATE(x) = \mathbb{E}[Y_1 - Y_0|X = x]$ captures how the average treatment effect changes across sub-populations with covariate value $X = x$. Under unconfounded-ness (Assumption 1 i) below), $\tau(x)$ is nonparametrically identified[2] by $\mathbb{E}[Y|D = 1, X = x] - \mathbb{E}[Y|D = 0, X = x]$ in the (quasi)-experiment Imbens and Rubin [2015].

Assumption 1. *Unconfounded-ness & Overlap*

i) $Y_1, Y_0 \perp\!\!\!\perp D|X$.

ii) For all $x \in \mathcal{X}$ *we have* $0 < \epsilon \leq \pi(x) \leq 1 - \epsilon < 1$

In the case of a randomized control trial, for example when treatment assignment is

[2]If the CATE only partially identified, like in the case on non-compliance based on unobservables, it is possible to follow a bounding approach for my robustness procedure. I leave this interesting case for future research.

completely randomized or is randomized conditional on covariates, Assumption 1 holds by design. In the case of (quasi)-experimental studies Assumption 1 i) requires the researcher to carefully evaluate the selection mechanism that governs program participation. Assumption 1 ii) is strict overlap. While strict overlap is not a necessary condition for identification, it will be important in the estimation of the robustness metric in Section 1.3.

In this paper, the goal is to study the robustness of claims concerning the ATE with respect to changes in the distribution of the covariates. Because the ATE is obtained by averaging $\tau(x)$ with weights proportional to F_X we have the following map between the covariate distributions and the ATE:

$$ATE : F_X \mapsto \int_{\mathcal{X}} \tau_{F_X}(x) dF_X(x) \tag{1.1}$$

The subscript F_X on $\tau(x)$ indicates that, in general, it's possible that the functional form of CATE depends on F_X. In this case, a change in the distribution of the covariates would effect the magnitude of ATE through two channels: a direct effect thorough the weights of F_X and an indirect effect through changing the functional form of $\tau_{F_X}(x)$. In this paper, I introduce the covariate shift assumption[3] to eliminate the indirect effect.

Assumption 2. ***(Covariate Shift)*** *Let X' denote the covariates in the new environment. Then:*

i $F_{Y'_d|X'}(y|x) = F_{Y_d|X}(y|x)$ *for* $d = \{0,1\}$, *for all* $x \in \mathcal{X}$ *and* $y \in \mathcal{Y}_d$ *and all distributions of* X'.

ii $\mathcal{X}' \subseteq \mathcal{X}$

Assumption 2 i) says that the causal link between the treatment variable D and the potential outcomes of interest Y_1 and Y_0 does not depend on the distribution of the observables. One could think of Assumption 2 as analogous to a policy invariance condition

[3]This assumption appears, for example also in Hsu et al. [2020] and Jeong and Namkoong [2020].

where the invariance in this case is with respect to the distribution of covariates.

Assumption 2 ii) says the support of the covariates in the new environments is contained in the support of the baseline environment. In practice, this limits the extrapolation to environments for which any value of the covariates could have been observed in the (quasi)-experimental setting as well. Because Assumption 2 guarantees that $\tau_{F_X}(x)$, the CATE, does not vary when F_X is replaced by any other distribution $F_{X'}$ it is not necessary to index $\tau(x)$ with F_X.[4] Then, the link between F_X and ATE reduces to integration against a fixed $\tau(x)$:

$$ATE : F_X \mapsto \int_{\mathcal{X}} \tau(x) dF_X(x) \tag{1.2}$$

To emphasize the dependence of the ATE on an arbitrary distribution of the covariates F_X, I occasionally write $ATE(F_X)$. Before presenting the general framework I give perhaps the simplest nontrivial example of a robustness exercise with respect to the distribution of the covariates.

Example 1. *Consider a binary covariate $X = \{0,1\}$. D is randomly assigned, trivially satisfying Assumption 1. By unconfounded-ness, $\mathbb{E}[Y_1|x=0], \mathbb{E}[Y_0|x=0], \mathbb{E}[Y_1|x=1], \mathbb{E}[Y_0|x=1]$ can all be identified. Consequently, the average treatment effect for the sub-populations $x=0$ and $x=1$, denoted $\tau(0)$ and $\tau(1)$ are also identified. Because X is Bernoulli, any distribution on $\{0,1\}$ is fully described by $P_X(x=1)=p_1$ so automatically $P_X(x=0)=1-p_1$.*

[4] This could be cast as an identification result which follows immediately from the Assumption 2. See Hsu et al. [2020], Lemma 2.1.

Suppose that, in the experiment $ATE \geq 0$. Note that:

$$\begin{aligned} ATE(F_X) &= \mathbb{E}[Y_1|x=0]\cdot(1-p_1)+\mathbb{E}[Y_1|x=1]\cdot p_1 \\ &\quad -\mathbb{E}[Y_0|x=0]\cdot(1-p_1)-\mathbb{E}[Y_0|x=1]\cdot p_1 \\ &= (\mathbb{E}[Y_1|x=0]-\mathbb{E}[Y_0|x=0])\cdot(1-p_1)+(\mathbb{E}[Y_1|x=1]-\mathbb{E}[Y_0|x=1])\cdot p_1 \\ &= \tau(0)\cdot(1-p_1)+\tau(1)\cdot p_1. \end{aligned}$$

A shift in the covariate distribution is simply a shift in the parameter p_1. Assume the treatment effects are sufficiently heterogeneous, namely $\tau(1) > 0 > \tau(0)$ so one group has positive effects from treatment and the other group has negative effects. What is the closest covariate distribution that invalidates the claim $ATE \geq 0$?

It suffices to find the weights on $x=0, x=1$ such that the ATE is 0. Expressing it in terms of p_1:

$$\tau(0)\cdot(1-p_1^*)+\tau(1)\cdot p_1^* = 0$$

A solution is given by:

$$p_1^* = \frac{-\tau(0)}{\tau(1)-\tau(0)} \in [0,1]$$

so the distance $|p_1^* - p_1| = |\frac{-\tau(0)}{\tau(1)-\tau(0)} - p_1|$ is largest shift in the covariates that still guarantees that the claim $ATE \geq 0$ holds.

Under what conditions we are always guaranteed to find a solution like p_1^* above? Is it unique? Can we always characterize the distance between p_1^* and p_1? If the space $\mathcal{X}$ is not discrete, a probability distribution on $\mathcal{X}$ cannot be described by a finite dimensional parameter without restricting the class of probability distributions on $\mathcal{X}$. How should one measure the distance between two distributions in general?

I start from this last question by introducing a notion of distance that does not

require any parametric restriction on probability distributions.[5] Here I introduce the KL-divergence distance:

Definition 2 (KL-divergence)**.** *Consider the KL-divergence between two distributions F_X and F'_X given by:*

$$D_{KL}(F'_X||F_X) := \int_{\mathcal{X}} \log\left(\frac{dF'_X}{dF_X}(x)\right) \frac{dF'_X}{dF_X}(x) dF_X(x) \tag{1.3}$$

where $\frac{dF'_X}{dF_X}$ is the Radon-Nikodym derivative of the distribution F'_X with respect to the experimental distribution F_X, provided that $P'_X \ll P_X$ for the respective probability measures.

There are several advantages to using the KL divergence to measure the distance between probability distributions: it is nonparametric, it has useful invariance properties and it delivers a closed form solution for the policy-maker's robustness problem introduced below. Both Ho [2020] and Christensen and Connault [2019] use the KL divergence to measure the distance between probability distributions in different contexts. Appendix 1.H discusses in detail how to use convex analysis to obtain a closed form solution for the policy-maker's robustness problem.

1.2.2 The policy-maker's problem: quantifying robustness

After isolating the link between the ATE and the distribution of covariates and choosing a distance measure between probability distributions, we can formalize the policy-maker's robustness problem. Consider the claim given by $ATE \geq \tilde{\tau}$: the ATE is larger than a desired threshold $\tilde{\tau}$. The sign of the inequality is without loss of generality, as claims of the type $ATE \leq \tilde{\tau}$ can be accommodated with an equivalent treatment. The threshold $\tilde{\tau}$ captures a minimal desirable aggregate effect that would make the intervention viable for the policy-maker. It could capture the average cost for the roll-out of the intervention or

[5] I discuss the details of parametric classes in Appendix 1.B, as special cases of the general procedure.

the value of ATE for a competing policy. In Example 1, $\tilde{\tau}$ was fixed at 0. The policy-maker is interested in the smallest shift from the (quasi)-experimental distribution, F_X, such that the claim $ATE \geq \tilde{\tau}$ is invalidated. Recall $\tau(x) = CATE(x)$. Formally the policy-maker wants to solve the following problem:

$$\inf_{dF'_X:\ dF'_X \ll dF_X; dF'_X(\mathcal{X})=1} D_{KL}(F'_X||F_X) \tag{1.4}$$

$$s.t.\ \int_{\mathcal{X}} \tau(x) dF'_X(x) \leq \tilde{\tau} \tag{1.5}$$

The optimization problem in Equation (1.4) searches across all distributions of the covariates that invalidate the policy-maker's claim $ATE \geq \tilde{\tau}$ (notice that the ATE for all the distributions in Equation (1.5) is constrained to be less than $\tilde{\tau}$) and selects, if they exist, the one(s) that are closest to the (quasi)-experimental distribution F_X, according to the KL distance in Equation (1.4). Notice also that $\tau(x)$ in Equation (1.5) is not indexed by F'_X because of the covariate shift assumption (Assumption 2). Here, the class of probability measures for the covariates is restricted to be absolutely continuous w.r.t the (quasi)-experimental measure dF_X[6] but no other restriction is imposed: the class of distributions is still nonparametric. Absolute continuity does restrict the distributions F'_X to be supported on $\mathcal{X}$. While it may appear as an unnecessary restriction, I view it as a very reasonable requirement: the feasible distributions in Equation (1.5) cannot put mass on a sub-population $X = x$ that could not theoretically be observed in the (quasi)-experimental setting. Clearly, treatment effect values for sub-populations with $X = x$ that can never be observed can lead to arbitrarily large average effects and the robustness exercise would not be very informative. We are now ready to define the *least favorable distribution* and the robustness metric.

[6]This is a refinement of Assumption 1. Namely, with a slight abuse of notation, requiring for instance that $dF_X, dF'_X \ll \lambda$ will deliver absolute continuity of dF'_X w.r.t dF_X. Restricting the support guarantees that dF'_X cannot put mass on areas where dF_X does not put mass.

Definition 3. *i) The least favorable distribution set* $\{F_X^*\}$ *is given by the expression below:*

$$\{F_X^*\} = \arg \min_{P'_X:\ P'_X \ll P_X; P'_X(\mathcal{X})=1} D_{KL}(F'_X||F_X) \tag{1.6}$$
$$s.t.\ \int_{\mathcal{X}} \tau(x) dF'_X(x) \leq \tilde{\tau}$$

where the set in Equation (1.6) *is allowed to be the empty set.*

ii) For a given $\tilde{\tau} \in \mathbb{R}$ *the robustness metric* $\delta^*(\tilde{\tau})$ *is given by:*

$$\delta^*(\tilde{\tau}) = D_{KL}(F_X^*||F_X). \tag{1.7}$$

The minimizer of Equation (1.4) is the *least favorable distribution*, the closest distribution of the covariates that invalidates the target claim. I define the KL-distance between the experimental distribution and the *least favorable distribution* as my metric $\delta^*(\tilde{\tau})$ which quantifies the robustness of the claim $ATE \geq \tilde{\tau}$. Observe that, if the (quasi)-experimental ATE satisfies the constraint in Equation (1.5), then we can always choose the *least favorable distribution* to be the (quasi)-experimental one, namely $F_X^* = F_X$ since it's feasible and $D_{KL}(F_X^*||F_X) = 0$. In words this means that the policy-maker's claim is already invalidated in the (quasi)-experiment. The problem is non-trivial when the $ATE(F_X) > \tilde{\tau}$ condition is satisfied for the (quasi)-experimental distribution F_X. In such a case, the (quasi)-experimental distribution F_X is excluded from the feasible set of Equation (1.5). As a result, the value of $D_{KL}(F_X^*||F_X)$ in Equation (1.4) must be strictly positive. Notice that, in Example 1, we imposed the requirement that the $ATE(p_1)$ in the experiment was larger than 0, to guarantee that the problem was indeed non-trivial.

If $\mathcal{X}$ is a set containing finitely many elements, the covariate distribution is discrete. In practice, there are many empirical applications in which covariates of interest are either discrete or have been discretized for privacy reasons. Any grouping of a continuous variables

in finitely many classes, gives rise to discrete distribution. For example, in the Oregon experiment, the recipients income may have been discretized into income groups. When the covariates space is discrete, we can get an important geometric insight in the structure of the robustness problem as formulated by Equations (1.4) and (1.5). The example below illustrates the case where $\mathcal{X}$ contains only 3 points. In this case, a probability distribution on $\mathcal{X}$ can be parametrized by 2 parameters and there is convenient visual representation of the robustness problem contained in Equations (1.4) and (1.5).

Example 4. *Consider the case* $\mathcal{X} = \{x_1, x_2, x_3\}$ *each value representing an income bin: high, medium and low respectively. Here the experimental distribution is represented by a triplet* (p_1, p_2, p_3)*. Because* $p_1 + p_2 + p_3 = 1$ *the whole space of probability distributions on* $\mathcal{X}$ *is 2-dimensional: it suffices to choose* p_1 *and* p_2 *to fully characterize a distribution. Suppose that conditional treatment effects are highest for lower income participants and are lowest for high income participants:* $\tau(x_1) = 1, \tau(x_2) = 2, \tau(x_3) = 3$*. The average cost of roll-out is equal to* $\tilde{\tau} = 1.8$*. The claim is* $ATE \geq \tilde{\tau}$ *meaning that the ATE should be higher than average cost. In the experiment ATE is equal to* $2.4 > 1.8$ *which satisfies the claim.*

The policy-maker's robustness problem in Example 4 is depicted in Figure 1.1. Since the functions in Equations (1.4) and (1.5) are differentiable in p_1 and p_2 the finite dimensional problem could be easily solved through the standard Karush-Kuhn-Tucker conditions. The level sets of the KL distance, the feasible set and the *least favorable distribution* are all indicated in Figure 1.1. The KL level set associated to $\delta^*(\tilde{\tau})$ is highlighted by a green contour. It includes the set of covariate distributions that are guaranteed to satisfy the policy-maker's claim. This region is conservative, in the sense that there exist covariate distributions that satisfy the policy-maker's claim but fall outside of the green contour. This feature reflects the definition of robustness as a minimization problem in Equations (1.4) and (1.5).

When $\mathcal{X}$ is not discrete, a representation like Figure 1.1 may not be possible.

Nonetheless one can still show that, given some conditions, a solution for F_X^* like the one in Figure 1.1 always exists, is unique, and can be characterized by a closed form expression, with virtually little difference from the finite dimensional case. This result also guarantees that the robustness metric $\delta^*(\tilde{\tau})$ is well defined for a wide range of $\tilde{\tau}$ values.

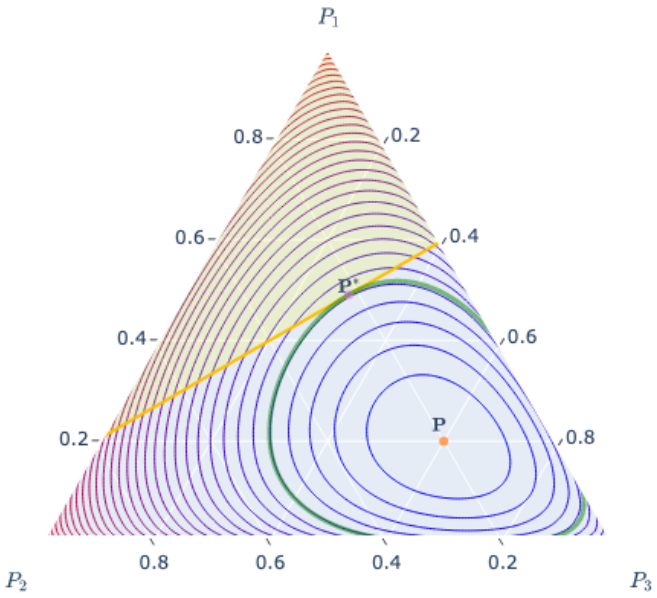


Figure 1.1: The triangle represents the collection of all arbitrary probability distribution triplets (p_1, p_2, p_3) on the discrete set (x_1, x_2, x_3) represented in barycentric coordinates. P denotes the experimental distribution, given by $(0.2, 0.2, 0.6)$. The $CATE(x_1, x_2, x_3) = (1, 2, 3)$ so the conditional treatment effect is greater in the highest income group. The yellow shaded region is the feasible set: the collection of covariate distributions with an $ATE \leq 1.8$, which invalidate the policy-maker's claim. The solid yellow line is the boundary of the feasible set. The contour lines from blue to red represent the level sets of the KL distance of any distribution in the triangle with respect to the experimental distribution P (bluer indicates a lower value for the KL divergence). The distribution $P^* = (0.491, 0.218, 0.291)$ is the *least favorable distribution*. It is the minimizer of the KL divergence, subject to the feasibility constraint (it lies on the orange line). The green boundary is the level set of KL that corresponds to $\delta^* \approx 0.296$. Any distribution closer than δ^*, within the green boundary is guaranteed to satisfy the policy-maker's claim.

1.2.3 A closed form solution for quantifying robustness

In this section I characterizes the solution for the policy-maker's robustness problem in Equations (1.4) and (1.5) in the general case. Some additional conditions are introduced below.

Assumption 3 (Bounded-ness)**.** *The conditional average treatment effect $\tau(X)$ is bounded P_X-almost surely over $\mathcal{X}$. In particular for some $M \in \mathbb{R}_+$ we have:*

$$\mathbb{P}_X\left(|\tau(X)| \leq M\right) = 1$$

Incidentally, for any covariate probability measure that is absolutely continuous w.r.t P_X, Assumptions 3 continues to hold. This is because $P_{X'}$ cannot put mass on the subsets of X that P_X considers negligible, which includes the subset of X where $\tau(x)$ is unbounded. Assumption 3 is automatically satisfied if $\tau(X)$ is bounded on $\mathcal{X}$. Bounded-ness is not very restrictive in a micro-econometrics framework where virtually all variables are bounded in the cross-section.

Consider the feasible set in Equation (1.5). While the set is guaranteed to be convex, it may be empty. If that is the case, the value of the minimization problem in Equation (1.4) is $+\infty$. I avoid such cases by guaranteeing that, for a given claim, an $ATE = \tilde{\tau}$ is attainable, for some distribution F'_X. This amounts to assuming that there is enough variation in $\tau(x)$ to induce an ATE of $\tilde{\tau}$ through changes in the distribution of the covariates. An extreme case where such requirement fails is described below.

Example 5 (Homogeneous treatment effects)**.** *Consider a situation of constant treatment effects. In this case $ATE(F_X) = \int_{\mathcal{X}} c \; dF_X = c$ so that the ATE is equal to c regardless of the distribution of the covariates.*

Not surprisingly, no heterogeneity in treatment effects translates in no threat to robustness. One can freely extrapolate the claim from the (quasi)-experimental environment

to any other environment. Constant treatment effects are a rather extreme case. A more realistic example concerns whether the minimal desired magnitude $\tilde{\tau}$ is outside of the range of variation of the heterogenous treatment effects. For example, suppose that $2 \leq \tau(x) \leq 5$ with probability equal to 1. Then, choosing $\tilde{\tau} = 1$ results in an empty feasible set of distributions, since no probability distribution may ever integrate against $\tau(x)$ to an ATE of 1. In this case, since the set of distributions in Equation (1.5) is empty, the infimum in Equation (1.4) evaluates to $+\infty$. So we see that enough heterogeneity of treatment effects is a necessary condition for robustness to be non-trivial. For estimation purposes it is convenient to consider a parameter space for the robustness measure that is a subset of $\mathbb{R}$ rather than $\mathbb{R} \cup \{+\infty\}$. The following assumption guarantees that the feasible set is not empty:

Assumption 4. *(Non-emptiness) Denote the interior S° of a set S to be the union of all open sets $O \subseteq S$. Let $L : F_X \to \int_{\mathcal{X}} \tau(x) dF_X(x)$ be the linear map defined on the set of probability distributions on $\mathcal{X}$ that are absolutely continuous w.r.t P_X, denoted as $\mathcal{P}_X \subset \mathcal{M}$. We require $\tilde{\tau} \in L^{\circ}(\mathcal{P}_X)$, that $\tilde{\tau}$ is in the interior of the range of L.*

Assumption 4 says that $\tilde{\tau}$ is in the interior of the range of the linear map L. In other words, there is enough observable heterogeneity in treatment effects that there exists a distribution of covariates that, when integrated against $\tau(x)$, it induces an $ATE = \tilde{\tau}$. Contrast this to the homogeneous treatment effect case in Example 5, where Assumption 3 fails. There, $L^{\circ}(\mathcal{P}_X) = \emptyset$. More generally, the length of $L(\mathcal{P}_X)$ measures how rich is the set of ATEs that could be produced by choosing an arbitrary distribution F_X. Assumption 4 is testable. For a given value for $\tilde{\tau}$, one could obtain an estimate of the $\tau(x)$ and test whether $\tilde{\tau}$ is smaller than $\sup_x \tau(x)$ or greater than $\inf_x \tau(x)$, depending on the sign of the claim of interest, using the procedure in Chernozhukov et al. [2013]. Testing Assumption 4 tests for whether treatment effects are sufficiently heterogeneous to invalidate the claim of interest through a covariate shift, which is more general than testing whether any form of

treatment effect heterogeneity is present. This is because, along the lines of the discussion above, treatment effects can indeed be heterogeneous but not heterogeneous enough to invalidate the policy-maker's claim. A rejection in the test means implies an infinite value for the robustness metric and signals that the policy-maker's claim can never be invalidated by covariates shifts.

Remark 6. *The interior condition cannot be relaxed. By Assumption 3, the image of $\mathcal{P}_X$ under L is a compact convex subset of $\mathbb{R}$, that is, an interval. If $\tilde{\tau}$ is at a an endpoint of this interval, the feasible set in Equation* (1.5) *may consist of only a point mass measure Because such a covariate measure is not absolutely continuous w.r.t. P_X, the feasible set is again empty and will necessarily result in an infinite value for the KL-divergence in Equation* (1.4)*.*

In Example 1 we imposed the condition $ATE(1) = \tau(0) < 0$ to guarantee that the problem has a solution. In the context of Example 1, $L(\mathcal{P}_X) = [\tau(0), \tau(1)]$, the image of L is the interval between the conditional average treatment effects at $x = 0$ and $x = 1$ since any $ATE(p)$ is a weighted average of $\tau(0)$ and $\tau(1)$. By requiring that $\tau(0) < 0 < \tau(1)$, $\tilde{\tau} = 0 \in L^{\circ}(\mathcal{P}_X)$ hence satisfies Assumption 4.

With Assumptions 3 and 4 we are now ready to introduce the key result that always delivers a closed form solution for the robustness metric. It says that the *least favorable distribution* set in Definition 3 is nonempty and it contains a unique distribution (P_X-almost everywhere). Moreover the robustness metric $\delta^*(\tilde{\tau})$ is finite and both it and the *least favorable distribution* have a closed form solution:

Lemma 7 (Closed form solution)**.** *Let Assumptions 1, 2, 3 and 4 hold. Then: i) The infimum in Equation* (1.4) *is achieved. Moreover F_X^*, is characterized, P_X-almost everywhere, by:*

$$\frac{dF_X^*}{dF_X}(x) = \frac{\exp(-\lambda(\tau(x) - \tilde{\tau}))}{\int_{\mathcal{X}} \exp(-\lambda(\tau(x) - \tilde{\tau}))dF_X(x)} \tag{1.8}$$

where $\frac{dF_X^*}{dF_X}$ *is the Radon-Nikodym derivative of* dF_X^* *with respect to* dF_X *and* λ *is the Lagrange multiplier implicitly defined by the equation:*

$$\int_{\mathcal{X}} \exp(-\lambda(\tau(x)-\tilde{\tau}))(\tau(x)-\tilde{\tau})dF_X(x) = 0 \tag{1.9}$$

ii) The value of the robustness metric $\delta^*(\tilde{\tau})$ *is given by:*

$$\delta^*(\tilde{\tau}) = D_{KL}(F_X^*||F_X) = -\log\left(\int_{\mathcal{X}} \exp(-\lambda(\tau(x)-\tilde{\tau}))dF_X(x)\right) \tag{1.10}$$

Proof. See Appendix 1.I. □

Lemma 7 greatly simplifies the computation of the robustness metric by essentially showing that the fully general robustness problem that searches over the nonparametric space of probability distribution is no-harder than the parametric cases in Examples 1 and 4. We can compare the closed form solution of Lemma 7 with the KKT solution one could derive for Example 1 and verify that the two solutions are indeed identical.

Example 8. *Return to the example of the discrete variable so* $X = \{0,1\}$*. First notice that the dominating measure here is the counting measure on* $\{0,1\}$*. We are therefore interested in simply the ratio* $\frac{p_1}{p_1^*}$ *since it completely characterizes* p_1^**. Because the problem is one dimensional, the unique minimizer is the one that satisfies the constraint:*

$$\tau(1)\cdot p_1^* + \tau(0)\cdot(1-p_1^*) = \tilde{\tau} \implies p_1^* = \frac{\tilde{\tau}-\tau(0)}{\tau(1)-\tau(0)} \tag{1.11}$$

Recall that in Example 1 $\tilde{\tau} = 0$*. On the other hand, from the solution provided by 7 we have:*

$$\frac{p_1^*}{p_1} = \frac{\exp(-\lambda(\tau(1)-\tilde{\tau}))}{\exp(-\lambda(\tau(1)-\tilde{\tau}))\cdot p_1 + \exp(-\lambda(\tau(0)-\tilde{\tau}))\cdot(1-p_1)} \tag{1.12}$$

where λ *is implicitly defined as in Equation (7).*

Fact 9. *Equations 1.12 and 1.11 are equivalent.*

Proof. See Appendix 1.I. □

Lemma 7 completely characterizes the robustness metric in terms of the (quasi)-experimental distribution $F_X(x)$ and the CATE, $\tau(x)$. This is important because both of them are nonparametrically identified from the (quasi)-experimental data. Hence, to give an answer to the policy-makers robustness problem, it is enough to estimate the treatment effect heterogeneity in $\tau(x)$. This result will deliver a very convenient estimation theory which I discuss in Section 1.3.

1.2.4 Locally infeasible problem

We have seen how the restriction in Assumption 4 is key to guarantee that a solution to Equation (1.4) exists and that the associated $\delta(\tilde{\tau})$ is finite. There is a partial extension to Lemma 7 with respect to a local violation of Assumption 4. Consider a sequence of $\tilde{\tau}_m$ converging to a boundary point $\tilde{\tau}_b$ of the range of $\tau(X)$. An example is depicted in Figure 1.2. Suppose the policy-maker's claim is given by: $ATE \leq \tilde{\tau}_m$.

For each $\tilde{\tau}_m$ within the range of variation of $\tau(X)$, the policy-maker's problem has a solution, $F^*_{X,m}$ given by Lemma 7. This is because there is a sub-population with covariates x such that $\tau(x) \geq \tilde{\tau}_m$. The *least favorable distribution* will increase the weight on this sub-population. If $\tilde{\tau}$ is on the boundary, for example $\tilde{\tau} = 3$ in Figure 1.2, the only sub-population that has $\tau(x) \geq \tilde{\tau}_b$ is $x = 0.6$, concentrated on a singleton. But distributions that put unit mass on singletons are not feasible in the policy-maker's problem. For $\tilde{\tau} = \tilde{\tau}_b$, the feasible set is empty so there is no solution. If one looks at the sequence of *least favorable distributions*, $F^*_{X,m}$, associated to the sequence $\tilde{\tau}_m \to \tilde{\tau}_b$, is there a limiting distribution to which the sequence $F^*_{X,m}$ converges in some sense?

Under some additional assumptions, one can show a type of concentration result for

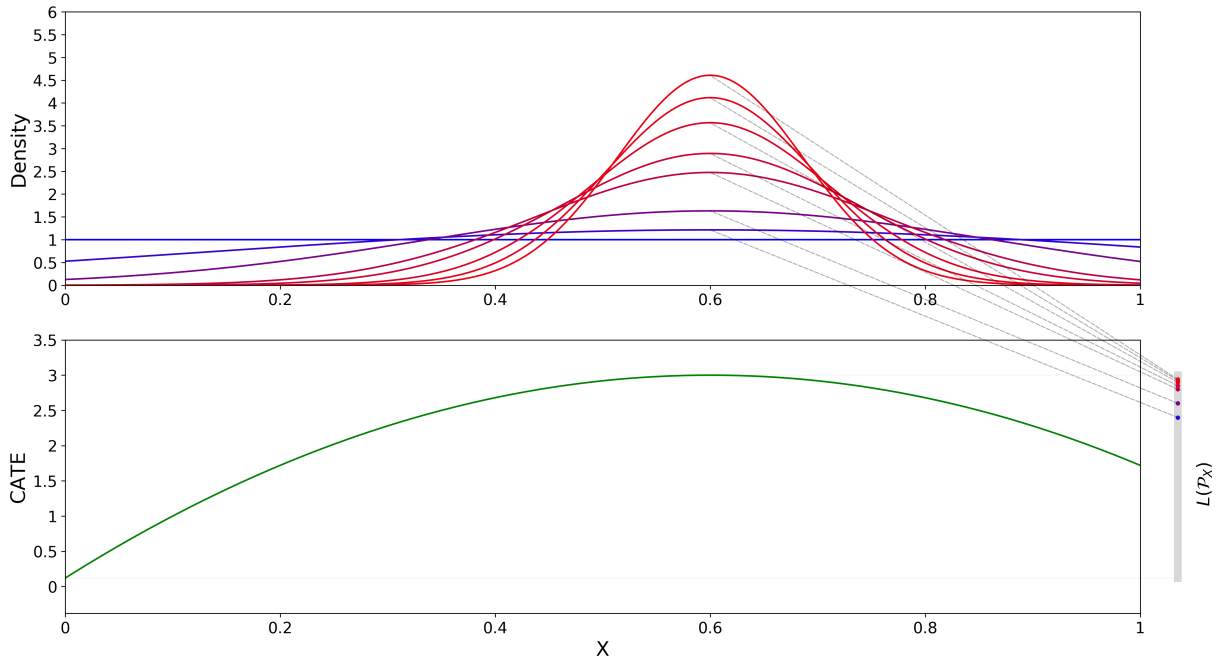


Figure 1.2: Local to boundary conditions. The lower panel displays the conditional average treatment effect, $\tau(x)$ for a univariate variable X. The experimental distribution is in blue: the uniform distribution. The gray segment on the left labelled $L(\mathcal{P}_X)$ is the image of the collection of probability distributions supported on $\mathcal{X}$ under the map $L : F_X \mapsto \int_{\mathcal{X}} \tau(x) dF_X(x)$. For every point in the interior, Lemma 7 holds and, for each $\tilde{\tau}_m$, there is an associated *least favorable distribution* $F^*_{X,m}$ displayed in the upper panel. As the sequence of $\tilde{\tau}_m$ approaches the boundary of $L(\mathcal{P}_X)$, the distributions concentrate around $x = \arg\max \tau(x) = 0.6$.

the sequence of solutions obtained by applying the closed-from solution formula in Lemma 7. If $\tau(x)$ is a single peaked function, that is, it achieves its maximum (or minimum) at a single point, we obtain convergence in distribution of the sequence $F^*_{X,m}$ to the Dirac distribution at the single peak, δ_{x_b}.

Proposition 10 (Local to boundary $\tilde{\tau}$)**.** *Let Assumptions 1-3 hold and let* $\tilde{\tau}_m \to \tilde{\tau}_b \in \partial L(\mathcal{P}_X)$*. Assume that the pre-image* $\tau^{-1}(\tilde{\tau}_b) = \mathcal{X}_b = \{x_b\} \in \mathcal{X}^\circ$ *is a singleton. Further, let* X *be compactly supported, with density* $f(x) < M$ *on* $\mathcal{X}$*. Then the sequence of least favorable distributions for the policy-maker's problem with parameter* $\tilde{\tau}_m$*, denoted* $F^*_{X,m}$*,*

converges weakly to δ_{x_b}, the Dirac delta distribution with point mass at x_b, that is:

$$\lim_{m\to\infty}\int_{\mathcal{X}} g(x)dF^*_{X,m}(x) \to \int_{\mathcal{X}} g(x)\delta_{x_b} := g(x_b)$$

for $g \in C_b(\mathcal{X})$, the space of all continuous, bounded functions on $\mathcal{X}$.

Proof. See Appendix 1.I. □

The point-mass distribution δ_{x_b} is not a solution to the policy-maker's problem with parameter $\tilde{\tau}_b$ because the feasible set never includes point mass distributions unless $\mathcal{X}$ is discrete. In this sense, Proposition 10 delivers the limit of the sequence of solutions in the sense of weak convergence. This is a weaker that the notion of convergence induced by D_{KL}. In particular when $dF_X \ll \lambda_{Leb}$ (the Lebesgue measure on $\mathbb{R}^k$), $D_{KL}(dF^*_{X,m}||\delta_{x_b}) = +\infty$ so the sequence of solutions $F^*_{X,m}$ does not converge to δ_{x_b} in D_{KL}.[7]

1.2.5 Interpreting robustness

In this section I offer some practical guidance on how to interpret the the robustness metric proposed in Definition 18. The first interpretation links the robustness metric to a bound on the probability of drawing a sample of size n for which the experimental conclusion is false. The second interpretation is a bench-marking exercise using available census covariates.

A probability interpretation using Sanov's theorem

One way to link the magnitude of the robustness metric $\delta^*(\tau)$ to an easily interpretable probability bound is through Sanov's theorem. In this section I provide the intuition through a finite dimensional example and give the interpretation. I discuss more

[7] In fact, Posner [1975] showed that D_{KL} is lower-semicontinuous, that is, if $P_n \to P$ weakly, then $\lim_{n\to\infty} D_{KL}(P_n||Q) \geq D_{KL}(P||Q)$. In this case we have $+\infty > 0$

details on the foundations of Sanov's theorem in Appendices 1.F. First, consider the setting of Example 4. Now suppose we collect a sample containing n i.i.d observations. Consider a generic sequence of the data of size n, $x = (x_1, x_2, \cdots, x_n)$. Each sequence is an ordered list of values $(High, Medium, Low)$. Define the *type* P_x of a sequence x as the proportion (relative to n) of realizations of a in x. This is $P_x(a) = \frac{N(a|x)}{n}$ where $N(a|x)$ is the number of times realization a shows up is in sequence x. We denote the collection of types as $\mathcal{P}_n$.[8]

For the present example, a result by Cover [1999] shows that while the number of sequences is of the order of 3^n, the number of types is bounded above by $|\mathcal{P}_n| \leq (n+1)^3$. We can look at the *types* that fall within a specific subset E of probability distributions. For example we can look at all the *types* that invalidate the experimental conclusion on the ATE. In this case $E := \{Q \in \mathcal{P}_X : \int_{\mathcal{X}} \tau(x) dQ \leq \tilde{\tau}\}$, the constraint in Equation 1.5. Notice that whether a sequence $x \in E$ or not depends only on its *type* P_x. Now, what is the probability that, drawing a sequence x according to P_X, such a sequence invalidates the experimental results, that is $x \in E$?. It turns out that Sanov's theorem provides a link between this probability and the metric of robustness $\delta^*(\tau)$.

Theorem 11. *(Sanov's theorem) Let $X_1, \cdots X_n$ be i.i.d distributed according to F_X. Let E be a convex set of probability distributions. Letting P_X^n be the product measure of n copies of P_X. Then*

$$P_X^n(E \cap \mathcal{P}_n) \leq e^{-nD_{KL}(P_X^* || P_X)}$$

where

$$P^* = \min_{Q \in E} D_{KL}(Q||P)$$

[8] One can think of a *type* P_x as keeping track of the proportion but forgetting the order. So for example the two sequences of size $n = 3$ given by $x = (High, Medium, Medium)$ and $x' = (Medium, High, Medium)$ are distinct: $x \neq x'$. But they have the same type: $P_{x'} = P_x$.

Moreover, if the set E is the closure of its interior then

$$\lim_{n\to\infty} \frac{1}{n} \log(P^n(E)) \to -D_{KL}(P^*|P)$$

Proof. The proof can be found in Cover [1999] Theorem 11.4.1. □

Note that $E := \{Q : \int_{\mathcal{X}} \tau(x) dQ \leq \tilde{\tau}\}$ is obtained through imposing a linear restriction on Q and therefore E is convex. Sanov's theorem remains true for larger classes of probability distributions, not necessarily confined to finitely supported X variables like discussed in Csiszár [1984]. Note that $\delta^*(\tilde{\tau}) = D_{KL}(P^*||P)$ is precisely the metric of robustness $\delta^*(\tau)$. It captures the smallest distance from the experimental distribution of the covariates that will fail to satisfy the conclusion, hence a bound for the probability that a sequence does not satisfy the policy-maker's conclusion is given by

$$P_X^n(E) \leq e^{-n\delta^*(\tilde{\tau})}$$

The fact that the probability bound depends on $\tilde{\tau}$ should not be surprising since $\tilde{\tau}$ also controls the inequality that defines the constraint set E. Notably the bound is non-asymptotic in that it applies for any n. The bound is monotonically decreasing in the magnitude of $\delta^*(\tau)$ and it becomes trivial when $\delta^*(\tau) = 0$. Of course if $\delta^*(\tau) = \infty$ we know that the set E does not contain any valid distributions, so it is reasonable that $P_X^n(E) = 0$. Below, we may revisit the discrete example to get a sense of the estimate that Sanov's theorem provides.

Example 4 (Continued)**.** *Recall X = income, $\mathcal{X} = \{High, Medium, Low\}$ and the experimental distribution is $F_X = (p_1, p_2, p_3) = (0.2, 0.2, 0.6)$. For a given n we can list the types of sequences of size n that can be generated. Here the count of $High$ and $Medium$ income individuals will completely determine the type of a sequence (since for fixed n,*

$\#Low = n - \#High - \#Medium$. *For* $n = 3$ *for example, there are 10 possible sequence types each corresponding to one of the sequences* $(3,0,0)$, $(2,0,1)$, $(2,1,0)$, $(1,2,0)$, $(1,1,1)$, $(1,0,2)$, $(0,3,0)$, $(0,2,1)$, $(0,1,2)$, $(0,0,3)$ *divided by 3. Therefore* $|\mathcal{P}_3| = 10$. *They are displayed below in barycentric coordinates as red points in the 2-simplex. The set E is also displayed in yellow.*

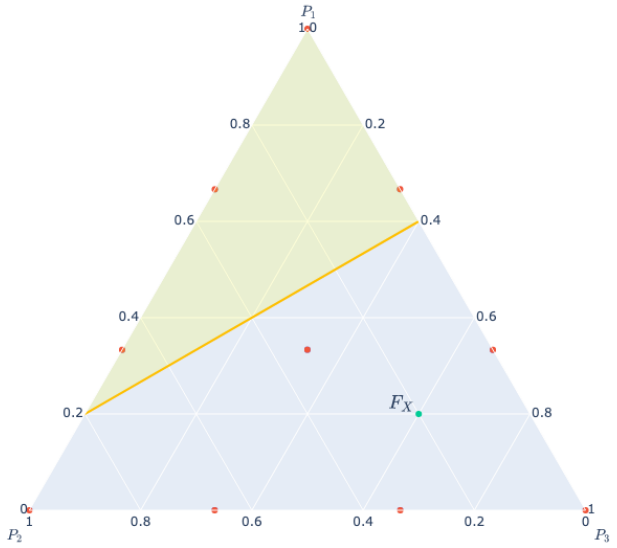


Figure 1.3: Distribution *types* for the 3 point space, $n = 3$

For $n = 10$ *there are* 110 *distinct sequence types, that is,* $|\mathcal{P}_{10}| = 110$. *They are displayed in the figure below.*

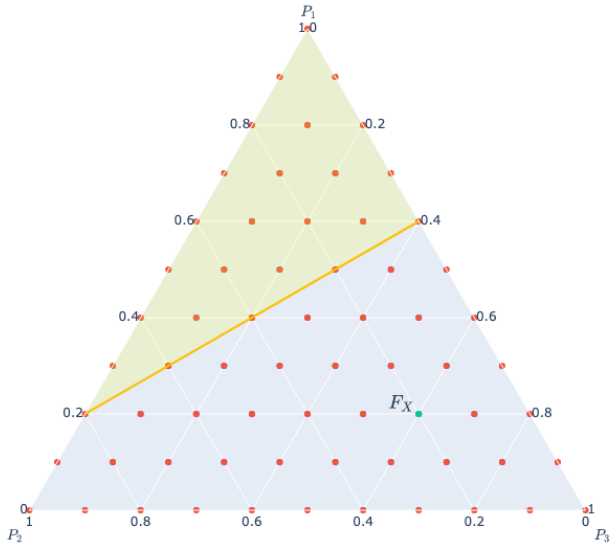


Figure 1.4: Distribution *types* for the 3 point space, $n = 10$

Note that each of the sequence types may contain many sequences. Because the draws from the distribution F_X are i.i.d, all sequences of the same type have the same probability under P_X. The result in Sanov's theorem gives a finite sample upper bound on the probability that a sequence $X_l = (X_{1,l}, \cdots, X_{n,l})$, drawn from the joint distribution P_X^n belongs to the set E. For $n = 3$ there are only 4 sequence types that are in E, namely $(3,0,0)$, $(2,0,1)$, $(2,1,0)$, $(1,2,0)$. What is the probability associate to them? $P_X^3(x_l \in E) = 0.128$. On the other hand, we know that $\delta^*(\tilde{\tau}) =$ so Sanov's theorem gives the upper bound $e^{-3 \cdot 0.2492} = 0.474$ so the bound is fairly loose. On the other hand, when $n = 10$, 26 out of 110 sequence types fall in the set E. The total probability associated with those sequences is 0.0174. Sanov's theorem gives an upper bound of 0.0827. Finally for $n = 30$ $P_X^{30}(x \in E) = 0.000083$, while Sanov's bound gives $P_X^{30}(x \in E) \leq 0.00057$. The bound is known to be optimal in the exponent for $\lim n \to \infty$.

Benchmarking robustness using census covariates

Several papers have proposed benchmarking the sensitivity to unobservable variables, which is often not computable, using observable variables. For example, Cinelli and Hazlett [2020] and Oster [2019] who use the explanatory power of observed covariates to benchmark for the explanatory power of unobserved covariates. This section suggests a similar approach for the robustness problem. In the context of this paper I would like to quantify whether a given value for the robustness parameter, δ^* is high or low. To this end I propose to leverage the subset of covariates in X_c, which are available in both the (quasi)-experimental environment and in the extrapolation environment to benchmark the robustness measure. At the population level it amounts to:

- computing the robustness metric δ^* through Equation (1.10)
- use the census information to compute $D_{KL}(P'_{X_c}||P_{X_c})$, the KL divergence between the distributions of the X_c covariates in the (quasi)-experimental population and the new population
- compare the two measures

If the variables in X_c collectively differ across the two environments by the same amount as X_e, observing $\delta^* > D_{KL}(P'_{X_c}||P_{X_c})$ suggests that the (quasi)-experimental claim can be extrapolated to the new environment. In words, it says that the distance, measured by the KL divergence between the observable census variables in the two environments would not be large enough to invalidate the claim drawn from the (quasi)-experimental evidence. In principle, one could develop a formal test that uses both δ^* and $D_{KL}(P'_{X_c}||P_{X_c})$ (under the assumption that the true distance in X_e is no larger than the true distance in X_c) to provide a pessimistic policy-maker with a clear rule on when to expand the policy given the census data. For now, transforming the heuristic exercise above in a full fledged two-sample test is beyond the scope of this paper and I leave it to future research.

1.2.6 A conditional limit theorem interpretation for F_X^*

We have seen that the value of $\delta^*(\tau)$ has a natural interpretation as a probability bound. What about the *least favorable distribution* F_X^*, the minimizer of Equation (1.4)? It turns out that an extension of the result by Sanov provides a new perspective for it. Adapting a version of Theorem 1 in Csiszár [1984], one obtains a striking result on the joint distribution of the data $(X_1, \cdots X_n)$:

Theorem 12. *(adapted from Csizar, 1984) Let Assumptions 2 - 4 hold. Set* $E = \{Q : \int_{\mathcal{X}} \tau(x) dQ \leq \tilde{\tau}\}$*, let* P_X *be the probability measure of i.i.d data. Denote the empirical distribution of* $X_1, \cdots, X_n$ *as* $\hat{F}_n$*. Then:*

(i) the random variables $X_1, \cdots, X_n$ *are asymptotically quasi-independent*[9] *conditional on the event that the empirical distribution* $\hat{F}_n \in E$

(ii) $P(X_i | \hat{F}_n \in E) \approx P^*(X_i)$ *for* $i = 1, \cdots, n$

Proof. The proof follows straightforwardly from Theorem 1 in Csiszár [1984] noting that, by Assumption 4, condition (2.18) in Csiszár [1984] is satisfied. For finitely supported X, an easier proof is given in Theorem 11.6.2 in Cover [1999]. □

In contrast to Theorem 11 which holds for any n, Theorem 12 is an asymptotic result: the approximation of the conditional law in *ii*) depends on the sample size n. The interpretation is the following, $P^{*n} := \Pi_{i=1}^n P^*$ is the approximate joint law of the covariates $X_1, \cdots X_n$, if we learned that the empirical distribution $\hat{F}_n$ does not satisfy the experimental conclusions. To visualize this, imagine drawing S-many repeated samples of n observations each from the covariate distribution. Then, combining the Sanov theorem in Section 1.2.5 together with the Csiszár [1984] conditional limit theorem tells us that:

(i) $\lim_{S \to \infty} \frac{1}{S} \sum_{l=1}^{S} \mathbb{1}[\hat{F}_{n,l} \in E] \leq e^{-n\delta^*(\tilde{\tau})}$

[9] See Definition 2.1 in Csiszár [1984].

(ii) $P_X^n(X_i|\hat{F}_{n,l} \in E) \approx P^{*n}(X_i)$ for any $i = 1, \cdots, n$ and $l = 1, \cdots, S$

Part (i) says that the proportion of samples of size n that fail to satisfy the experimental evidence is bounded above by $e^{-n\delta^*(\tilde{\tau})}$. This interpretation is closest to the robustness approach in Broderick et al. [2020] which is based on dropping a percentage of the sample. The difference is that their procedure focuses on a proportion of the fixed sample, whereas this result concerns the proportion all possible samples of size n that could be drawn from the joint distribution of P_X^n. A small value for the robustness metric $\delta^*(\tilde{\tau})$ will not control this probability very well. Part (ii) gives an approximate law for the joint distribution P_X^n of the collection of samples that invalidate the experimental results. This tells us that the F_X^* is not just a by-product of the optimization problem in Equations (1.4) and (1.5) but it gives the approximate law of the data if we happen to draw a sample l which does not satisfy the experimental results.

1.3 Estimation and Asymptotic Results

In this section I introduce a semi-parametric estimator for my robustness metric δ^*, according to Definition 3 ii) and I characterize its asymptotic properties. I show that the robustness metric can be estimated using a GMM criterion function which only depends on the (quasi)-experimental distribution and on the CATE $\tau(x)$, both of which are identified in the quasi experiment. The theory is based on constructing the nonparametric influence function correction for the de-biased GMM procedure in Chernozhukov et al. [2020] to account for flexible nonparametric estimation of $\tau(x)$. The proofs are in the Appendix 1.I.

1.3.1 An empirical estimate of the robustness metric δ^*

The closed form solution in Lemma 7 suggests a natural estimator based on empirical averages. In particular, one would like to replace Equation (1.10) with its sample analog using the Generalized Method of Moments (GMM) framework. Consider the quantities:

$$\nu_0 := \int_{\mathcal{X}} \exp(-\lambda_0(\tau(x) - \tilde{\tau}))dF_X(x)$$

where λ_0 is defined implicitly as the unique solution to:

$$\int_{\mathcal{X}} \exp(-\lambda_0(\tau(x) - \tilde{\tau}))(\tau(x) - \tilde{\tau})dF_X(x) = 0$$

The pair of parameters that solves the population moment condition is denoted by $\theta_0 = (\nu_0, \lambda_0)^T$. Ultimately, the robustness measure $\delta^* = -\log(\nu_0)$ is the parameter of interest. The asymptotic theory for δ^* follows directly from establishing the asymptotic theory for $\hat{\theta} = (\hat{\nu}, \hat{\lambda})^T$ hence, I will focus on these parameters in this section. The parameter space $\Theta \subseteq \mathbb{R}^2$ satisfies some constraints. First, observe that if the policy-maker's claim ($ATE \geq \tilde{\tau}$) holds with a strict inequality for the (quasi)-experimental distribution, then the true $\delta^* > 0$. This implies a restriction on $\nu_0 < 1$. Moreover, $\nu_0 > 0$ because by the properties of the exponential, the quantity $\exp(-\lambda(\tau(x) - \tilde{\tau}) > 0$ for all $x \in \mathcal{X}$. Hence, the restriction on ν is $0 \leq \nu_0 \leq 1$.

Let $W = (X, D, Y)$ be the data. Then, as in Newey and McFadden [1994] we can write the moment condition jointly for ν_0 and λ_0 as:

$$\mathbb{E}[g(W, \theta, \tau)] = \mathbb{E}\begin{bmatrix} \exp(-\lambda_0(\tau_0(X) - \tilde{\tau})) - \nu_0 \\ \exp(-\lambda_0(\tau_0(X) - \tilde{\tau}))(\tau_0(X) - \tilde{\tau}) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{1.13}$$

where $\tau_0(X)$ denotes the true value of CATE. Assumptions 1–4 guarantee that the parame-

ters of interest (λ_0, ν_0) are (globally) identified by Equation (1.13). Because the true value for $\tau_0(X)$ is an unknown but estimable population quantity, I consider a feasible version of Equation (1.13) that uses an estimate $\hat{\tau}(X)$ in place of $\tau_0(X)$. One could define the vector $\hat{\theta} = (\hat{\lambda}, \hat{\nu})^T$ is defined as the approximate solution to the empirical moment:

$$\mathbb{E}_n[g(W, \theta, \hat{\tau})] = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} \exp(-\hat{\lambda}(\hat{\tau}(X_i) - \tilde{\tau})) - \hat{\nu} \\ \frac{1}{n}\sum_{i=1}^{n} \exp(-\hat{\lambda}(\hat{\tau}(X_i) - \tilde{\tau}))(\hat{\tau}(X_i) - \tilde{\tau}) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{1.14}$$

where $\hat{\tau}(X)$ is a plug-in estimate of the conditional average treatment effect. While Assumption 1 guarantees nonparametric identification of $\tau_0(X)$, there are many ways that one could estimate it, both parametrically and nonparametrically. For example Athey et al. [2016] uses random forest, Hsu et al. [2020] uses a doubly robust score function.

One caveat of the estimator based on Equation (1.14) is that the identifying moment conditions provided in Equation (1.13) are not Neyman orthogonal with respect to the first-step estimator $\hat{\tau}(X)$. As a result, the first-step estimation of $\hat{\tau}(X)$ can, in general, have a first-order effect on the estimator for $\theta_0 = (\nu_0, \lambda_0)^T$, and consequently on the estimator for δ^*, and possibly lead to incorrect inferences on the robustness metric, a general problem discussed in Chernozhukov et al. [2018]. Deriving primitive conditions on this form of the moment condition requires *ad-hoc* conditions on the first-step nonparametric estimator that can be hard or inconvenient to check in practice. As an alternative, I use the debiased-GMM approach in Chernozhukov et al. [2020] that allows to choose flexible estimators for $\tau_0(X)$ while automatically correcting for the first-order bias.

1.3.2 Nonparametric influence function correction and de-biased GMM estimator

In this section, I derive the nonparametric correction for the GMM estimator of θ based on Equation (1.14). I map the causal quantities like $\tau(X)$ to the statistical functionals that identify them and then explicitely construct the nonparametric influence function for these functionals. Because these functionals are always implicitly regarded as mapping the distribution function of the data, F, to some space, it is natural to index the functional with a subscript F. For example the $\tau(X) = \tau_F(X)$ because depends of the distribution of the data F. The true distribution of the data will be denoted as F_0 and it is understood that $\tau_0(X) = \tau_{F_0}(X)$. Recall that $\tau_{F_0}(X)$ is a causal parameter which needs to be identified through the distribution of the data. By Assumption 1, $\tau_{F_0}(X)$ can be nonparametrically identified as the difference between the conditional means: $\tau_{F_0}(X) = \gamma_{1,F_0}(X) - \gamma_{0,F_0}(X)$ where $\gamma_{1,F}(X) := \mathbb{E}_F[Y|X, D = 1]$ and $\gamma_{0,F}(X) := \mathbb{E}_F[Y|X, D = 0]$. The left hand side features a causal quantity while the right hand side features two statistical quantities. The first step then has two functions that need to be estimated. For convenience, I gather them into a single vector-valued statistical functional $\gamma_F = (\gamma_{0,F}, \gamma_{1,F})^T$. When considering the de-biasing term to correct for the first-step estimation of $\tau_{F_0}(X)$, we actually need to consider the first-step correction with respect to the full vector γ_F.

Now consider a parametric sub-model for the distribution function, consisting of $F_r := (1-r) \cdot F_0 + rH$ where F_0 is the true baseline distribution function of the data and H is an arbitrary distribution function which satisfies Assumption 1. For any $r \in [0, 1], F_r$ is a mixture distribution and hence, it is also a valid distribution function. Moreover, if both F_0 and H satisfy Assumption 1 then F_r does as well. In order to de-bias the moment conditions in $\mathbb{E}[g(W, \theta, \tau_F)]$ with the approach of Chernozhukov et al. [2020] one needs to compute the nonparametric influence function with respect to τ_F. The nonparametric influence

function maps infinitesimal perturbations of F in the direction of H in a neighborhood of F_0, to perturbations in $\mathbb{R}^2$ (because there are 2 moment conditions). It does so *linearly* in H. In particular, the nonparametric influence function of $\mathbb{E}[g(W,\theta,\tau_F)]$ with respect to F, labelled $\phi(\cdot)$ is implicitly defined by the equation below:

$$\left.\frac{d\mathbb{E}[g(W,\theta,\gamma_{F_r})]}{dr}\right|_{r=0} = \int \phi(w,\gamma_{F_0},\theta,\alpha)dH(w) \tag{1.15}$$

Note that, other than the original arguments of $g(\cdot)$, which feature the vector of conditional means γ_{F_0}, $\phi(\cdot)$ is allowed to depend on additional nonparametric components, gathered in $\alpha(\cdot)$. In the next result I derive the nonparametric influence function explicitly.

Proposition 13. *The de-biased GMM nonparametric influence function based on moment function $g(\cdot)$ is:*

$$\begin{aligned}\phi(w,\theta,\gamma_0,\alpha_0) = &\begin{bmatrix} \exp\left(-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})\right)\cdot(-\lambda) \\ \exp\left(-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})\right)\cdot(1-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})) \end{bmatrix} \\ &\times\left(\frac{d(y-\gamma_{1,F_0}(x))}{\pi_{F_0}(x)} - \frac{(1-d)(y-\gamma_{0,F_0}(x))}{1-\pi_{F_0}(x)}\right)\end{aligned}$$

which could be written in the form:

$$\begin{aligned}\phi(w,\theta,\gamma_0,\alpha_0) = &\begin{bmatrix} \exp\left(-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})\right)\cdot(-\lambda) \\ \exp\left(-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})\right)\cdot(1-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})) \end{bmatrix} \\ &\times\left(\begin{bmatrix}\alpha_{1,F_0}(x)\\ \alpha_{0,F_0}(x)\end{bmatrix}^T \begin{bmatrix} d(y-\gamma_{1,F_0}(x)) \\ (1-d)(y-\gamma_{0,F_0}(x))\end{bmatrix}\right)\end{aligned}$$

with $\alpha_{F_0}(x) := \begin{bmatrix}\alpha_{1,F_0}(x)\\ \alpha_{0,F_0}(x)\end{bmatrix} = \begin{bmatrix} \frac{1}{\pi_{F_0}(x)} \\ \frac{1}{1-\pi_{F_0}(x)}\end{bmatrix}$.

There are two main multiplicative terms in $\phi(\cdot)$. The first term is the derivative of the moment conditions with respect to the first-step estimator. The second one is the variation of individual treatment effects about their conditional mean, appropriately weighted by the propensity score. One can immediately check that, by the law of iterated expectations, $\mathbb{E}_F[\phi(W,\theta,\gamma_0,\alpha_0)] = 0$ for any θ. Hence we can form the de-biased GMM moment functions by taking:

$$\psi(w,\gamma,\theta,\alpha) = g(w,\theta,\gamma) + \phi(w,\theta,\gamma,\alpha) \tag{1.16}$$

Notice that $\mathbb{E}_{F_0}[\psi(W,\theta,\gamma_0,\alpha_0)] = 0$ so an estimator for θ that uses the de-biased moment function $\psi(\cdot)$ instead of $g(\cdot)$ will preserve identification. Standard conditions can be given to guarantee $\mathbb{V}[\psi(W,\theta,\gamma_0,\alpha_0)] < \infty$ so that $\psi(W_i,\theta,\gamma_0,\alpha_0)$ is a valid influence function. As emphasized in Chernozhukov et al. [2020] the de-biased GMM form of $\psi(\cdot)$ corrects for the first order bias induced by replacing $\gamma_{1,F_0} - \gamma_{0,F_0}$, the statistical counterpart of the true τ_{F_0}, with a flexibly estimated $\hat{\gamma}_1 - \hat{\gamma}_0$. In particular, for $\sqrt{n}$-consistency of θ, the estimators for $\hat{\gamma}_1$ and $\hat{\gamma}_0$ only need to satisfy mild conditions on the L^2-rate of convergence in Assumption 5 below. This allows to characterize simple inference for the robustness measure $\hat{\delta}^*$ while allowing for flexible nonparametric estimation of γ_{1,F_0} and γ_{0,F_0} using a large collection of machine learning-based estimators which include, among others, random forest, boosting, and neural nets. In practice, machine learning methods can help when the covariate space is high-dimensional but the true $\tau_0(X)$ has a sparse representation.

The key property to guarantee de-biasing is given by the Neyman orthogonality of the new moment conditions with respect to the first-step estimator, established in the result below.

Proposition 14. *Equation* (1.16) *satisfies Neyman orthogonality.*

Proof. See Appendix 1.I. □

Consider now the empirical version of the de-biased GMM equations:

$$\hat{\psi}(\theta, \hat{\gamma}, \hat{\alpha}) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|I_k|} \sum_{i \in I_k} \Big(g(W_i, \theta, \hat{\gamma}_{-k}) + \phi(W_i, \tilde{\theta}, \hat{\gamma}_{-k}, \hat{\alpha}_{-k}) \Big)$$

The de-biased GMM estimator takes advantage of a cross-fitting procedure where the sample is split into K many folds. For each fold $k = 1, \cdots, K$, the nonparametric components in $\psi(\cdot)$, that is, the $\gamma(\cdot)$ and $\alpha(\cdot)$ functions, are estimated on the observations in the remaining $(K-1)$ folds which explains the indexing $-k$ in the subscripts of $\gamma(\cdot)$ and $\alpha(\cdot)$. Sample splitting reduces own-observation bias and, together with the Neymann orthogonality property established above, avoids complicated Donsker-type conditions that would potentially not be satisfied for some first-step estimators of $\hat{\gamma}$ and $\hat{\alpha}$, as discussed in Chernozhukov et al. [2020]. Finally note that $\tilde{\theta}$ is a consistent estimator for θ needed to evaluate ϕ. For example one could use the θ from the plug-in GMM which is consistent but may not be $\sqrt{n}$-consistent in general. The de-biased GMM estimator is given by:

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \hat{\psi}(\theta, \hat{\gamma}, \hat{\alpha}) \tag{1.17}$$

To establish $\sqrt{n}$-convergence of the GMM estimators for θ, some quality conditions on the L_2 rates of convergence of the first-step estimators for γ and α are required.

Assumption 5. *For any k, $\|\hat{\gamma}_{-k} - \gamma_0\|_L^2 = o_P(N^{-\frac{1}{4}}); \|\hat{\alpha}_{-k} - \alpha_0\|_L^2 = o_P(1)$.*

In Appendix 1.I, I use Assumptions 1 – 5 to prove the influence function representation for $\hat{\theta}$ to which a standard central limit theorem applies to establish the asymptotic normality of the de-biased GMM estimator for $\theta = (\nu, \lambda)^T$. This, in turn, allows to conduct inference on the parameter of interest, δ^* through a straightforward application of the delta method.

Theorem 15 (Asymptotic normality of θ)**.** *Let Assumptions 1–5. For $\hat{\theta}$ defined in Equation*

(1.17)*:*

$$\begin{aligned}
\sqrt{n}(\hat{\theta}-\theta_0) &\xrightarrow{d} \mathcal{N}(0,S)\\
S &:= (G)^{-1}\Omega(G')^{-1}\\
G &:= \mathbb{E}[D_\theta\psi(w,\theta,\gamma_0,\alpha_0)]\\
\Omega &:= \mathbb{E}[\psi(w,\theta_0,\gamma_0,\alpha_0)\psi(w,\theta_0,\gamma_0,\alpha_0)^T]
\end{aligned}$$

and $D_\theta\psi(\cdot)$ is the Jacobian of the augmented moment condition with respect to the parameters in θ.

Proof. See Appendix 1.I. □

The parameter of interest follows from a straightforward application of the parametric delta method.

Corollary 16 (Asymptotic normality of δ^*)**.** *Let $\hat{\delta}^* = -\log(\hat{\nu})$. Then*

$$\sqrt{N}(\hat{\delta}^* - \delta^*) \xrightarrow{d} \mathcal{N}\left(0, \frac{S_{11}}{\nu_0^2}\right)$$

where S_{11} is the (1,1) entry of the variance covariance matrix S in Theorem 15.

With the results of Theorem 15 one can obtain a point estimate δ^*, together with a confidence interval for a pre-specified coverage level. Because of the nature of the estimand, the researcher or the policy-maker, are likely to care especially about the lower bound for δ^*. This is because overestimating the δ^* implies that there is a distribution of the covariates within the estimated $\hat{\delta}^*$ that invalidates the policy-maker's claim. This defies the entire purpose of the robustness exercise. On the other hand, underestimating δ^* may result in unduly conservative characterization of the set of distributions for which the claim is valid, but it does not defy the purpose of the robustness exercise. A similar, asymmetric

approach is followed by Masten and Poirier [2020] who report a lower confidence region for their breakdown frontier rather than a confidence band.

1.3.3 Reporting features of the *least favorable distribution*

Lemma 7 gives an explicit formula for the *least favorable distribution* F_X^* and shows that it depends on λ_0 and $\tau(X)$. Because of the interpretation of F_X^* as the conditional law of the data that we have given in Section 1.F, the researcher may be interested in F_X^* directly. If $\mathcal{X} \subseteq \mathbb{R}^d$ is even moderately high dimensional, it may be very inconvenient to look at features of the estimated F_X^*. Moreover, the rate of convergence of as estimator of F_X^* can, in general, be nonparametric. This is because, under some conditions, it inherits the nonparametric rate of $\hat{\tau}(X)$. Rather, the researcher could report particular moments of F_X^* that are of interest. This exercise is analogous to reporting moments of the covariate distribution and compare them across treatment status to gauge at covariate balance, like in Rosenbaum and Rubin [1984]. The researchers may want to report moments of F_X^*, in addition to the robustness metric δ^*. For example they may want to report a vector of covariate means under the *least favorable distribution* F_X^* and compare it with the (quasi)-experimental distribution. In such a case, we would like to construct an estimator for the moments of interest and establish the asymptotic theory of these estimators. I give a convenient extension of Theorem 15, to include an arbitrary, finite dimensional collection $\zeta \in \mathbb{R}^s$ of moments of interest, along with the original θ parameters.

Theorem 17 (De-biased estimator of *least favorable* moments)**.** *Let $u : \mathbb{R}^d \to \mathbb{R}^s$, with $u \in (L^\infty(\mathcal{X}, \mu))^s$ for μ some dominating measure of P_X. Let $\zeta_0 = \mathbb{E}_{F_X^*}[u(X)] \in \mathbb{R}^s$. Define the following estimating equation for the parameters $(\hat{\theta}, \hat{\zeta})$, that is, the original parameters*

of interest, augmented by ζ, the additional moments of the least favorable distribution:

$$\hat{\psi}^u(\theta,\zeta,\hat{\gamma},\hat{\alpha}) := \frac{1}{K}\sum_{k=1}^{K}\frac{1}{|I_k|}\sum_{i\in I_k}\begin{bmatrix} g(W_i,\theta,\hat{\gamma}_{-k})+\phi(W_i,\theta,\hat{\gamma}_{-k},\hat{\alpha}_{-k}) \\ g^u(W_i,\theta,\zeta,\gamma_{-k})+\phi^u(W_i,\theta,\zeta,\hat{\gamma}_{-k},\hat{\alpha}_{-k})\end{bmatrix}$$

where $g(\cdot),\phi(\cdot),\gamma(\cdot)$ and $\alpha(\cdot)$ are the same as in Propositions 13 – 27 and $g^u(\cdot)$ and $\phi^u(\cdot)$, whose values are vectors in $\mathbb{R}^s$ are defined below.

$$\begin{aligned} g^u(W_i,\theta,\zeta,\gamma) &= u(X_i)\exp(-\lambda(\tau(X_i)-\tilde{\tau})-\nu\cdot\zeta \\ \phi^u(W_i,\theta,\zeta,\gamma,\alpha) &= u(X_i)\exp\left(-\lambda(\tau(X_i)-\tilde{\tau})\right)\cdot(-\lambda) \\ &\quad\times\left(\frac{D_i(Y_i-\gamma_1(X_i))}{\pi(X_i)}-\frac{(1-D_i)(Y_i-\gamma_0(X_i))}{1-\pi(X_i)}\right)\end{aligned}$$

$$(\hat{\theta},\hat{\zeta}) := \arg\min_{(\theta,\zeta)\in\mathbb{R}^{s+2}}\hat{\psi}^u(\theta,\zeta,\hat{\gamma},\hat{\alpha})^T\hat{\psi}^u(\theta,\zeta,\hat{\gamma},\hat{\alpha})+o_P(1) \tag{1.18}$$

Let Assumptions 1–5 hold. Then:

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{K}\sum_{i\in I_k}\psi^u(W_i,\theta,\zeta,\hat{\gamma}_{-k},\hat{\alpha}_{-k}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi^u(W_i,\theta,\zeta,\gamma_0,\alpha_0)+o_P(1)$$

Moreover

$$\begin{aligned} \sqrt{n}\left(\begin{bmatrix}\hat{\theta}-\theta_0 \\ \hat{\zeta}-\zeta_0\end{bmatrix}\right) &\xrightarrow{d}\mathcal{N}(0,S^u) \\ S^u &:= (G^u)^{-1}\Omega^u(G^{u'})^{-1} \\ G^u &:= \mathbb{E}[D_{\theta,\zeta}\ \psi^u(W,\theta,\zeta,\gamma_0,\alpha_0)] \\ \Omega^u &:= \mathbb{E}[\psi^u(w,\theta_0,\gamma_0,\alpha_0)^T\psi^u(w,\theta_0,\gamma_0,\alpha_0)]\end{aligned}$$

where $D_{\theta,\zeta}$ denotes the Jacobian matrix with respect to the parameters θ and ζ.

Proof. The proof follows the same structure of Theorem 15 and is omitted. □

1.3.4 Simulation data

I conclude this section with a small Monte-Carlo exercise featuring three different data generating processes (DGPs) with increasing degrees of observable heterogeneity. To capture the idea of possibly high-dimensional experimental data, I consider a setting with $k = 100$ covariates, all independent and each distributed uniformly on $[0,1]$ so that $\mathcal{X} = [0,1]^k$. To reflect the fact that only a few out of all available experimental covariates are important to predict the treatment effect, I construct $\tau(x)$ to be sparse: $\tau(x)$ is a function of of only 1,3 and 10 out of 100 covariates in DGP1, DGP2 and DGP3 respectively. In each DGP, the potential outcomes also depend on an additive unobservable noisy error term.[10] To show that it is the heterogeneity that drives the robustness, keeping the same baseline ATE for the three DGPs is fundamental. I choose the shape of $\tau(x)$ to induce the same ATE across the three DGPs, regardless of the heterogeneity of treatment effects, when evaluated with respect to the experimental distribution. I consider $M = 1000$ replications for each DGP and a sample size of $N = 10,000$. The first step $\tau(x)$ is estimated through K-fold cross-fitting, using either boosting or random forest to estimate $\gamma_1(x)$, $\gamma_0(x)$ and the propensity score $\pi_X(x)$. The number of trees and splitting criteria are tuned to the sample size through heuristic criteria. In practice one would use *within-fold* cross-validation to tune hyper-parameters. I estimate the implied $\hat{\delta}^*(\tilde{\tau})$, with $\tilde{\tau} = 1.3$ and evaluate its bias, variance and MSE against the true value δ^*. Fixing the ATE and the experimental distribution of the covariates guarantees that a change in the population value for δ^* is only capturing the change in heterogeneity. I report the estimates of δ^* using both the plug-in GMM and de-biased GMM approach below. Note that, because of K-fold cross

[10] In particular:

- DGP1: $Y_1 - Y_0 = \exp(X_1) + U_1 - U_0$;
- DGP2: $Y_1 - Y_0 = \exp(X_1) \cdot (X_2 + 0.5) \cdot (X_3 + 0.5) + U_1 - U_0$;
- DGP3: $Y_1 - Y_0 = \exp(X_1) \cdot (X_2 + 0.5) \cdot (X_3 + 0.5) \cdot \Pi_{j=4}^{10}(0.1 \cdot X_j + 0.95) + U_1 - U_0$.

(U_1, U_0) are uncorrelated normals with $\mu = 0, \sigma = 0.25$.

fitting, the own-observation bias in the plug-in GMM is attenuated. Still, the de-biased, GMM shows very good bias improvements over the plug-in approach.

Table 1.1: Monte Carlo Simulation reports the DGP, the population value for the robustness metrics, ML estimator used for the nonparametric components and MSE, Bias and Variance. Sample size $n = 10,000$, number of simulations $M = 1000$.

Data	$\delta^*(\tilde{\tau})$	Method	$\gamma(\cdot)$, $\alpha(\cdot)$ est	MSE	Bias^2	Variance
DGP1	0.4485	plug-in	Random Forest	$3.7568 \cdot 10^{-4}$	$0.1235 \cdot 10^{-4}$	$3.6334 \cdot 10^{-4}$
			Boosting	$1.6311 \cdot 10^{-3}$	$1.2056 \cdot 10^{-3}$	$0.4255 \cdot 10^{-3}$
		de-biased	Random Forest	$3.7148 \cdot 10^{-4}$	$0.1030 \cdot 10^{-4}$	$3.6117 \cdot 10^{-4}$
			Boosting	$1.5278 \cdot 10^{-3}$	$1.1038 \cdot 10^{-3}$	$0.4240 \cdot 10^{-3}$
DGP2	0.1344	plug-in	Random Forest	$5.0716 \cdot 10^{-3}$	$4.9474 \cdot 10^{-3}$	$0.1242 \cdot 10^{-3}$
			Boosting	$1.1218 \cdot 10^{-3}$	$1.0622 \cdot 10^{-3}$	$0.0597 \cdot 10^{-3}$
		de-biased	Random Forest	$3.6640 \cdot 10^{-3}$	$3.5616 \cdot 10^{-3}$	$0.1024 \cdot 10^{-3}$
			Boosting	$0.7309 \cdot 10^{-3}$	$0.6749 \cdot 10^{-3}$	$0.0560 \cdot 10^{-3}$
DGP3	0.1328	plug-in	Random Forest	$5.2825 \cdot 10^{-3}$	$5.1558 \cdot 10^{-3}$	$0.1267 \cdot 10^{-3}$
			Boosting	$1.4637 \cdot 10^{-3}$	$1.3991 \cdot 10^{-3}$	$0.0646 \cdot 10^{-3}$
		de-biased	Random Forest	$3.8369 \cdot 10^{-3}$	$3.7326 \cdot 10^{-3}$	$0.1043 \cdot 10^{-3}$
			Boosting	$0.9312 \cdot 10^{-3}$	$0.8716 \cdot 10^{-3}$	$0.0596 \cdot 10^{-3}$

Table 1.1 report the results. First, observe the reduction in the population value of $\delta^*(\tilde{\tau})$ as heterogeneity increases in the DGP. This is entirely driven by an increase in the heterogeneity of $\tau(x)$ since the ATE is the same across the three DGPs. This means that a smaller shift in the covariates is required to invalidate the policy-maker claim ($ATE \geq 1.3$). As a result, the robustness metric decreases. Moving from DGP1 to DGP2 and DGP3 the population value of the robustness metric drops from 0.4485 to 0.1344 to 0.1328. The decrease is most accentuated between DGP1 and DGP2 because of the functional form of $\tau(x)$.

In DGP1 we can see that the heuristic choice for the hyper-parameters in boosting likely results in under-fitting the data, leading to a bias an order of magnitude higher than the variance. For DGP1, the de-biasing procedure results in approximately 20% squared bias reduction which drives the reduction of approximately the same percentage in the Mean Squared Error. Variances are comparable between plug-in and de-biased

GMM. The random forest procedure is better overall for MSE criterion. In DGP2, the bias dominate the variance component, suggesting both random forest and boosting are under-fitting. This is likely do to the absence of a *within-fold* cross-validation step. In this case,the de-biased GMM reduces the squared bias by about 40% for both random forest and boosting methods. The variances are again very similar across plug-in and de-biased and boosting has about half of the variance of random forest. DGP3's heterogeneity increases slightly, reducing the associated $\delta^*(\tilde{\tau})$. Like in DGP2, the bias dominates the variance component regardless of the first-step estimation method. Similarly, the de-biased GMM approach results in substantial bias reduction in comparison to the plug-in GMM approach.

1.4 Empirical Application: How robust are the effect of the Oregon Medicaid expansion?

In this section, I apply my approach to study the robustness of health insurance policy with respect to shifts in the distribution of covariates. The key reference is Finkelstein et al. [2012], which uses experimental data to study the effect of the Oregon Medicaid expansion lottery on health-care consumption and financial outcomes. The positive results of the study are of great interest for any policy-maker who is potentially interested in implementing a similar intervention in their state. Because the populations of recipients are likely to differ across states, I propose to complement the experimental result in Finkelstein et al. [2012] with an estimate of my robustness metric δ^* to quantify the smallest shift in important experimental covariates needed to eliminate the positive effects of the insurance lottery.

1.4.1 Institutional context and heterogeneity

Between March and September 2008, the state of Oregon conducted a series of lottery draws that would award the selected individuals the option to enroll in the Oregon Health Plan (OHP) Standard. OHP Standard is a Medicaid expansion program available for Oregon adult residents that are between 19 and 64 years of age and have limited income and assets. Finkelstein et al. [2012] studies the effect of the insurance coverage on a set of metrics that include health-care utilization (number of prescription, inpatient, outpatient and ER visits), recommended preventive care (cholesterol and diabetes blood test, mammogram and pap-smear test) and measures of financial strain (outstanding medical debt, denied care, borrow/default). The study uses both administrative and survey data but only the survey data is publicly accessible through Finkelstein [2013]. The Online appendix of Finkelstein et al. [2012] discusses a variety of robustness concerns that center on external validity. For example they note that scaling up the experiment can induce a supply side change in providers' behavior. Additionally, they acknowledge substantial demographic differences between the study population in Oregon *versus* the potential recipients in other states. These differences include, for example, a smaller African American and a larger white sub-population in Oregon versus other states. From the survey data it appears that the Oregon lottery participants are older and their health metrics under-performs the national average. If these covariates are important in determining the treatment effects of the health insurance, the results of Oregon experiment may not be robust to a change in the distribution of covariates. This robustness is especially important to quantify if the experimental results are to be extrapolated for policy adoption in other states. I stress the fact that, in this context, the re-weighting procedure in Hartman [2020] or Hsu et al. [2020] is not applicable because it lacks the survey-specific health data that are likely to be most predictive of treatment effect heterogeneity. Absent full covariate data form other states, I proposed to study the robustness of the policy by augmenting each of the treatment effect

estimators in Finkelstein et al. [2012] with my robustness metric, which can be computed by exploiting the heterogeneity in the publicly available survey data Finkelstein [2013].

1.4.2 Robustness in the Oregon Medicaid Experiment

For the robustness exercise I focus on the Intention to Treat Effect (ITT) of the Oregon Medicaid Experiment lottery. As noted in Finkelstein et al. [2012], not all recipients who were awarded the option to enroll in the insurance program actually enrolled. For this reason Finkelstein et al. [2012] estimates both an ITT and a LATE estimate. One could argue that the ITT is the key parameter for a policy-maker interested in offering the same intervention. To map my framework to the application, recall that the ITT effect can be considered as an ATE where the treatment D is simply the "the option to enroll in the health insurance" so the robustness approach discussed in the paper carries over to the ITT with only notational changes. I consider hypotheses of the form $ITT_j \geq \tilde{\tau}$ or $ITT_j \leq \tilde{\tau}$ (depending on the outcome measure of interest) where j indexes a health-care utilization or a financial strain outcome, following the notation convention in Finkelstein et al. [2012]. As noted in Finkelstein et al. [2012] all health-care utilization outcomes are defined consistently so that a positive sign for ITT means an increase in utilization. Similarly, all financial strain outcomes are defined so that a negative sign for the ITT means a decrease in financial strain. I focus on 2 value of interest for $\tilde{\tau}$ for each of the outcome measures. One of the values is $\tilde{\tau} = 0$ which reflects the claim that the ITT is non-negative (for health-care utilization outcomes), or non-positive (for financial strain outcomes). The second value is $\tilde{\tau} = t_j = z_\alpha \sigma_j$ where σ_j is the standard deviation of the ITT for outcome j. t_j is the critical value for the t-statistic of a one sided test with null hypothesis $ITT \leq 0$ for some pre-specified α. As a result $\delta(t_j)$ proxies for the magnitude of a change in the covariate distribution that would make the ITT statistically not distinguishable from a non-positive

or non-negative outcome (respectively).[11] Because σ_j is in general not available, in the empirical procedure I use $\hat{\sigma}_j$ in place of σ_j. The researcher interested in different hypothesis may adapt the procedure easily by specifying a $\tilde{\tau}$ with a value different from the two discussed above.

For the application I group the outcome measure in three groups: measure of health-care utilization, measures of compliance with recommended preventive care and measures of financial strain. I replicate the estimates of the intention to treat effect (ITT) for outcome variables in each of the three groups in Finkelstein et al. [2012] from a reduced form regression of the outcome variable on the lottery indicator and controls. The regression includes survey waves indicators, household size indicators and interaction terms between the two as controls. Because the regression is fully saturated, the estimates for the ITT are nonparametric. In my robustness exercise I focus on covariates that appear critical for external validity and are likely to differ across states. Among others, Finkelstein et al. [2012] identifies gender, age, race, credit access, education and proxies for health status. To capture the potential heterogeneity, I estimate a Conditional Intention to Treat effect (CITT) with the set of covariates listed above.[12] Finally I use the estimated CITT to compute the measure of robustness δ^* for each of the outcome variables in the three categories and report it, together with the original ITT estimate, for both values of $\tilde{\tau}$ discussed above.[13] All outcomes are measured on the survey data Finkelstein [2013].

In Table 1.2, column 2, 3 and 4 contain respectively the experimental ITT for each outcome variable, the estimates for $\delta^*(0)$ and the estimates for $\delta^*(t_j)$. Here $t_j = \pm 1.645 \cdot \sigma_j$

[11] This interpretation is heuristic, in the sense that the standard deviation of the ITT estimate can depend on the distribution of the covariates as well. It is possible to impose an additional constraint on optimization problem, requiring that the variance of the treatment effects about the ITT remains the same. Such a construction fall into the case discussed in Appendix 1.C.

[12] From a technical standpoint, the CITT estimated with a discrete set of covariates is still a parametric estimator. In practice, it can be obtained by a fully saturated regression where the lottery indicator is interacted with all possible combinations of dummies.

[13] Comparable (survey weighted) ITT estimates can be found in column 2 labelled Reduced form, of 1.2. Discrepancies with the (unweighted) ITT effects I compute are due to survey weights.

depending on whether the experimental ITT is positive or negative. As an example, consider a measure of financial strain, like whether a patient had to borrow or skip a payment because of medical debt. The intention to treat effect is equal to -0.0515 with standard error 0.0060. $\delta^*(0) = 0.367$ represents the smallest distributional shift of the covariates that can induce an ITT equal to 0. The $\delta^*(t_j) = 0.265$ represents the smallest distributional shift in the covariates that can result in an $ITT = -1.645 \cdot 0.0060 = -0.0118$ which leads to not rejecting the hypothesis $H_0 : ITT \geq 0$. For any distributional shift that is smaller than $\delta^*(t_j)$ the statistical claim $H_0 : ITT \geq 0$ would be rejected.

Table 1.2: δ^* robustness metric for the health-care utilization and financial strain outcomes in Finkelstein et al. [2012]. ITT for measures of preventive care are indistinguishable from 0 for the experimental distribution so the robustness metric is trivial in this case. The measure is evaluated at $\tilde{\tau} = 0$ and $\tilde{\tau} = t_j = \pm 1.645\sigma_j$ for each outcome, depending on the relevant sign of the estimated ITT. The third group of outcomes, preventive care measures, all have statistically insignificant ITT, leading to a 0 robustness for all $\delta^*(t_j)$. I omit them in this table.

Outcome	Experimental ATE	$\delta^*(0)$	$\delta^*(t_j)$
health-care Utilization			
Prescriptions	0.1296 (0.044)	0.380 (0.007)	0.068 (0.002)
Out-patient visits	0.2986 (0.039)	1.552 (0.022)	0.965 (0.014)
ER visits	0.0064 (0.013)	0.009 (0.001)	0 n/a
In-patient visits	0.0081 (0.005)	0.119 (0.003)	0 n/a
Financial Strain			
Out of pocket expenses	−0.0622 (0.0069)	0.462 (0.030)	0.346 (0.023)
Outstanding expenses	−0.0529 (0.0070)	0.290 (0.0231)	0.204 (0.016)
Borrow/Skip payments	−0.0515 (0.0060)	0.367 (0.019)	0.265 (0.014)
Refused care	−0.011 (0.0040)	0.063 (0.006)	0.013 (0.002)

I highlight two benefits of this robustness metric. First, it allows a comparison of the robustness across outcomes because each δ^* has the same units and it is measured on the same covariate space. Second, the fourth column of Table 1.2 has a natural interpretation

as a breakdown point: what is the smallest perturbation of the distribution of covariates that will break statistical significance of the ITT? A policy-maker may consider findings with larger δ^* as more readily applicable to her own policy setting. From the δ metrics reported in Table 1.2 I notice that among the health-care utilization metrics, the ITT on outpatient visits is the most robust while the ITT on ER visits is the least robust. For the measures of financial strain the ITT on out of pocket expenses is the most robust and the ITT on instances of refused care because of medical debt is the least robust. If one had access to census data, one could choose a set of census variables of interest and compute the KL divergence between the distribution of the Oregon census variables and a target state's census variables. Then the researcher use this computed measure to benchmark the magnitude of the robustness metrics in Table 1 to assess whether the magnitude of each δ^* is high or low, relative to the observed differences in the census variables.

1.5 Conclusion

Robustness of (quasi)-experimental findings is an importance premise of evidence based policy-making. In this paper I propose a metric δ^* to quantify the robustness of (quasi)-experimental findings with respect to a shifts in the distribution of the covariates. I focus on claims on aggregate policy effects of the type ($ATE \geq \tilde{\tau}$). While I focus on ATE as a main object of interest, the extension to other linear policy parameters is straightforward. I characterize my robustness metric as the minimal distance, in terms of KL divergence, between the set of covariate distributions that invalidate the claim and the (quasi)-experimental covariates. My robustness metric gives a nonparametric, one-dimensional summary that links treatment effect heterogeneity, (quasi)-experimental findings and covariate shifts. Because the computation of the δ^* robustness metric for ATE requires computing CATE, I employ the debiased-GMM approach to allow for CATE to

be estimated using a large collection of machine learning techniques, which only need to satisfy mild requirements on their L^2 norm convergence rates. These include, for example, lasso, random forest, boosting, neural nets.

I apply my framework to assess the robustness of the results in Finkelstein et al. [2012] about the Oregon Medicaid Experiment. I consider a set of covariates including gender, race and lottery timing and find that the increase in outpatient visits and the decrease in out-of-pocket expenses are, respectively the most robust findings among the measure of health-care utilization and financial strain. For most other measures, relatively small shifts in the covariate distributions appear to invalidate the results.

1.6 Acknowledgements

Chapter 1, in full, is currently being prepared for submission for publication of the material. The dissertation author was the sole author of this material.

Appendix

1.A Another look at the Lagrange multiplier λ

The formulation of the optimization problem in Equation (1.4) concerns a policy-maker who wishes to maintain the claim $ATE \geq \tilde{\tau}$ so that the constraint set in Equation (1.5) takes the opposite direction of the inequality. The formulation with the Lagrange multiplier in Equation (1.9) is without loss of generality. If the policy-maker is interested in maintaining a claim of the type $ATE \leq \tilde{\tau}$, the Lagrange multiplier would enter Equation (1.9) with a negative sign, or equivalently, if we want to preserve Equation (1.9), the value of the Lagrange multiplier would be negative.

The Lagrange multiplier λ in Equation (1.9) can give insight in what happens moving from the experimental distribution to the *least favorable distribution*. Note that λ has the opposite sign as the difference between the (quasi)-experimental ATE and the target ATE. To see this, we consider how the target ATE relates to the CATE. For each given $\tilde{\tau}$ there is a partition of the covariate support $\mathcal{X}$ into three sets, depending on what will be down-weighted or up-weighted by the *least favorable distribution*. The weight is given by:

$$w(x) = \frac{\exp(-\lambda(\tau(x) - \tilde{\tau}))}{\int_{\mathcal{X}} \exp(-\lambda(\tau(x) - \tilde{\tau}))dF_X}$$

so we see, after simplifying, that $w(x) = 1$, i.e dF_X^* and dF_X coincide, iff:

$$\exp(-\lambda\tau(x)) = \int_{\mathcal{X}} \exp(-\lambda\tau(x)) dF_X$$

so the three sets are given by:

$$\mathcal{X}^- := \{x \in \mathcal{X} \ s.t. \ \exp(-\lambda\tau(x)) - \mathbb{E}_{P_X}[\exp(-\lambda\tau(x))] < 0\}$$
$$\mathcal{X}^+ := \{x \in \mathcal{X} \ s.t. \ \exp(-\lambda\tau(x)) - \mathbb{E}_{P_X}[\exp(-\lambda\tau(x))] > 0\}$$
$$\mathcal{X}^0 := \{x \in \mathcal{X} \ s.t. \ \exp(-\lambda\tau(x)) - \mathbb{E}_{P_X}[\exp(-\lambda\tau(x))] = 0\}$$

For example, suppose that the researcher wants to support the claim $ATE \geq \tilde{\tau}$, which holds for the *experimental* ATE. Then, in order to achieve a lower ATE the *least favorable distribution* will have to shift weight from $\mathcal{X}^+$ to $\mathcal{X}^-$. These sets in the partition will in general not coincide with the sets $\{x \in \mathcal{X} \ s.t. \ \tau(x) - \tilde{\tau} < 0\}$, $\{x \in \mathcal{X} \ s.t. \ \tau(x) = \tilde{\tau}\}$ and $\{x \in \mathcal{X} \ s.t. \ \tau(x) < \tilde{\tau}\}$. One case when they coincide is when F_X follows the normal distribution.

1.B Relating parametric forms of least favorable distributions with assumptions on CATE

Lemma 7 gives a solution to the policy-maker's problem that does not depend on a specific functional form for CATE nor on a parametric assumption for the experimental distribution F_X. Leveraging the closed form solution I show that if the conditional treatment effect function does follow a particular form and the experimental distribution belongs to a certain parametric family, we can guarantee that the *least favorable distribution* belongs to the same parametric family, up to a shift in the parameters.

Definition 18. *We say that a class of parametric distributions indexed by θ, denoted F_X^θ is* ***least-favorable closed*** *with respect to a parametric class of Conditional Average Treatment Effects, $\tau(x)_\eta$, indexed by $\eta \in H$ if for any θ and η, the least favorable distribution $F_X^* = F_X^{\theta^*}$ for some $\theta^* \in \Theta$. The choice of θ^* will in general also depend on features of η as well.*

This means that the *least favorable distribution* belongs to the same parametric class as the original, experimental distribution. This idea is similar to the conjugate prior construction where the posterior distribution belongs to the same class of priors if the likelihood is within a conjugate parametric class. The distributional shift can then be thought of as a parameter shift.

Proposition 19 (Quadratic-Normal least favorable closed-ness)**.** *The parametric class $\mathcal{N}(\mu, \sigma^2)$ is* ***least favorable closed*** *for quadratic Conditional Average Treatment Effects. That is, if $X \in \mathbb{R}^k$ follows the multivariate normal distribution $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is p.d. and $\tau(x) = \boldsymbol{x^T A x} + \boldsymbol{x^T \beta} + c$ for $\boldsymbol{\beta} \in \mathbb{R}^k$ then F_X^* is the measure induced by $X^* \sim \mathcal{N}(\boldsymbol{\mu^*}, \boldsymbol{\Sigma^*})$ with $\boldsymbol{\mu}^* = (\boldsymbol{\Sigma}^{-1} + 2\lambda \boldsymbol{A})^{-1}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \lambda \boldsymbol{\beta})$ and $\boldsymbol{\Sigma^*} = (\boldsymbol{\Sigma}^{-1} + 2\lambda \boldsymbol{A})^{-1}$, provided that $(\boldsymbol{\Sigma}^{-1} + 2\lambda \boldsymbol{A})^{-1}$ is p.d. The parameter λ is defined as in Equation* (1.9)*.*

Proof. See Appendix 1.I. □

Corollary 20 (Linear-Normal least favorable closed-ness)**.** *If $\tau(x) = \boldsymbol{x^T \beta}$ and $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then $X^* \sim \mathcal{N}(\boldsymbol{\mu^*}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu^*} = \boldsymbol{\mu} - \lambda \boldsymbol{\Sigma \beta}$.*

Proof. Follows from 19 when $\boldsymbol{A} = 0$. □

An extension of Proposition 19 could be shown to hold for the more general class of distributions in the exponential family given by $f(x|\theta) = g(\theta)h(x)\exp(\eta(\theta)^T T(x))$ but this is beyond the scope of this paper. The parametric example gives some additional insights in the geometry of the policy-maker's problem, which could also help to understand the analytical

expression for the least favorable distribution above. Consider the univariate case ($d = 1$) where F_X is the normal distribution with mean μ and standard deviation σ. The policy-maker's desired claim is $ATE \geq 0$. The conditional average treatment effects are linear in the only covariate, that is $\tau(x) = \pi X$ for some $\pi \in \mathbb{R}$. Because CATE is linear in X, the ATE is only a function of the population mean μ. As a result, the feasible set of the policy-maker's problem in Figure 1.B.1 is the half space $\mu \leq 0$. Proposition 19 allows us to reduce the problem to a finite dimensional problem which we can solve with the usual KKT conditions. Observe that $D_{KL}(\mathcal{N}(\mu^*, \sigma^*)||\mathcal{N}(\mu, \sigma)) = \frac{1}{2}\left(\log\left(\frac{\sigma^2}{\sigma^{*2}}\right) + \frac{\sigma^{*2}}{\sigma^2} - 1 + \frac{1}{\sigma^2} \cdot (\mu - \mu^*)^2\right)$. In that case:

$$\begin{aligned} \min_{(\mu^*, \sigma^*) \in \mathbb{R} \times \mathbb{R}_+} \quad & \frac{1}{2}\left(\log\left(\frac{\sigma^2}{\sigma^{*2}}\right) + \frac{\sigma^{*2}}{\sigma^2} - 1 + \frac{1}{\sigma^2} \cdot (\mu - \mu^*)^2\right) \\ s.t. \quad & \pi\mu^* \leq \tilde{\tau} \end{aligned}$$

where the constraint is simplified because of the linear functional form of CATE and linearity of the expectation operator.

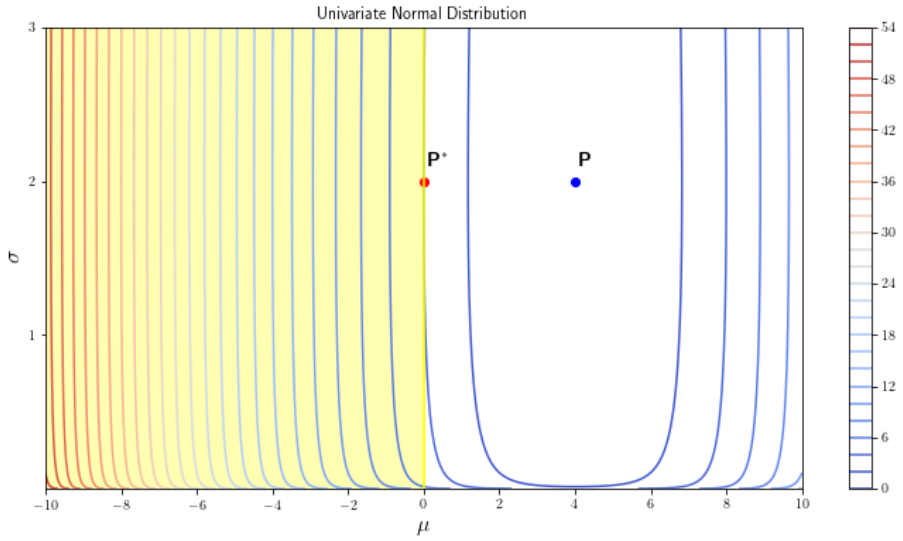


Figure 1.B.1: Univariate Normal Distribution, Linear CATE. Each point in the graph represents a normal distribution parametrized by its mean and standard deviation $\mathcal{N}(\mu, \sigma^2)$. The starting distribution, the experimental is taken to be $\mathrm{P} = \mathcal{N}(4,2)$. The contour lines represent the KL divergence with respect to the experimental distribution. The policy-maker's desired claim is $ATE \geq 0$. The feasible set shaded in yellow represents all univariate normal distributions that satisfy $ATE \leq 0$. When CATE is linear (that is $\tau(x) = \pi X$), the only parameter that contributes to the ATE is the mean μ so the feasible set is parallel to the σ axis. As a result, the *least favorable* distribution, labelled as P^*, amounts to a mean shift from $\mu = 4$ to $\mu^* = \frac{\tilde{\tau}}{\pi} = 0$ and no shift in the σ parameter.

The KKT conditions imply:

$$\mu^* = \mu - \lambda \pi \sigma^2$$

$$\sigma^* = \sigma$$

$$\lambda = \frac{1}{\pi \sigma^2} \left(\mu - \frac{\tilde{\tau}}{\pi} \right)$$

The *least favorable distribution* amounts to a mean shift of the prescribed magnitude and no change in the variance. Contrast the example above with the case where the CATE is allowed to be quadratic. Proposition 19 still applies, hence the problem can still be formulated as minimizing over the paramteric space (μ^*, σ^*). This time though, the variance

of the covariate X matters in determining the ATE and the feasible set reflects this.

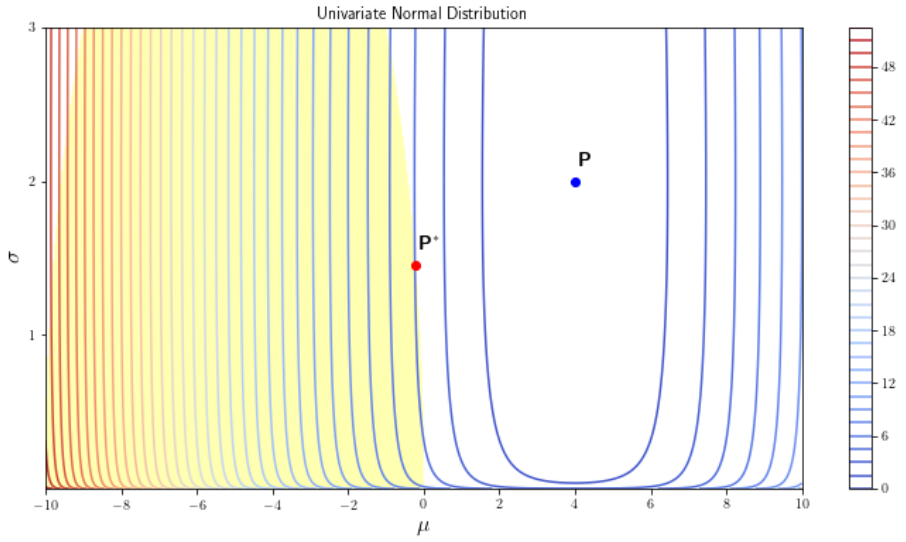


Figure 1.B.2: Univariate Normal Distribution: Quadratic CATE. The setting is identical as in Figure 1.B.1 but here is quadratic, $\tau(x) = 0.8 \cdot X^2 + 8 \cdot X$. As a result, $ATE(\mu, \sigma) = 0.8 \cdot (\mu^2 + \sigma^2) + 8\mu$ so both parameters of the covariate distribution contribute to determining the ATE. The feasible set in yellow has nonflat curvature. The *least favorable distribution*, labelled as P*, features a parameter shift in both the mean and the variance.

1.C Constrained Classes

An applied researcher may wish to restrict the class of distributions for problem 1 by imposing additional constraints. For example, they may want to fix the certain moments of the experimental distribution.[14] The computational price to pay for each additional constraint is one additional Lagrange multiplier per constraint as detailed out in Ho [2020].

[14]Note that finitely many moment restrictions would still amount to searching the KL infimum within a infinite dimensional class of probability distributions, and, as such, the nonparametric nature of the problem persists.

For example, for a known moment function $q : \mathcal{X} \to \mathbb{R}$ we want:

$$\int_{\mathcal{X}} q(X) dF_X = \int_{\mathcal{X}} q(X) dF'_X$$

This requirement restricts the space of feasible probability distributions because it asks that the *least favorable distribution* preserves the additional moment. From the perspective of robustness, the value of the problem δ^* for the constrained problem must be larger or equal than the value for the unconstrained problem. That is:

$$\begin{aligned} \inf_{dF'_X:\ dF'_X \ll dF_X; P'_X(\mathcal{X})=1} & D_{KL}(F'_X||F_X) \qquad \leq \inf_{F'_X:\ F'_X \ll F_X; P'_X(\mathcal{X})=1} D_{KL}(F'_X||F_X) \\ s.t. & \int_{\mathcal{X}} \tau(x) dF'_X(x) \leq \tilde{\tau} \qquad \int_{\mathcal{X}} \tau(x) dF'_X(x) \leq \tilde{\tau} \\ & \int_{\mathcal{X}} q(x) dF'_X(x) = q \end{aligned}$$

Assumptions about moment preservation aides the robustness to external validity of a causal claim.

In case of additional constraints the solution to the KL problem takes the form:

$$\frac{dF^*_X}{dF_X} = \frac{\exp(-\lambda(\tau(x) - \tilde{\tau})) \prod_{l=1}^{L} \exp(-\mu_l(q(x) - \tilde{q}))}{\int_{\mathcal{X}} \exp(-\lambda(\tau(x) - \tilde{\tau})) \prod_{l=1}^{L} \exp(-\mu_l(q(x) - \tilde{q}))}$$

and each Lagrange multiplier can be solved by:

$$\int_{\mathcal{X}} \exp(-\mu_l(q(x) - \tilde{q}))(q(x) - \tilde{q}) dF_X = 0$$

For estimation, the additional restrictions result in L many additional parameters, one for each Lagrange multiplier that needs to be computed. One could adapt the estimation framework in Section 1.3 and have $\theta \in \Theta \subseteq \mathbb{R}^{L+2}$ gathers the original parameters α and λ as well as the Lagrange multipliers for the population optimization problem $\mu_1, \mu_2, \cdots, \mu_L$.

At the cost of a more cumbersome notation, all the asymptotic results in Section 1.3 apply.

1.D Partial identification of CATE

In this section, I consider the case where the main ingredient needed to identify the robustness metric, $\tau(x)$ is only partially identified. This situation is important in practice. For example, with one-sided noncompliance $\tau(x)$ is only partially identified. In this section I will show that one can still recover bounds for $\delta^*(\tilde{\tau})$ that are robust to this partial identification.

In section 1.2.2, the covariate shift assumption allowed us to write the ATE as a linear functional of the covariate distribution, greatly simplifying the treatment. This linear functional is fixed because $\tau(x)$ is identifiable.

Suppose we can set identify $\tau \in \mathcal{T}$. For example $\tau(x)$ could be identified up to a finite dimensional parameter or one could have an identification region where any $\tau \in \tau$ satisfies $\underline{\tau}(x) \leq \tau(x) \leq \overline{\tau}(x)$, that is, there are identification bands bounding any $\tau \in \mathcal{T}$ above and below. Then we can compute a conservative version of the robustness metric define below:

$$\underline{\delta}^*(\tilde{\tau}) := \inf_{\tau \in \mathcal{T}} \inf_{dF'_X:\ dF'_X \ll dF_X; dF'_X(\mathcal{X})=1} D_{KL}(F'_X||F_X)$$
$$s.t.\ \int_{\mathcal{X}} \tau(x) dF'_X(x) \leq \tilde{\tau}$$

Because now $\tau(\cdot)$ is not identified, the problem above considers the least favorable among the ones in the set $\mathcal{T}$. Because τ controls the shape of the feasible set we can rewrite it as

$$\underline{\delta}^*(\tilde{\tau}) := \inf_{dF'_X:\ dF'_X \ll dF_X; dF'_X(\mathcal{X})=1} D_{KL}(F'_X||F_X)$$
$$s.t.\ \int_{\mathcal{X}} \tau(x) dF'_X(x) \leq \tilde{\tau} \text{ for some } \tau \in \mathcal{T}$$

Now regard the constraint set as a collection of $\mathcal{F}_\tau := \{F'_X : \int_{\mathcal{X}} \tau(x) dF'_X(x) \leq \tilde{\tau}\}$ for a given τ. It is immediate to notice that, if $\tau'(x) \leq \tau(x)$ point-wise, then $\mathcal{F}_\tau \subseteq \mathcal{F}_{\tau'}$. That is, if a CATE that is dominated point-wise (or in fact F_X almost everywhere) the constraint set admits a larger class of distributions. As a result, for $\underline{\tau}$ we have, for any $\tau \in \mathcal{T}$, $\mathcal{F}_\tau \subseteq \mathcal{F}_{\underline{\tau}}$. But this greatly simplifies the problem since now it is enough to write:

$$\underline{\delta}^*(\tilde{\tau}) := \inf_{dF'_X:\ dF'_X \ll dF_X; dF'_X(\mathcal{X})=1} D_{KL}(F'_X||F_X)$$
$$s.t.\ \int_{\mathcal{X}} \underline{\tau}(x) dF'_X(x) \leq \tilde{\tau}$$

so now the problem can be solved for the lower bound of the identified set. Again, this interpretation of delta amounts to considering robustness to the lack of identification the CATE. A similar argument applies for the reverse inequality ($ATE \leq \tilde{\tau}$) and $\overline{\tau}$.

1.E Re-evaluating policies over time

In the main paper, the policy-maker is concerned with extrapolating experimental results to different policy contexts. In the application, this takes the form of extrapolating the Medicaid extension program to other states. In this section I show that we can have an alternative interpretation that emphasizes changes over time rather than across regions. According to this interpretation, the measure of robustness δ^* captures the minimal change in demographic trends that is needed to invalidate a particular policy conclusion.

Consider a time horizon $t = -1, 0, 1, 2, \cdots, T$. Suppose that a policy is implemented at time 0. For the covariate distribution at time 0, $F_{X,0}$ the policy meets the target $\tilde{\tau}$, that is, $ATE_{F_{X,0}} \geq \tilde{\tau}$. Now, we may worry that over time, the covariate distribution might change from F_0 in such a way that does not justify the policy any longer.

How does the covariate shift assumption translate in thus context? It requires that the causal effect $\tau_{F_{X,0}}(\cdot) = \tau_{F_{X,t}}(\cdot)$ for all $t = 1, 2, \cdots, T$. That is, the CATE for whichever time horizon it is defined, does not change for new cohorts who are newly treated.

Here, a natural benchmark for comparison is given by the change between the reference point and the pre-policy period $t = -1$. This benchmark is given by $\delta_{benchmark} = D_{KL}(F_{X,-1}||F_{X,0})$. In this case, if one finds $\delta^*(\tau) > \delta_{benchmark}$ then the policy-maker may be comforted by observing that the amount of variation needed to invalidate the claim is larger than the natural variation that can be elicited from the time trends. Of course, one could decide to formalize this notion since we could seek to jointly characterize the asymptotic distribution of the vector of estimators $(\hat{\delta}^*(\tilde{\tau}), \hat{\delta}_{benchmark})^T$ which is beyond the scope of this paper.

1.F An interpretation of the robustness metric based on Sanov's theorem

In this section, I discuss some further details on the interpretation of $\delta^*(\tau)$ based on Sanov's theorem the I have introduced in Section 1.F. The treatment in this section will be restricted to a covariate space $\mathcal{X}$ that is supported on finitely many points, reflecting the discussion of the method of *types* in Cover [1999]. Suppose there are k covariates, $X_1, \cdots, X_k$, each taking values in $\mathcal{X}_j$ with $|\mathcal{X}_j|$ finite. Let $m := \sum_{j=1}^{k} |\mathcal{X}_j|$. The set of probability distributions on $\mathcal{X} = \Pi_{j=1}^{k} \mathcal{X}_j$ can be identified with the $(D-1)$-dimensional simplex in $\mathbb{R}^m$. For a fixed sample size n consider the set of all sequences of data

$x = (x_1, x_2, \cdots, x_n)$ taking values in $\mathcal{X}^n$ and define the *type* P_x of a sequence x as the relative proportion of each possible realization a in $\mathcal{X}$, that is, $P_x(a) = \frac{N(a|x)}{n}$ where $N(a|x)$ is the number of times realization a shows up is in sequence x. Let $\mathcal{P}_n$ be the collection of types. Cover [1999] version of Sanov theorem allows for a more general set E, not necessarily convex at the price of an additional multiplicative polynomial term $(n+1)^m$ in the number of observation. If the set E is a convex set, the upper bound can be tightened to $P^n(E \cap \mathcal{P}_n) \leq e^{-nD_{KL}(P^*||P)}$ and the polynomial term in n is dropped. Note that $E := \{Q : \int_{\mathcal{X}} \tau(x) dQ \leq \tilde{\tau}\}$ is obtained through imposing a linear restriction on Q and therefore E is convex. Sanov's theorem remains true for larger classes of probability distributions, not necessarily confined to finitely supported X like discussed in Csiszár [1984] but the method of types leans itself for a discussion on discrete spaces.

1.G Some additional results

Proposition 21. *Let $\epsilon > 0$. Then for $\tilde{\tau} > \inf_{\mathcal{X}} \tau(x) + \epsilon$, $\delta^*(\tilde{\tau})$ in Definition 3 is decreasing in $\tilde{\tau}$.*

Proof. First denote the feasible set $E(\tilde{\tau}) := \{F_X \in \mathcal{F} : \int_{\mathcal{X}} \tau(x) dF_X(x) \leq \tilde{\tau}\}$. Then, $G_X \in E(\tilde{\tau}) \iff \int_{\mathcal{X}} \tau(x) dF_X(x) \leq \tilde{\tau} < \tilde{\tau}'$ for any $\tilde{\tau}' > \tilde{\tau}$ so $G_X \in E(\tilde{\tau}')$. But then $E(\tilde{\tau}) \subseteq E(\tilde{\tau}')$. Hence, because we are minimizing on a larger set of distributions $\delta^*(\tilde{\tau}) := \inf_{G_X \in E(\tilde{\tau})} D_{KL}(G_X||F_X) \geq \inf_{G_X \in E(\tilde{\tau}')} D_{KL}(G_X||F_X) =: \delta^*(\tilde{\tau}')$. If the feasible set E has the reverse inequality, it follows immediately that $\delta^*(\tau)$ is increasing in $\tilde{\tau}$.

□

1.H General φ-divergence metrics and *least favorable* closed classes.

In this section I extend the theory of least favorable classes by considering different φ divergence measures. To this end I leverage the thorough treatment of φ divergences in Christensen and Connault [2019]. The Kullback-Leibler divergence is a special case of a more general construction, known as φ-divergence. It is introduced below:

Definition 22 (φ-divergence)**.** *Consider the φ-divergence between F_X and F'_X given by:*

$$D_\varphi(F'_X||F_X) := \int \varphi\left(\frac{dF'_X}{dF_X}\right) dF_X \tag{A1.19}$$

where φ is a convex function with $\varphi(1) = 0$ and $\frac{dF'_X}{dF_X}$ is the Radon-Nikodym derivative of the probability distribution F'_X with respect to the probability distribution of F_X, provided that $P'_X \ll P_X$ for the respective measures. For example the choices $\varphi(t) = t\log t$ and $\varphi(t) = \frac{1}{2}|t-1|$ give rise to the KL-divergence and to the total variation divergence (TV) respectively.

There may be a reason to choose a different φ-divergence metric instead of the KL-divergence. Under suitable conditions, the construction of the proposed robustness metric will change in magnitude, since now the (pseduo)-metric on the space of distributions of the covariates is different. A closed form solution analogous to Lemma 7 is available. The characterization of the δ^* now depends on $\varphi(\cdot)$. In particular it is fully characterized in terms of the Fenchel-conjugate of φ and its derivative.

Definition 23 (Fenchel-Conjugate)**.** *Given a topological vector space X and convex function $\varphi : X \to \mathbb{R}$, the Fenchel-conjugate $\varphi^* : X^* \to \mathbb{R}$, defined on the dual space of X, is defined*

by:

$$\varphi^* : x^* \mapsto \sup_{x \in X} \langle x^*, x \rangle - \varphi(x) \tag{A1.20}$$

Then we can have a generalization of the policy-maker's problem in Equations (1.4) and (1.5) for an arbitrary φ divergence in 22:

$$\inf_{P'_X:\ P'_X \ll P_X; P'_X(\mathcal{X})=1} D_\varphi(F'_X || F_X) \tag{A1.21}$$

$$s.t.\ \int_{\mathcal{X}} \tau(x) dF'_X(x) \leq \tilde{\tau} \tag{A1.22}$$

From the KKT Theorem (Theorem 1, Ch.8, Sec. 3 in Luenberger [1997]) we can write the problem as:

$$\sup_{\lambda \in \Lambda} \sup_{\xi} \left(\inf_{P'_X:\ P'_X \ll P_X; P'_X(\mathcal{X})=1} D_\varphi(F'_X || F_X) + \lambda \int_{\mathcal{X}} (\tau(x) - \tilde{\tau}) dF'_X(x) + \xi \left(\int_{\mathcal{X}} dF'_X - 1 \right) \right) \tag{A1.23}$$

where and ξ is the Lagrange multiplier for integration to unity, λ is the Lagrange multiplier for the policy-maker's claim. The convexity conditions for Theorem 1, Ch.8, Sec. 3 in Luenberger [1997] are immediate to verify. The interior condition, analogous to a Slater condition, is satisfied by Assumption 4. Note that convex cone where the Lagrange multiplier takes values is $\mathbb{R}_+$ (or $\mathbb{R}_-$ if the policy-maker's claim is $ATE \leq \tilde{\tau}$ instead). In Equation (1.9) the Lagrange multiplier λ is a 1-dimensional parameter. Notice that after fixing the experimental distribution, $D_{KL}(\cdot || F_X)$ is convex in its first argument. With a careful rewriting we can express the inner problem as:

$$\inf_{P'_X:\ P'_X \ll P_X; P'_X(\mathcal{X})=1} \int_{\mathcal{X}} \left(\varphi\left(\frac{dF'_X}{dF_X}(x) \right) - (-\lambda(\tau(x) - \tilde{\tau}) - \xi) \frac{dF'_X}{dF_X}(x) \right) dF_X(x) - \xi$$

and recognize that, if we can pass the infimum under the integral sign, we can substitute the expression for the Fenchel-conjugate of φ, switching the sign of the infimum.

$$\begin{aligned}&\inf_{P'_X:\ P'_X \ll P_X; P'_X(\mathcal{X})=1} \int_{\mathcal{X}} \left(\varphi\left(\frac{dF'_X}{dF_X}(x) \right) - (-\lambda(\tau(x)-\tilde{\tau})-\xi) \frac{dF'_X}{dF_X}(x) \right) dF_X(x) - \xi \\ &= -\int_{\mathcal{X}} \varphi^*(-\lambda(\tau(x)-\tilde{\tau})-\xi) dF_X(x) - \xi\end{aligned}$$

Substituting this back into the outside problem one obtains:

$$\sup_{\lambda \in \Lambda} \sup_{\xi} \int_{\mathcal{X}} -\varphi^*(-\lambda(\tau(x)-\tilde{\tau})-\xi)) dF_X(x) - \xi$$

which can be maximized with respect to ξ and delivers the first order condition, evaluated at ξ^*:

$$\int_{\mathcal{X}} \dot{\varphi}^*(-\lambda(\tau(x)-\tilde{\tau})-\xi^*) dF_X = 1 \tag{A1.24}$$

where $\dot{\varphi}^*(\cdot)$ is the derivative of $\varphi^*(\cdot)$ with respect to its argument. Observe that the Fenchel-conjugate of $\varphi(t) = t\log(t)$ is given by $\varphi(t^*) = \exp(t^*-1)$. Solving for ξ^* here delivers:

$$\xi^* = \log\left(\int_{\mathcal{X}} \exp(-\lambda(\tau(x)-\tilde{\tau}-1)) dF_X(x) \right)$$

Now differentiating with respect to λ we obtain

$$\int_{\mathcal{X}} \dot{\varphi}^*(-\lambda^*(\tau(x)-\tilde{\tau})-\xi^*)(\tau(x)-\tilde{\tau}+\dot{\xi}^*_\lambda) dF_X(x) - \dot{\xi}^*_\lambda = 0 \tag{A1.25}$$

where $\dot{\xi}^*_\lambda$ is the derivative of ξ^* with respect to λ and λ^* is the value that implicitely solves the moment condition in Equation (A1.25). Observe that plugging Equation (A1.24) into Equation (A1.25) allows to simplify it to:

$$\int_{\mathcal{X}} \dot{\varphi}^*(-\lambda^*(\tau(x)-\tilde{\tau})-\xi^*)(\tau(x)-\tilde{\tau}) dF_X = 0$$

since the two terms in $\dot{\xi}^*_\lambda$ cancel out. Moreover, if $\varphi(\cdot)$ is the KL divergence like in the main body of the paper, then

$$\int_{\mathcal{X}} \dot{\varphi}^*(-\lambda^*(\tau(x)-\tilde{\tau})(\tau(x)-\tilde{\tau})dF_X \cdot \exp(-\xi^*) = 0$$

so the additional term $\exp(-\xi^*) > 0$ can be dropped and Equation (A1.25) recovers Equation (1.9).

1.I Proofs

First I introduce a few basic results for optimization problems like the one in Equations (1.4-1.5). Consider the set of probability distributions on $\mathcal{X}$, $\mathcal{P}_X := \{P_X : \int_{\mathcal{X}} dP_X = 1\}$. Under the L_1 norm, $\mathcal{P}_X$ is a complete metric space and it is convex. Namely, if $P_1, P_2 \in \mathcal{P}_X$ then $P_\alpha = \alpha P_1 + (1-\alpha)P_2 \in \mathcal{P}_X$ is a mixture distribution. Moreover, if there is a dominating measure μ such that $f_1 = \frac{dP_1}{d\mu}$ and $f_2 = \frac{dP_2}{d\mu}$ are the Radon-Nikodym derivatives then $\frac{dP_\alpha}{d\mu} = \alpha f_1 + (1-\alpha) f_2$. Now consider the constraint given in Equation (1.5). For any two P_1 and P_2 that satisfy the constraint, P_α for any $\alpha \in [0,1]$ will satisfy it as well. Hence the constraint set given by Equation (1.5) is a convex subset of $\mathcal{P}_X$. If such a set is non-empty, then, because $D_{KL}(\cdot||F_X)$ is a strictly convex function on a convex set, the infimization problem in Equation (1.4) has a unique solution (P_X-almost everywhere) and the infimum is achieved. Lemma 7 characterizes such a solution P_X-almost everywhere.

1.I.1 Proof of Lemma 7

The proof is based on a result that appeared first in Donsker and Varadhan [1975]. More recently Ho [2020] has used a similar argument to characterize global sensitivity in a Bayesian setting.

Lemma 7 (Closed form solution)**.** *Let Assumptions 1, 2, 3 and 4 hold. Then: i) The infimum in Equation* (1.4) *is achieved. Moreover F_X^*, is characterized, P_X-almost everywhere, by:*

$$\frac{dF_X^*}{dF_X}(x) = \frac{\exp(-\lambda(\tau(x)-\tilde{\tau}))}{\int_{\mathcal{X}} \exp(-\lambda(\tau(x)-\tilde{\tau}))dF_X(x)} \tag{1.8}$$

where $\frac{dF_X^*}{dF_X}$ is the Radon-Nikodym derivative of dF_X^* with respect to dF_X and λ is the Lagrange multiplier implicitly defined by the equation:

$$\int_{\mathcal{X}} \exp(-\lambda(\tau(x)-\tilde{\tau}))(\tau(x)-\tilde{\tau})dF_X(x) = 0 \tag{1.9}$$

ii) The value of the robustness metric $\delta^*(\tilde{\tau})$ is given by:

$$\delta^*(\tilde{\tau}) = D_{KL}(F_X^*||F_X) = -\log\left(\int_{\mathcal{X}} \exp(-\lambda(\tau(x)-\tilde{\tau}))dF_X(x)\right) \tag{1.10}$$

First note that, by the Radon-Nikodym theorem, $\frac{dF_X^*}{dF_X}$ exists and $\text{supp}\left(\frac{dF'_X}{dF_X}\right) \subset \mathcal{X}$. Recall $\tau(x) = \mathbb{E}[Y_1|X=x] - \mathbb{E}[Y_0|X=x]$. Then:

$$\inf_{F'_X:\ P'_X \ll P_X; P'_X(\mathcal{X})=1} D_{KL}(F'_X||F_X)$$
$$s.t.\ \int_{\mathcal{X}} \tau(x)dF'_X(x) = \tilde{\tau}$$

is equivalent to:

$$\inf_{F'_X:\ P'_X \ll P_X} D_{KL}(F'_X||F_X)$$
$$s.t.\ \int_{\mathcal{X}} \tau(x)\frac{dF'_X}{dF_X}dF_X(x) = \tilde{\tau}$$
$$P'_X(\mathcal{X}) = 1$$

I adapt a lemma from Donsker and Varadhan [1975]:

Lemma 24. *Let F_X^* satisfy $\frac{dF_X^*}{F_X} = \frac{\exp(-\lambda(\tau(x)-\tilde{\tau}))}{\int_{\mathcal{X}}\exp(-\lambda(\tau(x)-\tilde{\tau}))dF_X}$. For any probability measure $\tilde{F}_X$ such that $\tilde{F}_X \ll F_X$ we have:*

$$\log\left(\int_{\mathcal{X}}\exp(-\lambda(\tau(x)-\tilde{\tau}))dF_X\right) = -\left[\int_{\mathcal{X}}\lambda(\tau(x)-\tilde{\tau})d\tilde{F}_X(x) + D_{KL}(\tilde{F}_X||F_X)\right] + D_{KL}(\tilde{F}_X||F_X^*)$$

Proof. First by definition of the KL-divergence we have:

$$\begin{aligned}
D_{KL}(\tilde{F}_X||F_X^*) &= \int_{\mathcal{X}}\log\left(\frac{d\tilde{F}_X}{dF_X^*}\right)d\tilde{F}_X \\
&= \int_{\mathcal{X}}\log\left(\frac{\frac{d\tilde{F}_X}{dF_X}}{\frac{dF_X^*}{dF_X}}\right)d\tilde{F}_X \\
&= \int_{\mathcal{X}}\left(\log\left(\frac{d\tilde{F}_X}{dF_X}\right) - \log\left(\frac{dF_X^*}{dF_X}\right)\right)d\tilde{F}_X \\
&= \int_{\mathcal{X}}\log\left(\frac{d\tilde{F}_X}{dF_X}\right)d\tilde{F}_X - \int_{\mathcal{X}}\log\left(\frac{\exp(-\lambda(\tau(x)-\tilde{\tau})}{\int_{\mathcal{X}}\exp(-\lambda(\tau(x)-\tilde{\tau})dF_X}\right)d\tilde{F}_X \\
&= D_{KL}(\tilde{F}_X||F_X) + \int_{\mathcal{X}}\lambda(\tau(x)-\tilde{\tau})d\tilde{F}_X \\
&+ \int_{\mathcal{X}}\log\left(\int_{\mathcal{X}}\exp(-\lambda(\tau(x)-\tilde{\tau}))dF_X\right)d\tilde{F}_X \\
&= D_{KL}(\tilde{F}_X||F_X) + \int_{\mathcal{X}}\lambda(\tau(x)-\tilde{\tau})d\tilde{F}_X + \log\left(\int_{\mathcal{X}}\exp(-\lambda(\tau(x)-\tilde{\tau}))dF_X\right)
\end{aligned}$$

since $\tilde{F}_X \ll F_X^* \ll F_X$ and simple algebra. Rearranging we get:

$$\log\left(\int_{\mathcal{X}}\exp(-\lambda(\tau(x)-\tilde{\tau})dF_X\right) = D_{KL}(\tilde{F}_X||F_X^*) - \left[\int_{\mathcal{X}}\lambda(\tau(x)-\tilde{\tau})d\tilde{F}_X + D_{KL}(\tilde{F}_X||F_X)\right]$$

□

Proof. i) From the lemma above we have:

$$\log\left(\int_{\mathcal{X}}\exp(-\lambda(\tau(x)-\tilde{\tau}))dF_X\right) = D_{KL}(\tilde{F}_X||F_X^*) - D_{KL}(\tilde{F}_X||F_X) - \int_{\mathcal{X}}\lambda(\tau(x)-\tilde{\tau})d\tilde{F}_X$$

Now observe that, since the term $\log\left(\int_{\mathcal{X}} \exp(-\lambda(\tau(x)-\tilde{\tau})dF_X\right)$ does not depend on $\tilde{F}_X$ we must have:

$$\begin{aligned}\arg\min_{\tilde{F}_X \ll F_X} D_{KL}(\tilde{F}_X||F_X^*) &= \arg\max_{\tilde{F}_X \ll F_X} -\int_X \lambda(\tau(x)-\tilde{\tau})d\tilde{F}_X - D_{KL}(\tilde{F}_X||F_X) \\ &= \arg\min_{\tilde{F}_X \ll F_X} \int_X \lambda(\tau(x)-\tilde{\tau})d\tilde{F}_X + D_{KL}(\tilde{F}_X||F_X)\end{aligned}$$

but clearly $F_X^* = \arg\min_{\tilde{F}_X \ll F_X} D_{KL}(\tilde{F}_X||F_X^*)$ so we must have

$$F_X^* = \arg\min_{\tilde{F}_X \ll F_X} D_{KL}(\tilde{F}_X||F_X) + \lambda \int_X (\tau(x)-\tilde{\tau})d\tilde{F}_X$$

which is the desired result. ii) Observe that $D_{KL}(F_X^*||F_X^*) = 0$ hence the value of the minimization problem:

$$\begin{aligned}&\min_{\tilde{F}_X \ll F_X} D_{KL}(\tilde{F}_X||F_X) + \lambda \int_X (\tau(x)-\tilde{\tau})d\tilde{F}_X \\ &= \min_{\tilde{F}_X \ll F_X} D_{KL}(\tilde{F}_X||F_X^*) - \log\left(\int_{\mathcal{X}} \exp(-\lambda(\tau(x)-\tilde{\tau}))dF_X\right) \\ &= -\log\left(\int_{\mathcal{X}} \exp(-\lambda(\tau(x)-\tilde{\tau}))dF_X\right)\end{aligned}$$

□

1.I.2 Proof of Fact 9

Proof. First, in this setting $F_X^* \ll F_X$ simply implies $p_1 = 0 \implies p_1^* = 0$. Excluding such a trivial case, 1.12 characterizes $\frac{p_1^*}{p_1}$. First we solve for the Lagrange multiplier λ in 1.12 by

noting that:

$$\begin{aligned}\tilde{\tau} &= \int_{\mathcal{X}} \tau(x) dF_X^* \\ &= \frac{\exp(-\lambda(\tau(1)-\tilde{\tau}))\tau(1)p_1 + \exp(-\lambda(\tau(0)-\tilde{\tau}))\tau(0)(1-p_1)}{\exp(-\lambda(\tau(1)-\tilde{\tau}))p_1 + \exp(-\lambda(\tau(0)-\tilde{\tau}))(1-p_1)}\end{aligned}$$

rearranging the denominator and since $\tilde{\tau}$ is a constant, we obtain

$$\begin{aligned}&\exp(-\lambda(\tau(1)-\tilde{\tau}))\tau(1)p_1 + \exp(-\lambda(\tau(0)-\tilde{\tau}))\tau(0)(1-p_1) \\ &= \exp(-\lambda(\tau(1)-\tilde{\tau}))\tilde{\tau}p_1 + \exp(-\lambda(\tau(0)-\tilde{\tau}))\tilde{\tau}(1-p_1)\end{aligned}$$

which gives the condition:

$$\exp(-\lambda(\tau(1)-\tilde{\tau}))(\tau(1)-\tilde{\tau})p_1 + \exp(-\lambda(\tau(0)-\tilde{\tau}))(\tau(0)-\tilde{\tau})(1-p_1) = 0$$

And isolating each side and taking logs we obtain:

$$-\lambda(\tau(1)-\tau(0)) = \log\left(\frac{(\tilde{\tau}-\tau(0))(1-p_1)}{(\tau(1)-\tilde{\tau})p_1}\right)$$

so that

$$-\lambda = \frac{1}{(\tau(1)-\tau(0))}\log\left(\frac{(\tilde{\tau}-\tau(0))(1-p_1)}{(\tau(1)-\tilde{\tau})p_1}\right)$$

Finally, replacing $-\lambda$ in 1.11 we have:

$$\frac{p_1^*}{p_1} = \frac{\exp\left(\log\left(\frac{(\tilde{\tau}-\tau(0))(1-p_1)}{(\tau(1)-\tilde{\tau})p_1}\right)\frac{\tau(1)-\tilde{\tau}}{\tau(1)-\tau(0)}\right)}{\exp\left(\log\left(\frac{(\tilde{\tau}-\tau(0))(1-p_1)}{(\tau(1)-\tilde{\tau})p_1}\right)\frac{\tau(1)-\tilde{\tau}}{\tau(1)-\tau(0)}\right)p_1 + \exp\left(\log\left(\frac{(\tilde{\tau}-\tau(0))(1-p_1)}{(\tau(1)-\tilde{\tau})p_1}\right)\frac{\tau(0)-\tilde{\tau}}{\tau(1)-\tau(0)}\right)(1-p_1)}$$

Finally rearranging and combining terms we have:

$$
\begin{aligned}
p_1^* &= \frac{\exp\left(\log\left(\frac{(\tilde{\tau}-\tau(0))(1-p_1)}{(\tau(1)-\tilde{\tau})p_1}\right)\frac{\tau(1)-\tilde{\tau}}{\tau(1)-\tau(0)}\right)p_1}{\exp\left(\log\left(\frac{(\tilde{\tau}-\tau(0))(1-p_1)}{(\tau(1)-\tilde{\tau})p_1}\right)\frac{\tau(1)-\tilde{\tau}}{\tau(1)-\tau(0)}\right)p_1+\exp\left(\log\left(\frac{(\tilde{\tau}-\tau(0))(1-p_1)}{(\tau(1)-\tilde{\tau})p_1}\right)\frac{\tau(0)-\tilde{\tau}}{\tau(1)-\tau(0)}\right)(1-p_1)} \\
&= \frac{\left(\frac{(\tilde{\tau}-\tau(0))(1-p_1)}{(\tau(1)-\tilde{\tau})p_1}\right)^{\frac{\tau(1)-\tilde{\tau}}{\tau(1)-\tau(0)}}p_1}{\left(\frac{(\tilde{\tau}-\tau(0))(1-p_1)}{(\tau(1)-\tilde{\tau})p_1}\right)^{\frac{\tau(1)-\tilde{\tau}}{\tau(1)-\tau(0)}}p_1+\left(\frac{(\tilde{\tau}-\tau(0))(1-p_1)}{(\tau(1)-\tilde{\tau})p_1}\right)^{\frac{\tau(0)-\tilde{\tau}}{\tau(1)-\tau(0)}}(1-p_1)} \\
&= \frac{1}{1+\left(\frac{(\tilde{\tau}-\tau(0))(1-p_1)}{(\tau(1)-\tilde{\tau})p_1}\right)^{\frac{\tau(0)-\tilde{\tau}}{\tau(1)-\tau(0)}-\frac{\tau(1)-\tilde{\tau}}{\tau(1)-\tau(0)}}\frac{(1-p_1)}{p_1}} \\
&= \frac{1}{1+\left(\frac{(\tilde{\tau}-\tau(0))(1-p_1)}{(\tau(1)-\tilde{\tau})p_1}\right)^{-1}\frac{(1-p_1)}{p_1}} \\
&= \frac{1}{1+\frac{\tilde{\tau}-\tau(0)}{\tau(1)-\tau(0)}} \\
&= \frac{1}{\frac{\tau(1)-\tau(0)}{\tilde{\tau}-\tau(0)}} \\
&= \frac{\tilde{\tau}-\tau(0)}{\tau(1)-\tau(0)}
\end{aligned}
$$

which, with $\tilde{\tau}=0$, is the solution obtained in Equation (1.11). □

1.I.3 Proof of Proposition 10

Proposition 10 (Local to boundary $\tilde{\tau}$)**.** *Let Assumptions 1-3 hold and let $\tilde{\tau}_m \to \tilde{\tau}_b \in \partial L(\mathcal{P}_X)$. Assume that the pre-image $\tau^{-1}(\tilde{\tau}_b) = \mathcal{X}_b = \{x_b\} \in \mathcal{X}^\circ$ is a singleton. Further, let X be compactly supported, with density $f(x) < M$ on $\mathcal{X}$. Then the sequence of least favorable distributions for the policy-maker's problem with parameter $\tilde{\tau}_m$, denoted $F^*_{X,m}$, converges weakly to δ_{x_b}, the Dirac delta distribution with point mass at x_b, that is:*

$$\lim_{m\to\infty}\int_{\mathcal{X}} g(x)dF^*_{X,m}(x) \to \int_{\mathcal{X}} g(x)\delta_{x_b} := g(x_b)$$

for $g \in C_b(\mathcal{X})$, the space of all continuous, bounded functions on $\mathcal{X}$.

Proof. First observe that by Lemma 7 and the fact that each $\tau_m \in L^{\circ}(\mathcal{P}_X)$ we can construct the sequence of *least favorable distributions* $F^*_{m,X}$ satisfying:

$$\frac{dF^*_{m,X}}{dF_X}(x) = \frac{\exp(-\lambda_m(\tau(x) - \tilde{\tau}_m))}{\int_{\mathcal{X}} \exp(-\lambda_m(\tau(x) - \tilde{\tau}_m))dF_X}$$
$$\lambda_m : \quad \int_{\mathcal{X}} \exp(-\lambda_m(\tau(x) - \tilde{\tau}_m))(\tau(x) - \tilde{\tau}_m)dF_X = 0$$

□

Without loss of generality consider the case where $\tilde{\tau}_b = \max_{\mathcal{X}} \tau(x)$. First notice that the sequence of λ_m defined above is decreasing and unbounded below. To see that it's decreasing observe that implicitly differentiating $\lambda(\tilde{\tau})$:

$$\begin{aligned}
&\frac{\partial}{\partial \tilde{\tau}} \int_{\mathcal{X}} \exp(-\lambda_m(\tau(x) - \tilde{\tau}_m))(\tau(x) - \tilde{\tau}_m)dF_X(x) \\
&= -\frac{\partial \lambda}{\partial \tilde{\tau}}(\tilde{\tau}) \int_{\mathcal{X}} \exp(-\lambda(\tilde{\tau})(\tau(x) - \tilde{\tau}))(\tau(x) - \tilde{\tau})^2 dF_X \\
&+ \lambda(\tilde{\tau}) \int_{\mathcal{X}} \exp(-\lambda(\tilde{\tau})(\tau(x) - \tilde{\tau}))(\tau(x) - \tilde{\tau})dF_X \\
&- \int_{\mathcal{X}} \exp(-\lambda(\tilde{\tau})(\tau(x) - \tilde{\tau}))dF_X = 0
\end{aligned}$$

by the Dominated Convergence Theorem with envelope $g = \exp(2M) \cdot 2M$. Note that by the definition of $\lambda(\tilde{\tau})$ the second term is equal to 0. Isolating the derivative of λ with respect to $\tilde{\tau}$ we have:

$$\frac{\partial \lambda}{\partial \tilde{\tau}}(\tilde{\tau}) = -\frac{\int_{\mathcal{X}} \exp(-\lambda(\tilde{\tau})(\tau(x) - \tilde{\tau}))dF_X}{\int_{\mathcal{X}} \exp(-\lambda(\tilde{\tau})(\tau(x) - \tilde{\tau}))(\tau(x) - \tilde{\tau})^2 dF_X} < 0$$

so $\lambda(\tilde{\tau})$ is strictly decreasing on its domain. Suppose $\lambda_m \geq -B$ for all $m \in N$, with $B > 0$. Then:

$$\int_{\mathcal{X}} \exp(-\lambda_m(\tau(x) - \tilde{\tau}_m))(\tau(x) - \tilde{\tau}_m)dF_X \leq \quad \int_{\mathcal{X}} \exp(B(\tau(x) - \tilde{\tau}_m))(\tau(x) - \tilde{\tau}_m)dF_X$$

so taking the limit fro $m \to \infty$, if $P_X(\tau(x) \neq \tilde{\tau}_b) > 0$:

$$\begin{aligned}
&\lim_{m\to\infty} \int_{\mathcal{X}} \exp(-\lambda_m(\tau(x) - \tilde{\tau}_m))(\tau(x) - \tilde{\tau}_m) dF_X \\
&\leq \lim_{m\to\infty} \int_{\mathcal{X}} \exp(B(\tau(x) - \tilde{\tau}_m))(\tau(x) - \tilde{\tau}_m) dF_X(x) \\
&\leq \int_{\mathcal{X}} \exp(B(\tau(x) - \tilde{\tau}_b))(\tau(x) - \tilde{\tau}_b) dF_X(x) < 0
\end{aligned}$$

Then, there exist $m^* \in \mathbb{N}$ such that $\int_{\mathcal{X}} \exp(\lambda_{m^*}(\tau(x) - \tilde{\tau}_{m^*}))(\tau(x) - \tilde{\tau}_{m^*}) dF_X < 0$ which is a contradiction. So λ_m must be unbounded below. Because it's a strictly decreasing, unbounded below sequence, it must be the case that $\lambda_m \to -\infty$ as $\tilde{\tau}_m \to \tilde{\tau}_b$. Now we show convergence in distribution to δ_{x_b}. Let $\varphi(\cdot) \in \mathcal{C}_b$. We want to show:

$$\lim_{m\to\infty} \int_{\mathcal{X}} \varphi(x) dF^*_{X,m}(x) \to \int_{\mathcal{X}} \varphi(x) \delta_{x_b}(x) = \varphi(x_b)$$

We have:

$$\begin{aligned}
\int_{\mathcal{X}} \varphi(x) dF^*_{X,m}(x) &= \int_{\mathcal{X}} \varphi(x) \frac{\exp(-\lambda_m(\tau(x) - \tilde{\tau}_b)) dF_X(x)}{\int_{\mathcal{X}} \exp(-\lambda_m(\tau(x) - \tilde{\tau}_b)) dF_X(x)} \\
&= \int_{\mathcal{X}} \varphi(x) \frac{\exp(-\lambda_m(\tau(x) - \tilde{\tau}_b)) dF_X(x)}{\int_{\mathcal{X}} \exp(-\lambda_m(\tau(x) - \tilde{\tau}_b)) dF_X(x)}
\end{aligned}$$

Noticing that $\lambda_m < 0$. Consider the change of variables $y = \sqrt{-\lambda_m}(x_b - x)$. Then $x =$

$x_b - \frac{y}{\sqrt{-\lambda_m}}$, $dx = -\frac{dy}{\sqrt{-\lambda_m}}$. By the change of variable formula:

$$
\begin{aligned}
&\int_{\mathcal{X}} \varphi(x) \frac{\exp(-\lambda_m(\tau(x) - \tilde{\tau}_b)) f(x) dx}{\int_{\mathcal{X}} \exp(-\lambda_m(\tau(x) - \tilde{\tau}_b)) f(x) dx} \\
&= \frac{\int_{\mathbb{R}^k} \varphi\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) \exp\left(-\lambda_m \left(\tau\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) - \tau(x_b)\right)\right) f\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) \mathbb{1}_{\mathcal{Y}(\lambda_m)}(y) \frac{1}{\sqrt{-\lambda_m}} dy}{\int_{\mathbb{R}^k} \exp\left(-\lambda_m \left(\tau\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) - \tau(x_b)\right)\right) f\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) \mathbb{1}_{\mathcal{Y}(\lambda_m)}(y) \frac{1}{\sqrt{-\lambda_m}} dy} \\
&= \frac{\int_{\mathbb{R}^k} \varphi\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) \exp\left(-\lambda_m \left(\tau\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) - \tau(x_b)\right)\right) f\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) \mathbb{1}_{\mathcal{Y}(\lambda_m)}(y) dy}{\int_{\mathbb{R}^k} \exp\left(-\lambda_m \left(\tau\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) - \tau(x_b)\right)\right) f\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) \mathbb{1}_{\mathcal{Y}(\lambda_m)}(y) dy}
\end{aligned}
$$

Note that, if X is compactly supported then $f(x) = 0$ outside of a compact set $K \subseteq \mathbb{R}^k$ hence. Moreover, if $f(x) < M$ we have the dominating function given by:

$$
\begin{aligned}
&\varphi\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) \exp\left(-\lambda_m \left(\tau\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) - \tau(x_b)\right)\right) f\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) \mathbb{1}_{\mathcal{Y}(\lambda_m)}(y) dy \\
&\leq \|\varphi\|_\infty M \mathbb{1}_K(y)
\end{aligned}
$$

on $\mathbb{R}^k$ and $\int_{\mathbb{R}^k} \|\varphi\|_\infty M \mathbb{1}_K(x) dx = \|\varphi\|_\infty \cdot M \cdot \mathrm{vol}(K) < +\infty$. hence the assumptions of the Dominated Convergence theorem hold. Then we have:

$$
\begin{aligned}
&= \lim_{m \to \infty} \int_{\mathbb{R}^k} \varphi\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) \exp\left(-\lambda_m \left(\tau\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) - \tau(x_b)\right)\right) \\
&\times f\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) \mathbb{1}_{\mathcal{Y}(\lambda_m)}(y) dy \\
&= \int_{\mathbb{R}^k} \lim_{m \to \infty} \varphi\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) \exp\left(-\lambda_m \left(\tau\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) - \tau(x_b)\right)\right) \\
&\times f\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) \mathbb{1}_{\mathcal{Y}(\lambda_m)}(y) dy
\end{aligned}
$$

Now consider Taylor expanding $\tau(\cdot)$ around x_b. Because x_b is a maximizer, the Jacobian

$J_\tau(x_b) : \mathbb{R}^k \to \mathbb{R}$ is the zero matrix, from first order conditions. Hence:

$$\begin{aligned}
&\exp\left(-\lambda_m\left(\tau\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) - \tau(x_b)\right)\right) \\
&= \exp\left(-\lambda_m\left(\tau(x_b) - J_\tau(x_b)\left(\frac{y}{\sqrt{-\lambda_m}}\right) + \frac{1}{2}\cdot\frac{1}{-\lambda_m} y^T H_\tau(x_b) y - \tau(x_b)\right)\right) \\
&= \exp\left(\frac{1}{2} y^T H_\tau(x_b) y + o(1)\right)
\end{aligned}$$

where $H_\tau(x_b)$ is the $k \times k$ Hessian matrix of τ, evaluated at the maximizer x_b. Moreover:

$$\begin{aligned}
&\int_{\mathbb{R}^k} \lim_{m\to\infty} \varphi\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) \exp\left(-\lambda_m\left(\tau\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) - \tau(x_b)\right)\right) f\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) \mathbb{1}_{\mathcal{Y}(\lambda_m)}(y) dy \\
&= \int_{\mathbb{R}^k} \varphi(x_b) \exp\left(\frac{1}{2} y^T H(x_b) y\right) f(x_b) dy \\
&= \varphi(x_b) \int_{\mathbb{R}^k} \exp\left(\frac{1}{2} y^T H(x_b) y\right) f(x_b) dy
\end{aligned}$$

Now the denominator can be treated identically to have:

$$\begin{aligned}
&\int_{\mathbb{R}^k} \lim_{m\to\infty} \exp\left(-\lambda_m\left(\tau\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) - \tau(x_b)\right)\right) f\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) \mathbb{1}_{\mathcal{Y}(\lambda_m)}(y) dy \\
&= \int_{\mathbb{R}^k} \exp\left(\frac{1}{2} y^T H(x_b) y\right) f(x_b) dy
\end{aligned}$$

Now because x_b is a maximizer, $H(x_b)$ is negative definite so the quantities above are finite

and the numerator is greater than 0. Finally:

$$
\begin{aligned}
&\lim_{m\to\infty}\int_{\mathcal{X}}\varphi(x)dF^*_{X,m}(x)\\
&=\lim_{m\to\infty}\frac{\int_{\mathcal{X}}\varphi(x)\exp(-\lambda_m(\tau(x)-\tilde{\tau}_b))f(x)dx}{\int_{\mathcal{X}}\exp(-\lambda_m(\tau(x)-\tilde{\tau}_b))f(x)dx}\\
&=\frac{\lim_{m\to\infty}\int_{\mathcal{X}}\varphi(x)\exp(-\lambda_m(\tau(x)-\tilde{\tau}_b))f(x)dx}{\lim_{m\to\infty}\int_{\mathcal{X}}\exp(-\lambda_m(\tau(x)-\tilde{\tau}_b))f(x)dx}\\
&=\frac{\varphi(x_b)\int_{\mathbb{R}^k}\exp\left(\frac{1}{2}y^TH(x_b)y\right)f(x_b)dy}{\int_{\mathbb{R}^k}\exp\left(\frac{1}{2}y^TH(x_b)y\right)f(x_b)dy}\\
&=\varphi(x_b)
\end{aligned}
$$

Since $\varphi(\cdot)\in\mathcal{C}_b$ was arbitrary, by the Portmanteau theorem, $dF^*_{X,m}\xrightarrow{d}\delta_{x_b}$.

In the general case where $\mathcal{X}_b$ is not a singleton, it seems that the *least favorable distribution* still concentrates around the uniform distribution on the $\mathcal{X}_b$, rather than *any* distribution like the figure below suggests. I leave this interesting case for future work.

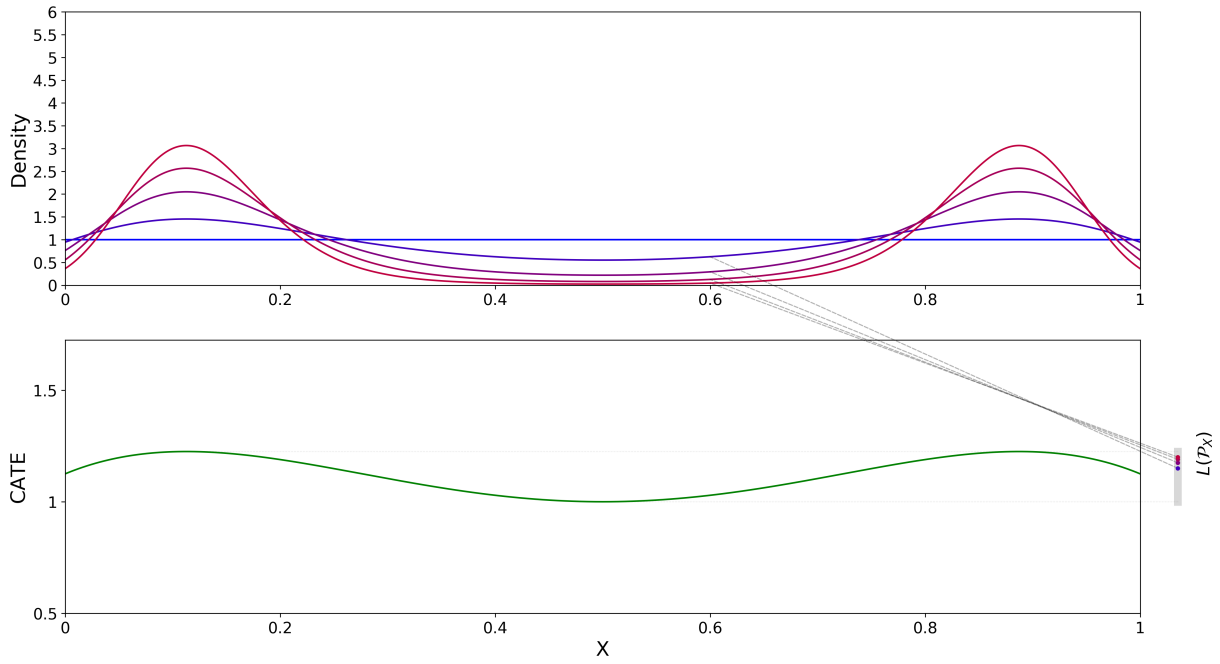


Figure 1.I.1: Here $\tau(x)$ is quadratic, experimental distribution is uniform and there are two peaks. It appears that the *least favorable distribution* concentrates around both peaks.

1.I.4 Proof of Proposition 19

Proposition 19 (Quadratic-Normal least favorable closed-ness)**.** *The parametric class* $\mathcal{N}(\mu,\sigma^2)$ *is* ***least favorable closed*** *for quadratic Conditional Average Treatment Effects. That is, if* $X \in \mathbb{R}^k$ *follows the multivariate normal distribution* $X \sim \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ *where* $\boldsymbol{\Sigma}$ *is p.d. and* $\tau(x) = \boldsymbol{x^T A x} + \boldsymbol{x^T \beta} + c$ *for* $\boldsymbol{\beta} \in \mathbb{R}^k$ *then* F_X^* *is the measure induced by* $X^* \sim \mathcal{N}(\boldsymbol{\mu^*},\boldsymbol{\Sigma^*})$ *with* $\boldsymbol{\mu}^* = (\boldsymbol{\Sigma}^{-1} + 2\lambda\boldsymbol{A})^{-1}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \lambda\boldsymbol{\beta})$ *and* $\boldsymbol{\Sigma^*} = (\boldsymbol{\Sigma}^{-1} + 2\lambda\boldsymbol{A})^{-1}$*, provided that* $(\boldsymbol{\Sigma}^{-1} + 2\lambda\boldsymbol{A})^{-1}$ *is p.d. The parameter* λ *is defined as in Equation* (1.9)*.*

Proof. Suppose $\mathcal{X} = \mathbb{R}^k$, $X \sim \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\sigma})$ and $\tau(x) = \boldsymbol{x^T A x} + \boldsymbol{x^T \beta} + c$. By Lemma 7 the Radon-Nikodym derivative of the least favorable distribution is given by Equation (1.8) so the distribution of F_X^* must have density:

$$
\begin{aligned}
d\mu_X^* &:= \frac{\exp(-\lambda(\tau(x)-\tilde{\tau}))\frac{\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^k\det(\boldsymbol{\Sigma})}}dx}{\displaystyle\int_{\mathcal{X}}\exp(-\lambda(\tau(x)-\tilde{\tau}))\frac{\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^k\det(\boldsymbol{\Sigma})}}dx} \\
&= \frac{\exp(-\lambda(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}+\boldsymbol{x}^T\boldsymbol{\beta}+c-\tilde{\tau}))\frac{\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^k\det(\boldsymbol{\Sigma})}}dx}{\displaystyle\int_{\mathcal{X}}\exp(-\lambda(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}+\boldsymbol{x}^T\boldsymbol{\beta}+c-\tilde{\tau}))\frac{\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^k\det(\boldsymbol{\Sigma})}}dx} \\
&= \frac{\frac{\exp(-\lambda(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}+\boldsymbol{x}^T\boldsymbol{\beta}+c-\tilde{\tau})-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))}{\sqrt{(2\pi)^k\det(\boldsymbol{\Sigma})}}dx}{\displaystyle\int_{\mathcal{X}}\frac{\exp(-\lambda(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}+\boldsymbol{x}^T\boldsymbol{\beta}+c-\tilde{\tau})-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))}{\sqrt{(2\pi)^k\det(\boldsymbol{\Sigma})}}dx} \\
&= \frac{\exp(-\lambda(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}+\boldsymbol{x}^T\boldsymbol{\beta}+c-\tilde{\tau})-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))dx}{\displaystyle\int_{\mathcal{X}}\exp(-\lambda(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}+\boldsymbol{x}^T\boldsymbol{\beta}+c-\tilde{\tau})-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))dx} \\
&= \frac{\exp(-\frac{1}{2}(\boldsymbol{x}-(\boldsymbol{\Sigma}^{-1}+2\lambda\boldsymbol{A})^{-1}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\lambda\boldsymbol{\beta}))(\boldsymbol{\Sigma}^{-1}+2\lambda\boldsymbol{A}))(x-(\boldsymbol{\Sigma}^{-1}+2\lambda\boldsymbol{A})^{-1}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\lambda\boldsymbol{\beta}))dx}{\displaystyle\int_{\mathcal{X}}\exp(-\frac{1}{2}(x-(\boldsymbol{\Sigma}^{-1}+2\lambda\boldsymbol{A})^{-1}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\lambda\boldsymbol{\beta}))(\boldsymbol{\Sigma}^{-1}+2\lambda\boldsymbol{A}))(x-(\boldsymbol{\Sigma}^{-1}+2\lambda\boldsymbol{A})^{-1}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\lambda\boldsymbol{\beta}))dx} \\
&\times \frac{\exp(\lambda c+\lambda\tilde{\tau}-\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\frac{1}{2}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\lambda\boldsymbol{\beta})(\boldsymbol{\Sigma}^{-1}+2\lambda\boldsymbol{\beta})^{-1}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\lambda\boldsymbol{\beta}))}{\exp(\lambda c+\lambda\tilde{\tau}-\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\frac{1}{2}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\lambda\boldsymbol{\beta})(\boldsymbol{\Sigma}^{-1}+2\lambda\boldsymbol{\beta})^{-1}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\lambda\boldsymbol{\beta}))} \\
&= \frac{\exp(-\frac{1}{2}(\boldsymbol{x}-(\boldsymbol{\Sigma}^{-1}+2\lambda\boldsymbol{A})^{-1}(\boldsymbol{\Sigma}^{-1}\mu-\lambda\boldsymbol{\beta}))(\boldsymbol{\Sigma}^{-1}+2\lambda\boldsymbol{A}))(x-(\boldsymbol{\Sigma}^{-1}+2\lambda\boldsymbol{A})^{-1}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\lambda\boldsymbol{\beta}))dx}{\displaystyle\int_{\mathcal{X}}\exp(-\frac{1}{2}(\boldsymbol{x}-(\boldsymbol{\Sigma}^{-1}+2\lambda\boldsymbol{A})^{-1}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\lambda\boldsymbol{\beta}))(\boldsymbol{\Sigma}^{-1}+2\lambda\boldsymbol{A}))(\boldsymbol{x}-(\boldsymbol{\Sigma}^{-1}+2\lambda\boldsymbol{A})^{-1}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\lambda\boldsymbol{\beta}))dx}
\end{aligned}
$$

from which we can recognize the form of the normal distribution with mean $\boldsymbol{\mu}^*$ and variance covariance matrix $\boldsymbol{\Sigma}^*$. The steps above follow from completing the square and from the properties of $\exp(\cdot)$. □

1.I.5 Proof of Proposition 13

Proposition 13. *The de-biased GMM nonparametric influence function based on moment function $g(\cdot)$ is:*

$$\phi(w,\theta,\gamma_0,\alpha_0) = \begin{bmatrix} \exp\Big(-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})\Big)\cdot(-\lambda) \\ \exp\Big(-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})\Big)\cdot(1-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})) \end{bmatrix} \times \left(\frac{d(y-\gamma_{1,F_0}(x))}{\pi_{F_0}(x)} - \frac{(1-d)(y-\gamma_{0,F_0}(x))}{1-\pi_{F_0}(x)}\right)$$

which could be written in the form:

$$\phi(w,\theta,\gamma_0,\alpha_0) = \begin{bmatrix} \exp\Big(-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})\Big)\cdot(-\lambda) \\ \exp\Big(-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})\Big)\cdot(1-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})) \end{bmatrix} \times \left(\begin{bmatrix} \alpha_{1,F_0}(x) \\ \alpha_{0,F_0}(x) \end{bmatrix}^T \begin{bmatrix} d(y-\gamma_{1,F_0}(x)) \\ (1-d)(y-\gamma_{0,F_0}(x)) \end{bmatrix}\right)$$

with $\alpha_{F_0}(x) := \begin{bmatrix} \alpha_{1,F_0}(x) \\ \alpha_{0,F_0}(x) \end{bmatrix} = \begin{bmatrix} \frac{1}{\pi_{F_0}(x)} \\ \frac{1}{1-\pi_{F_0}(x)} \end{bmatrix}$.

Proof. Let $F_r = (1-r)F_0 + rH$ for an arbitrary distribution H that satisfies unconfoundedness. Then F_r is a distribution because it's a convex combination of two distributions, and it satisfies unconfounded-ness. Therefore we can refer to the identification results:

$$\mathbb{E}_{F_r}[Y_1|X] = \mathbb{E}_{F_r}[Y|D=1,X]$$
$$\mathbb{E}_{F_r}[Y_0|X] = \mathbb{E}_{F_r}[Y|D=0,X]$$

and derive the distributional derivative of $\mathbb{E}_{F_r}[Y|D=1,X] - \mathbb{E}_{F_r}[Y|D=0,X]$ with respect to r and evaluate it at $r=0$. Alternatively one may start with the propensity score

weighting identification result below:

$$\mathbb{E}_{F_r}\left[\frac{Y\cdot D}{\pi_{F_r}(X)}-\frac{Y\cdot(1-D)}{\pi_{F_r}(X)}\middle| X\right]=\mathbb{E}_{F_r}[Y_1-Y_0|X]$$

and proceed as above to derive the distributional derivative of $\mathbb{E}_F[g(W,\theta,\gamma(F_r))]$. The second approach is more cumbersome so we present the proof for the regression adjustment method but note that both would be valid approaches to find the nonparametric influence function. Computing the derivative of the moment condition with respect to r and evaluating it at $r=0$ we have:

$$\begin{aligned}
\frac{d\mathbb{E}[g(W,\theta,\gamma(F_r))]}{dr}\bigg|_{r=0} &= \frac{d}{dr}\mathbb{E}\begin{bmatrix}\exp(-\lambda_0(\gamma_{1,F_r}(x)-\gamma_{0,F_r}(x)-\tilde{\tau}))-\nu]\\ \exp\left(-\lambda_0(\gamma_{1,F_r}(x)-\gamma_{0,F_r}(x)-\tilde{\tau})\right)(\gamma_{1,F_r}(x)-\gamma_{0,F_r}(x)-\tilde{\tau})]\end{bmatrix}\bigg|_{r=0}\\
&= \int_{\mathcal{X}}\frac{d}{dr}\begin{bmatrix}\exp\left(-\lambda_0\cdot(\gamma_{1,F_r}(x)-\gamma_{0,F_r}(x)-\tilde{\tau})\right)\\ \exp\left(-\lambda_0\cdot(\gamma_{1,F_r}(x)-\gamma_{0,F_r}(x)-\tilde{\tau})\right)(\gamma_{1,F_r}(x)-\gamma_{0,F_r}(x)-\tilde{\tau})\end{bmatrix}f_0(x)dx\bigg|_{r=0}\\
&= \int_{\mathcal{X}}\begin{bmatrix}\exp\left(-\lambda_0\cdot(\gamma_{1,F_r}(x)-\gamma_{0,F_r}(x)-\tilde{\tau})\right)\cdot(-\lambda_0)\\ \exp\left(-\lambda_0\cdot(\gamma_{1,F_r}(x)-\gamma_{0,F_r}(x)-\tilde{\tau})\right)\cdot(1-\lambda\cdot(\gamma_{1,F_r}(x)-\gamma_{0,F_r}(x)-\tilde{\tau}))\end{bmatrix}\\
&\quad\times\frac{\partial}{\partial r}(\gamma_{1,F_r}(x)-\gamma_{0,F_r}(x))f_0(x)dx
\end{aligned}$$

In order to characterize the contribution of the functional we have:

$$
\begin{aligned}
&\frac{\partial}{\partial r}(\gamma_{1,F_r}(x)-\gamma_{0,F_r}(x)) \\
&=\frac{\partial}{\partial r}\int_{\mathcal{Y}}\frac{y}{\int_{\mathcal{Y}}(1-r)dF_0(y,1,x)+rdH(y,1,x)}((1-r)dF(y,1,x)+rdH(y,1,x)) \\
&-\frac{\partial}{\partial r}\int_{\mathcal{Y}}\frac{y}{\int_{\mathcal{Y}}(1-r)dF_0(y,0,x)+rdH(y,0,x)}((1-r)dF(y,0,x)+rdH(y,0,x)) \\
&=\frac{\int_{\mathcal{Y}}y\cdot[dH(y,1,x)-dF_0(y,1,x)]\int_{\mathcal{Y}}(1-r)dF_0(y,1,x)+rdH(y,1,x)}{\left(\int_{\mathcal{Y}}(1+r)dF_0(y,1,x)+rdH(y,1,x)\right)^2} \\
&-\frac{\int_{\mathcal{Y}}y[dH(y,1,x)-dF_0(y,1,x)]((1-r)dF_0(y,1,x)-dH(y,1,x))}{\left(\int_{\mathcal{Y}}(1+r)dF_0(y,1,x)+rdH(y,1,x)\right)^2} \\
&-\frac{\int_{\mathcal{Y}}y\cdot[dH(y,0,x)-dF_0(y,0,x)]\int_{\mathcal{Y}}(1-r)dF_0(y,0,x)+rdH(y,0,x)}{\left(\int_{\mathcal{Y}}(1+r)dF_0(y,0,x)+rdH(y,0,x)\right)^2} \\
&+\frac{\int_{\mathcal{Y}}y[dH(y,0,x)-dF_0(y,0,x)]((1-r)dF_0(y,0,x)-dH(y,0,x))}{\left(\int_{\mathcal{Y}}(1+r)dF_0(y,0,x)+rdH(y,0,x)\right)^2}
\end{aligned}
$$

Below $f_0(d,x)=\int_{\mathcal{Y}}dF_0(y,d,x)$ and the same holds for $h(\cdot)$. Evaluating this expression at $r=0$ one obtains:

$$
\int y\cdot\frac{dH(y,1,x)}{f_0(1,x)}-\int y\cdot\frac{h(1,x)\cdot dF_0(y,1,x)}{f_0(1,x)^2}-\int y\cdot\frac{dH(y,0,x)}{f_0(0,x)}+\int y\cdot\frac{h(0,x)\cdot dF_0(y,0,x)}{f_0(0,x)^2}
$$

Combining this with the derivative of the moment condition with respect to the γ we have:

$$
\begin{aligned}
\frac{d\mathbb{E}[g(W,\theta,\gamma(F_r))]}{dr}=\int_{\mathcal{Y}\times\{0,1\}\times\mathcal{X}}&\begin{bmatrix}\exp\left(-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})\right)\cdot(-\lambda)\\ \exp\left(-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})\right)\cdot(1-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau}))\end{bmatrix} \\
&\times\left(\frac{d(y-\gamma_{1,F_0}(x))}{\pi_{F_0}(x)}-\frac{(1-d)(y-\gamma_{0,F_0}(x))}{1-\pi_{F_0}(x)}\right)dH(y,d,x)
\end{aligned}
$$

or $\frac{d\mathbb{E}[g(W,\theta,\gamma(F_r))]}{dr} = \int_{\mathcal{Y}\times\{0,1\}\times\mathcal{X}} \phi(w,\theta,\gamma(F_0),\alpha(F_0))dH(w)$ for

$$\begin{aligned}
\phi(w,\theta,\gamma,\alpha) &= \begin{bmatrix} \exp\left(-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})\right)\cdot(-\lambda) \\ \exp\left(-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})\right)\cdot(1-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})) \end{bmatrix} \\
&\times \left(\frac{d(y-\gamma_{1,F_0}(x))}{\pi_F(x)} - \frac{(1-d)(y-\gamma_{0,F_0}(x))}{1-\pi_F(x)}\right) \\
&= \begin{bmatrix} \exp\left(-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})\right)\cdot(-\lambda) \\ \exp\left(-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})\right)\cdot(1-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})) \end{bmatrix} \\
&\times \left(\begin{bmatrix} \alpha_{1,F_0}(x) \\ \alpha_{0,F_0}(x) \end{bmatrix}^T \begin{bmatrix} d(y-\gamma_{1,F_0}(x)) \\ (1-d)(y-\gamma_{0,F_0}(x)) \end{bmatrix}\right)
\end{aligned}$$

and $\alpha_{F_0}(X) := \begin{bmatrix} \alpha_{1,F_0}(x) \\ \alpha_{0,F_0}(x) \end{bmatrix} = \begin{bmatrix} \frac{1}{\pi_{F_0}(X)} \\ \frac{1}{1-\pi_{F_0}(X)} \end{bmatrix}$. Note that above $\phi(\cdot)$ is the Riesz representer of the linear functional $\left.\frac{d\mathbb{E}[g(W,\theta,\gamma(F_r))]}{dr}\right|_{r=0} : \mathcal{H} \to \mathbb{R}^2$ which maps H to $\mathbb{R}^2$.

Observe that $\mathbb{E}_{F_0}[\phi(W,\theta,\gamma_0(X),\alpha_0(X)] = 0$ by the law of iterated expectations. Moreover, for any distribution F, $\mathbb{E}_F\left[\left.\frac{D(Y-\mathbb{E}_F[Y|D=1,X])}{\pi_F(X)} - \frac{(1-D)(Y-\mathbb{E}_F[Y|D=0,X]}{1-\pi_F(X)}\right| X\right] = 0$. □

1.I.6 Proof of Proposition 14

Proposition 14. *Equation* (1.16) *satisfies Neyman orthogonality.*

Proof. To show that they are Neyman orthogonal we verify the conditions for Theorem 1 in Chernozhukov et al. [2020] in the Appendix. Let $\gamma_{1,F}(X), \gamma_{0,0}(X)$ denote $\mathbb{E}_F[Y|D=1,X], \mathbb{E}_F[Y|D=0,X]$ respectively.

i) Equation (1.15) holds. This has been verified above.

ii) $\int_{\mathcal{Y}_0\times\mathcal{Y}_1\times\mathcal{X}} \phi(w,\gamma(F_r),\theta,\alpha(F_r))F_r(dw) = 0$ for all $r \in [0,\tilde{r})$:

This is immediate by the law of iterated expectations

$$
\begin{aligned}
&\mathbb{E}_{F_r}[\phi(W,\gamma(F_r),\theta,\alpha(F_r)] \\
&= \mathbb{E}_{F_r}\left[\mathbb{E}_{F_r}[\phi(W,\gamma(F_r),\theta,\alpha(F_r)|X]\right] \\
&= \mathbb{E}_{F_r}\left[v(X)\cdot\mathbb{E}_{F_r}\left[\left(\frac{d(y-\gamma_{1,F_r}(X))}{\pi_{F_r}(X)} - \frac{(1-d)(y-\gamma_{1,F_r}(X))}{1-\pi_{F_r}(X)}\right)\middle| X\right]\right] \\
&= \mathbb{E}_{F_r}[v(X)\cdot 0] \\
&= 0
\end{aligned}
$$

for $v(X) = \begin{bmatrix} \exp\left(-\lambda\cdot(\gamma_{1,F_r}(x)-\gamma_{0,F_r}(x)-\tilde{\tau})\right)\cdot(-\lambda) \\ \exp\left(-\lambda\cdot(\gamma_{1,F_r}(x)-\gamma_{0,F_r}(x)-\tilde{\tau})\right)\cdot(1-\lambda\cdot(\gamma_{1,F_r}(x)-\gamma_{0,F_r}(x)-\tilde{\tau})) \end{bmatrix}$

iii) $\int_{\mathcal{Y}_0\times\mathcal{Y}_1\times\mathcal{X}}\phi(w,\gamma(F_r),\theta,\alpha(F_r))H(dw)$ and $\int_{\mathcal{Y}_0\times\mathcal{Y}_1\times\mathcal{X}}\phi(w,\gamma(F_r),\theta,\alpha(F_r))F_0(dw)$ are continuous at $r=0$.

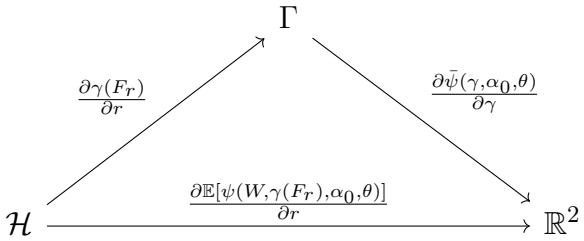
For a given H, we show that function $b: r\mapsto\int_{\mathcal{Y}_0\times\mathcal{Y}_1\times\mathcal{X}}\phi(w,\gamma(F_r),\theta,\alpha(F_r))H(dw)$ is continuous at $r=0$. Take a sequence $r_m\to r=0$, then $\phi_n(w):=\phi(w,\gamma(F_{r_m}),\theta,\alpha(F_{r_m}))$ converges H-almost everywhere to $\phi_0(w):=\phi(w,\gamma(F_0),\theta,\alpha(F_0))$. Moreover we have $\phi_m(w)\leq F(w)$ for all $m\in\mathbb{N}$ with $F\in L^1(H)$. By the dominated convergence theorem we have: $b(r_m)\to b(0)$ which is the desired result.

An analogous argument applies to the integral with respect to F_0. As a consequence of Theorems 1,2 and 3 in Chernozhukov et al. [2020] $\psi(w,\gamma,\theta,\alpha)$ is Neyman orthogonal. We can also verify Neyman orthogonality directly from the form of the $\bar{\psi}$ function. In particular:

$$
\begin{aligned}
&\frac{\partial}{\partial r}\mathbb{E}[\psi(W,\theta,\gamma_{F_r},\alpha_{F_r})]\Big|_{r=0} \\
&= \frac{\partial}{\partial r}\mathbb{E}[g(W,\theta,\gamma)+\phi(W,\theta,\gamma,\alpha)]\Big|_{r=0} \\
&= \mathbb{E}\Bigg[\frac{\partial}{\partial r}\begin{bmatrix}\exp\left(-\lambda_0\cdot(\gamma_{1,F_r}(X)-\gamma_{0,F_r}(X)-\tilde{\tau})\right) \\ \exp\left(-\lambda_0\cdot(\gamma_{1,F_r}(X)-\gamma_{0,F_r}(X)-\tilde{\tau})\right)(\gamma_{1,F_r}(X)-\gamma_{0,F_r}(X)-\tilde{\tau})\end{bmatrix} \\
&+ \frac{\partial}{\partial r}\Bigg(\begin{bmatrix}\exp\left(-\lambda\cdot(\gamma_{1,F_r}(X)-\gamma_{0,F_r}(X)-\tilde{\tau})\right)\cdot(-\lambda) \\ \exp\left(-\lambda\cdot(\gamma_{1,F_r}(X)-\gamma_{0,F_r}(X)-\tilde{\tau})\right)\cdot(1-\lambda\cdot(\gamma_{1,F_r}(X)-\gamma_{0,F_r}(X)-\tilde{\tau}))\end{bmatrix} \\
&\times\left(\frac{D(Y-\gamma_{1,F_r}(X))}{\pi_{F_r}(X)}-\frac{(1-D)(Y-\gamma_{0,F_r}(X))}{1-\pi_{F_r}(X)}\right)\Bigg)\Bigg] \\
&= \mathbb{E}\Bigg[\begin{bmatrix}\exp\left(-\lambda\cdot(\gamma_{1,F_0}(X)-\gamma_{0,F_0}(X)-\tilde{\tau})\right)\cdot(-\lambda) \\ \exp\left(-\lambda\cdot(\gamma_{1,F_0}(X)-\gamma_{0,F_0}(X)-\tilde{\tau})\right)\cdot(1-\lambda\cdot(\gamma_{1,F_0}(X)-\gamma_{0,F_0}(X)-\tilde{\tau}))\end{bmatrix} \\
&\times\left(\frac{\partial\gamma_{1,F_r}(X)}{\partial r}-\frac{\partial\gamma_{0,F_r}(X)}{\partial r}\right)\Big|_{r=0} \\
&-\begin{bmatrix}\exp\left(-\lambda\cdot(\gamma_{1,F_0}(X)-\gamma_{0,F_0}(X)-\tilde{\tau})\right)\cdot(-\lambda) \\ \exp\left(-\lambda\cdot(\gamma_{1,F_0}(X)-\gamma_{0,F_0}(X)-\tilde{\tau})\right)\cdot(1-\lambda\cdot(\gamma_{1,F_0}(X)-\gamma_{0,F_0}(X)-\tilde{\tau}))\end{bmatrix} \\
&\times\left(\frac{D}{\pi_{F_0}(X)}\cdot\frac{\partial\gamma_{1,F_r}(X)}{\partial r}\Big|_{r=0}-\frac{(1-D)}{1-\pi_{F_0}(X)}\cdot\frac{\partial\gamma_{0,F_r}(X)}{\partial r}\Big|_{r=0}\right) \\
&+\begin{bmatrix}\exp\left(-\lambda\cdot(\gamma_{1,F_0}(X)-\gamma_{0,F_0}(X)-\tilde{\tau})\right)\cdot(\lambda)^2 \\ \exp\left(-\lambda\cdot(\gamma_{1,F_0}(X)-\gamma_{0,F_0}(X)-\tilde{\tau})\right)\cdot(-\lambda)\cdot(2-\lambda\cdot(\gamma_{1,F_0}(X)-\gamma_{0,F_0}(X)-\tilde{\tau}))\end{bmatrix} \\
&\times\left(\frac{\partial\gamma_{1,F_r}(X)}{\partial r}-\frac{\partial\gamma_{0,F_r}(X)}{\partial r}\right)\Big|_{r=0}\times\left(\frac{D(Y-\gamma_{1,F_0}(x))}{\pi_{F_0}(X)}-\frac{(1-D)(Y-\gamma_{0,F_0}(X))}{1-\pi_{F_0}(X)}\right)\Bigg] \\
&+\begin{bmatrix}\exp\left(-\lambda\cdot(\gamma_{F_0}(X)-\tilde{\tau})\right)\cdot(-\lambda) \\ \exp\left(-\lambda\cdot(\gamma_{1,F_0}(X)-\gamma_{0,F_0}(X)-\tilde{\tau})\right)\cdot(1-\lambda\cdot(\gamma_{1,F_0}(X)-\gamma_{0,F_0}(X)-\tilde{\tau}))\end{bmatrix} \\
&\times\left(D(Y-\gamma_{1,F_0}(X))\cdot\frac{\partial}{\partial r}\left(\frac{1}{\pi_{F_r}(X)}\right)\Big|_{r=0}-(1-D)(Y-\gamma_{0,F_0}(X))\cdot\frac{\partial}{\partial r}\left(\frac{1}{1-\pi_{F_r}(X)}\right)\Big|_{r=0}\right) \\
&= 0
\end{aligned}
$$

The last equality follows by the law of iterated expectations. The first and second term cancel out since $\mathbb{E}\left[\frac{D}{\pi_{F_0}(X)}\middle|X\right]=1, \mathbb{E}\left[\frac{1-D}{1-\pi_{F_0}(X)}\middle|X\right]=1$. The third term is 0 be-

cause the nonparametric influence function is centered at 0 conditional on X. Moreover, $\mathbb{E}\left[D(Y-\mathbb{E}[Y|D=1,X]\Big|X\right]=0$ and $\mathbb{E}\left[(1-D)(Y-\mathbb{E}[Y|D=0,X]\Big|X\right]=0$ so whenever $\frac{\partial}{\partial r}\left(\frac{1}{\pi_{F_r}(X)}\right)\Big|_{r=0}$ and $\frac{\partial}{\partial r}\left(\frac{1}{1-\pi_{F_r}(X)}\right)\Big|_{r=0}$ are integrable, the fourth term is also 0, since they are measurable with respect to $\sigma(X)$. So $\frac{\partial}{\partial r}\mathbb{E}[\psi(W,\theta,\gamma_{F_r},\alpha_{F_r})]\Big|_{r=0}=0$. Observe that this result implies Neyman orthogonality with respect to the γ and α functions separately as well. To show the Neyman orthogonality with respect to γ and to set up the further results contained in Theorem 3 in Chernozhukov et al. [2020], we build the following construction. Consider the linear space of square integrable functions of X (with respect to some dominating measure), denoted as $\Gamma=L^2(\mathcal{X})$. $\mathcal{H}$ is the closed set of distributions which is a closed subset of the Banach space $L^1(\mathcal{Y}_0\times\mathcal{Y}_1\times\mathcal{X},\mu)$ under some appropriate dominating measure μ. Denote the Hadamard differential of the conditional mean function at F_0 as $\frac{\partial\gamma(F_r)}{\partial r}:\mathcal{H}\to\Gamma$. Denote the Hadamard differential for $\bar{\psi}(\gamma(F_r),\alpha_0,\theta)$ at F_0 as $\frac{\partial\mathbb{E}[\psi(W,\gamma(F_r),\alpha(F_r),\theta)]}{\partial r}:\mathcal{H}\to\mathbb{R}^2$. Finally denote the Hadamard differential of $\bar{\psi}(\gamma,\theta)$ with respect to γ as $\frac{\partial\bar{\psi}(\gamma,\alpha,\theta)}{\partial\gamma}:\Gamma\to\mathbb{R}^2$. Then the following diagram commutes by Proposition 20.9 in Van der Vaart [2000].



By Neymann orthogonality with respect to the distribution F_r, $\frac{\partial\mathbb{E}[\psi(W,\gamma(F_r),\alpha_0,\theta)]}{\partial r}\equiv 0$. $\frac{\partial\bar{\psi}(\gamma,\theta)}{\partial\gamma}$ is onto Γ which satisfies Chernozhukov et al. [2020] Theorem 3 condition iv). Then, by linearity of the Hadamard derivative and the commutativity of the above diagram it must be the case that $\frac{\partial\bar{\psi}(W,\gamma,\alpha_0,\theta)}{\partial\gamma}\equiv 0$. That is, the Hadamard derivative is the 0 function from $\Gamma\to\mathbb{R}^2$. Note that this is the case because $\frac{\partial\gamma(F_r)}{\partial r}$ is onto $L^2(\mathcal{X})$. According to the

above calculations we have, for $\delta_H := \frac{\partial \gamma_{1,F_r}}{\partial r} - \frac{\partial \gamma_{0,F_r}}{\partial r}\Big|_{r=0} \in L^2(\mathcal{X})$. Then as specified above: $\frac{\partial \mathbb{E}[\bar{\psi}(\theta,\alpha_0,\gamma)]}{\partial \gamma}(\delta_H)$ is a linear map from $L^2(X) \to \mathbb{R}^2$ in δ_H. In particular it maps to $0 \in \mathbb{R}^2$ for any $\delta_H(X)$, so it's the 0 map. Hence we verified Neyman orthogonality with respect to γ directly. □

1.I.7 Proof of Theorem 15

Lemma 25. *For $\bar{\psi}(\theta,\gamma,\alpha) = \mathbb{E}[\psi(w,\theta,\gamma,\alpha)]$ we have:*

i) $\bar{\psi}(\gamma,\alpha_0,\theta_0)$ is twice continuously Frechet differentiable in a neighborhood of γ_0.

ii) If Λ is bounded then $\forall \theta \in \Theta$, $\bar{\psi}(\gamma,\alpha_0,\theta) \leq \bar{C}\|\gamma - \gamma_0\|_{L_2}^2$.

Proof. Endow the spaces Γ with the $L^2(\mathcal{X},\mu)$ norm and $\mathbb{R}^2$ with the standard Euclidean norm $\|\cdot\|$. We directly compute the directional derivative of $\bar{\psi}(\theta,\gamma,\alpha)$ with respect to γ.

$$
\begin{aligned}
&\frac{\partial}{\partial r}\bar{\psi}(\gamma,\theta,\alpha_0) \\
=&\mathbb{E}\Bigg[\left[\frac{\exp\left(-\lambda\cdot((1-r)(\gamma_{1,0}(X)-\gamma_{0,0}(X))+r(\gamma_1(X)-\gamma_0(X))-\tilde{\tau}\right)\cdot(\lambda)^2}{\exp\left(-\lambda\cdot((1-r)(\gamma_{1,0}(X)-\gamma_{0,0}(X))+r(\gamma_1(X)-\gamma_0(X))-\tilde{\tau}\right)\cdot(-\lambda)\cdot(2-\lambda\cdot((1-r)(\gamma_{1,0}(X)-\gamma_{0,0}(X))+r(\gamma_1-\gamma_0)-\tilde{\tau}))}\right] \\
&\times\left(\frac{D(Y-(1-r)\gamma_{1,0}(X)-r\gamma_1(X))}{\pi_{F_0}(X)} - \frac{(1-D)(Y-(1-r)\gamma_{0,0}(X)-r\gamma_0(X))}{1-\pi_{F_0}(X)}\right)[(\gamma_1-\gamma_{1,0})-(\gamma_0-\gamma_{0,0})]\Bigg]
\end{aligned}
$$

where we emphasized linearity in $[(\gamma_1 - \gamma_{1,0}) - (\gamma_0 - \gamma_{0,0})]$, the discrepancy between the estimated CATE and the true one. The second order Frechet derivative, if it exists, is a bi-linear operator given below, obtained by differentiating the first order Frechet derivative with respect to r. Then:

$$
\begin{aligned}
&\frac{\partial}{\partial r}\frac{\partial \bar{\psi}(\gamma,\theta,\alpha_0)}{\partial r} \\
=&\mathbb{E}\Bigg[\Bigg\{\begin{bmatrix} \exp(-\lambda((1-r)(\gamma_{1,0}(X)-\gamma_{0,0}(X))+r(\gamma_1(X)-\gamma_0(X))-\tilde{\tau}))(-\lambda)^3 \\ \exp(-\lambda((1-r)(\gamma_{1,0}(X)-\gamma_{0,0}(X))+r(\gamma_1(X)-\gamma_0(X))-\tilde{\tau}))(-\lambda)^2(3-(1-r)(\gamma_{1,0}(X)-\gamma_{0,0}(X))+r(\gamma_1(X)-\gamma_0(X))-\tilde{\tau})) \end{bmatrix} \\
&\times\left(\frac{D(Y-(1-r)\gamma_{1,0}(X)-r\gamma_1(X))}{\pi_{F_0}(x)}-\frac{(1-D)(Y-(1-r)\gamma_{0,0}(X)-r\gamma_0(X))}{1-\pi_{F_0}(x)}\right) \\
&\times[(\gamma_1-\gamma_{1,0})-(\gamma_0-\gamma_{0,0});(\gamma_1-\gamma_{1,0})-(\gamma_0-\gamma_{0,0})] \\
&+\begin{bmatrix} \exp\left(-\lambda\cdot((1-r)(\gamma_{1,0}(X)-\gamma_{0,0}(X))+r(\gamma_1(X)-\gamma_0(X))-\tilde{\tau})\right)\cdot(\lambda)^2 \\ \exp\left(-\lambda\cdot((1-r)(\gamma_{1,0}(X)-\gamma_{0,0}(X))+r(\gamma_1(X)-\gamma_0(X))-\tilde{\tau})\right)\cdot(-\lambda)\cdot(2-\lambda\cdot((1-r)(\gamma_{1,0}(X)-\gamma_{0,0}(X))+r(\gamma_1-\gamma_0)-\tilde{\tau})) \end{bmatrix} \\
&\times[(\gamma_1-\gamma_{1,0})-(\gamma_0-\gamma_{0,0})]\left(\frac{D}{\pi_{F_0}(X)}[\gamma_1(X)-\gamma_{1,0}(X)]-\frac{1-D}{1-\pi_{F_0}(X)}[\gamma_0(X)-\gamma_{0,0}(X)]\right)\Bigg\}\Bigg]
\end{aligned}
$$

Evaluated at $r=0$ the second order directional derivatives are:

$$
\begin{aligned}
&\mathbb{E}\Bigg[\begin{bmatrix} \exp\left(-\lambda\cdot((\gamma_{1,0}(X)-\gamma_{0,0}(X))-\tilde{\tau})\right)\cdot(\lambda)^2 \\ \exp\left(-\lambda\cdot(\gamma_{1,0}(X)-\gamma_{0,0}(X))-\tilde{\tau})\right)\cdot(-\lambda)\cdot(2-\lambda\cdot((\gamma_{1,0}(X)-\gamma_{0,0}(X))-\tilde{\tau})) \end{bmatrix} \\
&\times[(\gamma_1(X)-\gamma_{1,0}(X))-(\gamma_0(X)-\gamma_{0,0}(X));(\gamma_1(X)-\gamma_{1,0}(X))-(\gamma_0(X)-\gamma_{0,0}(X))]\Bigg\}\Bigg]
\end{aligned}
$$

by the law of iterated expectations. We emphasized that the above expression, is bi-linear [15] in $(\gamma_1(X)-\gamma_{1,0}(X))-(\gamma_0(X)-\gamma_{0,0}(X)$. If the bi-linear map is continuous at $(\gamma_{1,0},\gamma_{0,0})$ with respect to the operator norm then $\bar{\psi}$ is Frechet differentiable at $(\gamma_{1,0},\gamma_{0,0})$ and the directional derivative and the Frechet derivative coincide. A sufficient condition is given by:

$$
\left\|\frac{\partial^2}{\partial r^2}\bar{\psi}(\gamma,\theta,\alpha_0)\right\|_{L_2}<\infty
$$

[15] Denote the space of linear maps from Banach spaces X to Y as $B(X,Y)$. It is itself a Banach space. Then one may identify $B(L^2(\mathcal{X})^2,B(L^2(\mathcal{X})^2;\mathbb{R}^2))$ with $B(L^2(\mathcal{X})^2\times L^2(\mathcal{X})^2;\mathbb{R}^2)$. Then the second order Frechet derivative is a bi-linear map from $L^2(\mathcal{X})^2\times L^2(\mathcal{X})^2\mathbb{R}^2$.

which translates to

$$\left\| \left[\begin{bmatrix} \exp\left(-\lambda \cdot ((\gamma_{1,0}(X) - \gamma_{0,0}(X)) - \tilde{\tau})\right) \cdot (\lambda)^2 \\ \exp\left(-\lambda \cdot (\gamma_{1,0}(X) - \gamma_{0,0}(X)) - \tilde{\tau})\right) \cdot (-\lambda) \cdot (2 - \lambda \cdot ((\gamma_{1,0}(X) - \gamma_{0,0}(X)) - \tilde{\tau})) \end{bmatrix} \right. \right.$$
$$\left. \left. \times [(\gamma_1(X) - \gamma_{1,0}(X)) - (\gamma_0(X) - \gamma_{0,0}(X)); (\gamma_1(X) - \gamma_{1,0}(X)) - (\gamma_0(X) - \gamma_{0,0}(X))] \right] \right\|_{L_2} < \infty$$

Then Frechet differentiability follows from Holder's inequality with $p = q = 2$. Under a slightly stronger condition which holds uniformly over $r \in [0, 1]$ one can obtain stronger results. Then Theorem 3 ii) in Chernozhukov et al. [2020] can be applied and we have:

$$\bar{\psi}(\gamma, \alpha_0, \theta_0) \leq C \|\gamma_1(X) - \gamma_{1,0}(X) - (\gamma_0(X) - \gamma_{0,0}(X))\|_{L^2}^2 \leq C \left\| \begin{bmatrix} \gamma_1(X) - \gamma_{1,0}(X) \\ \gamma_0(X) - \gamma_{0,0}(X) \end{bmatrix} \right\|_{L^2, E}^2$$

where the E denotes the Euclidean norm on $\mathbb{R}^2$. More generally consider $C(\lambda)$ defined below:

$$C(\lambda) := \left\| \sup_{r \in (0,1)} \left\{ \begin{bmatrix} \exp\left(-\lambda \cdot ((1-r)(\gamma_{1,0}(X) - \gamma_{0,0}(X)) + r(\gamma_1(X) - \gamma_0(X)) - \tilde{\tau})\right) \\ \exp\left(-\lambda \cdot ((1-r)(\gamma_{1,0}(X) - \gamma_{0,0}(X)) + r(\gamma_1(X) - \gamma_0(X)) - \tilde{\tau})\right) \end{bmatrix} \right. \right.$$
$$\left. \left. \begin{bmatrix} (\lambda)^2 & 0 \\ 0 & (-\lambda)(2 - \lambda \cdot ((1-r)(\gamma_{1,0}(X) - \gamma_{0,0}(X)) + r(\gamma_1 - \gamma_0) - \tilde{\tau})) \end{bmatrix} \right\| \right\|_E$$

For a general bound here the constant depends on $C(\lambda)$. If Λ is compact then we can afford a representation of the theorem which is uniform across values for λ_0 which gives a much stronger version of the approximating function in λ and gets rid of some terms. For $\bar{C} = \sup_{\lambda \in \Lambda} C(\lambda)$ then $\psi(\gamma, \theta, \alpha_0) \leq C \|\gamma - \gamma_0\|_{L_2}^2$ and Frechet differentiability in a neighborhood of λ_0 follows in a straightforward way from the continuity of $C(\lambda)$ and the compactness of Λ. □

Remark 26. *Compactness of Λ would follow, for example, from Assumption 4 which restricts λ to be finite. We note that a condition in the form of $\bar{C} < \infty$ is sufficient and does not require compactness of Λ.*

Lemma 27 ($\sqrt{n}$ - consistency)**.** *proposition Let Assumption 5 hold. Then*

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i \in I_k} g(W_i, \theta, \hat{\gamma}_{-k}) + \phi(W_i, \tilde{\theta}_{-k}, \hat{\gamma}_{-k}, \hat{\alpha}_{-k}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(W_i, \theta, \gamma_0, \alpha_0) + o_P(1)$$

Proof. The proof mirrors the blueprint of Theorem 15 in Chernozhukov et al. [2020]. We have:

$$\begin{aligned}
& g(W_i, \theta_0, \hat{\gamma}_{-k}) + \phi(W_i, \hat{\gamma}_{-k}, \tilde{\theta}_{-k}, \hat{\alpha}_{-k}) - \psi(W_i, \gamma_0, \theta_0, \alpha_0) \\
&= \underbrace{g(W_i, \theta_0, \hat{\gamma}_{-k}) - g(W_i, \theta_0, \gamma_0)}_{\hat{R}_{1i,-k}} \\
&+ \underbrace{\phi(W_i, \theta_0, \hat{\gamma}_{-k}, \alpha_0) - \phi(W_i, \theta_0, \gamma_0, \alpha_0)}_{\hat{R}_{2i,-k}} \\
&+ \underbrace{\phi(W_i, \tilde{\theta}_{-k}, \gamma_0, \hat{\alpha}_{-k}) - \phi(W_i, \theta_0, \gamma_0, \alpha_0)}_{\hat{R}_{3i,-k}} \\
&+ \underbrace{\phi(W_i, \tilde{\theta}_{-k}, \hat{\gamma}_{-k}, \hat{\alpha}_{-k}) - \phi(W_i, \tilde{\theta}, \gamma_0, \hat{\alpha}_{-k}) + \phi(W_i, \hat{\gamma}_{-k}, \alpha_0, \theta_0) - \phi(W_i, \gamma_0, \alpha_0, \theta_0)}_{\hat{\Delta}_{i,-k}} \\
&+ g(W_i, \theta_0, \gamma_0) + \phi(W_i, \theta_0, \gamma_0, \alpha_0) \\
&- \psi(W_i, \theta_0, \gamma_0) \\
&= \hat{R}_{1i,-k} + \hat{R}_{i2,-k} + \hat{R}_{i3,-k} + \hat{\Delta}_{i,-k}
\end{aligned}$$

Conditioning on the set not used in the nonparametric estimation we have:

$$\begin{aligned}\mathbb{E}[\hat{R}_{1i,-k}+\hat{R}_{2i,-k}|I_k^c] &= \int_{\mathcal{X}}(g(w,\theta_0,\hat{\gamma}_{-k},\alpha_0)+\phi(w,\theta_0,\hat{\gamma}_{-k},\alpha_0))dF_0(w)\\ &= \int_{\mathcal{X}}\psi(w,\theta_0,\hat{\gamma}_{-k},\alpha_0)dF_0(w)\\ &= \bar{\psi}(\theta_0,\hat{\gamma}_{-k},\alpha_0)\end{aligned}$$

The third term's expected value, conditional on the subsample is given by $\mathbb{E}[\hat{R}_{i3,-k}|I_k] = \int_{\mathcal{X}}\phi(W_i,\tilde{\theta}_{-k},\gamma_0,\hat{\alpha}_{-k})dF_0(w) = 0$. Finally consider the term:

$$\frac{1}{\sqrt{n}}\sum_{i\in I_c}\hat{R}_{1i,-k}+\hat{R}_{i2,-k}+\hat{R}_{i3,-k}-\mathbb{E}[\hat{R}_{1,-k}+\hat{R}_{2,-k}|I_k^c]+\mathbb{E}[\hat{R}_{1,-k}+\hat{R}_{2,-k}|I_k^c]$$

Now by Kennedy et al. [2020] Lemma 2 we have:

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i\in I_c}\hat{R}_{1i,-k}+\hat{R}_{i2,-k}-\mathbb{E}[\hat{R}_{1,-k}+\hat{R}_{2,-k}|I_k^c] &= O_P(\|\psi(W_i,\theta_0,\hat{\gamma}_k,\alpha_0)-\psi(W_i,\theta_0,\gamma_0,\alpha_0)\|_L^2)\\ &= O_P(\|\hat{\gamma}_k-\gamma_0\|_L^2)\end{aligned}$$

where the last equality follows form proposition 25 ii).

Again by Kennedy et al. [2020] Lemma 2

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i\in I_k}\hat{R}_{i3,-k}-\mathbb{E}[\hat{R}_{i3,-k}|I_k] &= O_P(\|\phi(W_i,\tilde{\theta}_{-k},\gamma_0,\hat{\alpha}_{-k})-\phi(W_i,\theta_0,\gamma_0,\alpha_0)\|_{L^2})\\ &= O_P(\|\hat{\alpha}-\alpha_0|_L^2)+O_P(\|\tilde{\theta}-\theta_0\|_{\mathbb{R}^2})\end{aligned}$$

since $\phi(\cdot)$ is linear in α and differentiable in θ. Then Assumption 5 guarantees that these last two terms are $o_P(1)$. Furthermore, by Proposition 25 ii) for n sufficiently large we

have:

$$\mathbb{E}[\hat{R}_{1,-k} + \hat{R}_{2,-k}|I_k] \leq \sqrt{n}C\|\hat{\gamma}_k - \gamma_0\|^2$$

for $\bar{C}$ given in proposition 25. A similar argument shows $\frac{1}{\sqrt{n}}\sum_{i\in I_k^c}\Delta_{i,-k} = o_P(1)$. If that's the case, we conclude that:

$$\frac{1}{\sqrt{n}}\sum_{i\in I_k} g(W_i, \theta_0, \hat{\gamma}_{-k}) + \phi(W_i, \tilde{\theta}_k, \hat{\gamma}_k, \hat{\alpha}_{-k}) = \frac{1}{\sqrt{n}}\sum_{i\in I_k}\psi(W_i, \gamma_0, \theta_0, \hat{\alpha}_0) + o_P(1)$$

□

Lemma 28 (Jacobian consistency)**.** *For Jacobian G of the debiased moment conditions:*

$$G = \mathbb{E}[D\psi(w, \theta_0, \gamma_0, \alpha_0)] = \mathbb{E}\left[\frac{\partial}{\partial\theta}\psi(w, \theta_0, \gamma_0, \alpha_0)\right] \tag{A1.26}$$

and $\hat{\theta} \xrightarrow{p} \theta_0$ we have $\|\frac{\partial\hat{\psi}(\hat{\theta})}{\partial\theta} - G\| = o_P(1)$.

Proof. First observe that at $\gamma = \gamma_0$ and $\alpha = \alpha_0$:

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial}{\partial\theta}\psi(w, \theta, \gamma, \alpha)\right] &= \mathbb{E}\left[\frac{\partial}{\partial\theta}g(w, \theta, \gamma, \alpha)\right] + \mathbb{E}\left[\frac{\partial}{\partial\theta}\phi(w, \theta, \gamma, \alpha)\right] \\
&= \mathbb{E}\left[\frac{\partial}{\partial\theta}g(w, \theta, \gamma)\right] + 0 \\
&= \mathbb{E}\left[\frac{\partial}{\partial\theta}g(w, \theta, \gamma)\right]
\end{aligned}$$

by the law of iterated expectations. (N.B: if α_0 is the propensity score than this holds in a neighborhood of the true F_0). Now, to show the result we verify the conditions in Lemma

17 of Chernozhukov et al. [2020]. First notice that for $\frac{\partial g(w,\theta,\gamma)}{\partial \theta}$, each of the functions:

$$\theta \mapsto -1$$

$$\theta \mapsto 0$$

$$\theta \mapsto -\exp(-\lambda(\tau(x)-\tilde{\tau}))(\tau(x)-\tilde{\tau})$$

$$\theta \mapsto -\exp(-\lambda(\tau(x)-\tilde{\tau}))(\tau(x)-\tilde{\tau}))^2$$

is continuously differentiable in θ at θ_0. The first two are constants and the other two derivatives are, respectively:

$$\exp(-\lambda(\tau(x)-\tilde{\tau}))(\tau(x)-\tilde{\tau})^2$$

$$\exp(-\lambda(\tau(x)-\tilde{\tau}))(\tau(x)-\tilde{\tau}))^3$$

Hence if $\mathbb{E}[\exp(-\lambda_0(\tau(x)-\tilde{\tau}))(\tau(x)-\tilde{\tau})^2] < \infty$ and $\mathbb{E}[\exp(-\lambda_0(\tau(x)-\tilde{\tau}))(\tau(x)-\tilde{\tau})^3] < \infty$. In particular Assumption 2 is a sufficient condition for locally bounded derivatives which satisfies Assumption 4 ii) in Chernozhukov et al. [2020]. Assumption 4 iii), namely $\int(\frac{\partial g_j}{\partial \theta_l}(w,\theta,\hat{\gamma_k}) - \frac{\partial g_j}{\partial \theta_l}(w,\theta,\gamma_0))dF_0(w)$ follows from the continuous mapping theorem and continuity of the the maps above with respect to $\gamma(\cdot) = \tau(\cdot)$ in the $\|\cdot\|_{L_2}$ norm. □

We are finally ready to prove 15 using the lemmas above.

Theorem 15 (Asymptotic normality of θ)**.** *Let Assumptions 1–5. For $\hat{\theta}$ defined in Equation*

(1.17)*:*

$$\begin{aligned}
\sqrt{n}(\hat{\theta}-\theta_0) &\xrightarrow{d} \mathcal{N}(0,S)\\
S &:= (G)^{-1}\Omega(G')^{-1}\\
G &:= \mathbb{E}[D_\theta\psi(w,\theta,\gamma_0,\alpha_0)]\\
\Omega &:= \mathbb{E}[\psi(w,\theta_0,\gamma_0,\alpha_0)\psi(w,\theta_0,\gamma_0,\alpha_0)^T]
\end{aligned}$$

and $D_\theta\psi(\cdot)$ is the Jacobian of the augmented moment condition with respect to the parameters in θ.

Denote $\hat{G} = \frac{\hat{g}(w,\hat{\theta},\hat{\gamma})}{\partial\theta}$. First note that by Lemma 28 we have $\|\hat{G}-G\| = o_P(1)$. Then, like in Chernozhukov et al. [2018] we have:

$$\begin{aligned}
\hat{G}^{-1}-G^{-1} &= (G+\hat{\Delta}_n)^{-1}-G^{-1}\\
&= (G+\hat{\Delta}_n)^{-1}(GG^{-1})-(G+\hat{\Delta}_n)G^{-1}\\
&= (G+\hat{\Delta}_n)^{-1}(G-(G+\hat{\Delta}_n))G^{-1}\\
&= (G+\hat{\Delta}_n)^{-1}\hat{\Delta}_nG^{-1}
\end{aligned}$$

Then like in Chernozhukov et al. [2018] from the basic matrix inequalities we have:

$$\begin{aligned}
\|\hat{G}^{-1}-G^{-1}\| &= \|(G+\hat{\Delta}_n)^{-1}\hat{\Delta}_nG^{-1}\|\\
&= \|(G+\hat{\Delta}_n)^{-1}\|\cdot\|\hat{\Delta}_n\|\cdot\|G^{-1}\|\\
&= O_P(1)\cdot o_P(1)\cdot O_P(1)\\
&= o_P(1)
\end{aligned}$$

Now by the central limit theorem and Lemma 27 we have:

$$\begin{aligned}
&\frac{1}{|K|}\sum_{k\in K}\left(\frac{1}{\sqrt{n}}\sum_{i\in I_k} g(W_i,\theta,\gamma_0)+\phi(W_i,\tilde{\theta}_{-k},\hat{\gamma}_{-k},,\hat{\alpha}_{-k})\right)\\
&=\frac{1}{|K|}\sum_{k\in K}\frac{1}{\sqrt{n}}\sum_{i\in I_k}\psi(W_i,\theta,\gamma_0,\alpha_0)+o_P(1)\xrightarrow{d}\mathcal{N}(0,\Omega)
\end{aligned}$$

where $\Omega=\mathbb{E}[\psi(w,\theta_0,\gamma_0,\alpha_0)\psi(w,\theta_0,\gamma_0,\alpha_0)]$. Finally observe that a standard GMM Taylor linearization gives:

$$\begin{aligned}
\sqrt{n}\begin{bmatrix}\nu-\nu_0\\ \lambda-\lambda_0\end{bmatrix}=&\left\{\frac{\partial}{\partial\theta}\hat{\psi}(w,\theta_0,\hat{\gamma},\hat{\alpha})'V\frac{\partial}{\partial\theta}\hat{\psi}(w,\theta_0,\hat{\gamma},\hat{\alpha})\right\}^{-1}\frac{\partial}{\partial\theta}\hat{\psi}(w,\theta_0,\hat{\gamma},\hat{\alpha})'V\\
&\times\frac{1}{|K|}\sum_{k\in K}\left(\frac{1}{\sqrt{n}}\sum_{i\in I_k} g(W_i,\theta,\hat{\gamma}_{-k})+\phi(W_i,\tilde{\theta}_{-k},\hat{\gamma}_{-k})\right)\\
&=(G'VG)^{-1}G'V\left(\frac{1}{|K|}\sum_{k\in K}\frac{1}{\sqrt{n}}\sum_{i\in I_k}\psi(W_i,\theta,\gamma_0,\alpha_0)\right)+o_P(1)\xrightarrow{d}\mathcal{N}(0,S)
\end{aligned}$$

which is the desired result.

1.I.8 Auxiliary Lemmas

Lemma 29. *(Kennedy et al. [2020]-Lemma 2)*

Let $\hat{g}(\cdot)$ be a function estimated from the I_k^c sample and evaluated on the I_k sample. Then $(\mathbb{P}_n-\mathbb{P})(\hat{g}-g_0)=O_P\left(\frac{|\hat{g}-g_0|}{\sqrt{n}}\right)$.

Proof. The proof follows from independence of I_k and I_k^c, the computation of conditional variance and Markov's inequality. See Kennedy et al. [2020] for a detailed treatment. □

1.J Additional Figures and Examples

In this section I include some additional visualizations and examples:

Example 30. *To visualize Corollary 20 consider the case where the dimension of the covariate space is* $k=2$. *The original data is normal* $\mathcal{N}(\mu,\Sigma)$ *with* $\mu=(4,3)^T$ $\Sigma=\begin{bmatrix} 2 & 0.5 \\ 0.5 & 2 \end{bmatrix}$. $\tau(x)=X^T\beta$ *is linear with* $\beta=(4,1)^T$. *Experimental* $ATE=18.98$. *Target* $ATE=15$.

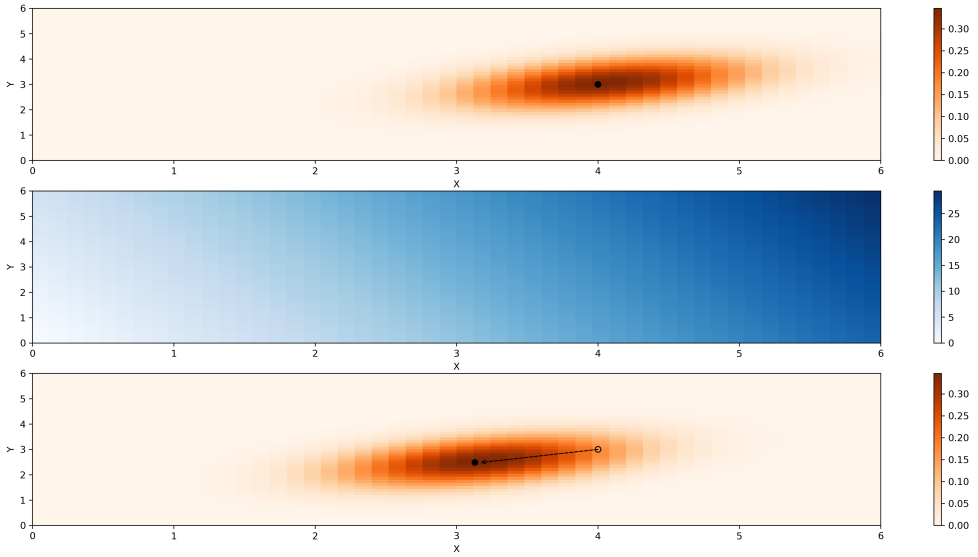


Figure 1.J.1: least favorable distribution for normally distributed data. First panel in red shows the density of $\mathcal{N}(\mu,\Sigma)\sim\mathcal{N}\left(\mu=\begin{bmatrix}4\\3\end{bmatrix};\Sigma=\begin{bmatrix}2 & 0.5\\ 0.5 & 2\end{bmatrix}\right)$, the experimental distribution. The second panel shows $CATE=X^T\beta$, linear in X with $\beta=(4,1)^T$. The third panel shows the parameter shift of the *least favorable distribution*.

Here $\lambda_0=0.396$. $\mu^*=(3.1288,2.4852)$. *The KL divergence, for two multivariate normal distributions* $(\mu_1,\Sigma_1),(\mu_2,\Sigma_2)$ *is given by:*
$KL(X_1||X_2)=\frac{1}{2}\left[\log\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right)-k+(\mu_2-\mu_1)^T\Sigma_2^{-1}(\mu_2-\mu_1)+tr(\Sigma_2^{-1}\Sigma_1)\right]$. *One could always compute the value of the KL divergence applying the nonparametric formula*

$$\delta^*=\int_{\mathcal{X}}\exp(-\lambda_0(\tau(X)-\tilde{\tau}))d\mu_X$$

or in this case, the "parametric" formula given by the KL divergence between two normal distributions.[16] *In this example both ways of computing the correspond to* $\delta^*=0.789$

[16]The "parametric" formula to compute the KL divergence would not be valid in general since the *least*

corresponding to the mean shift illustrated above.

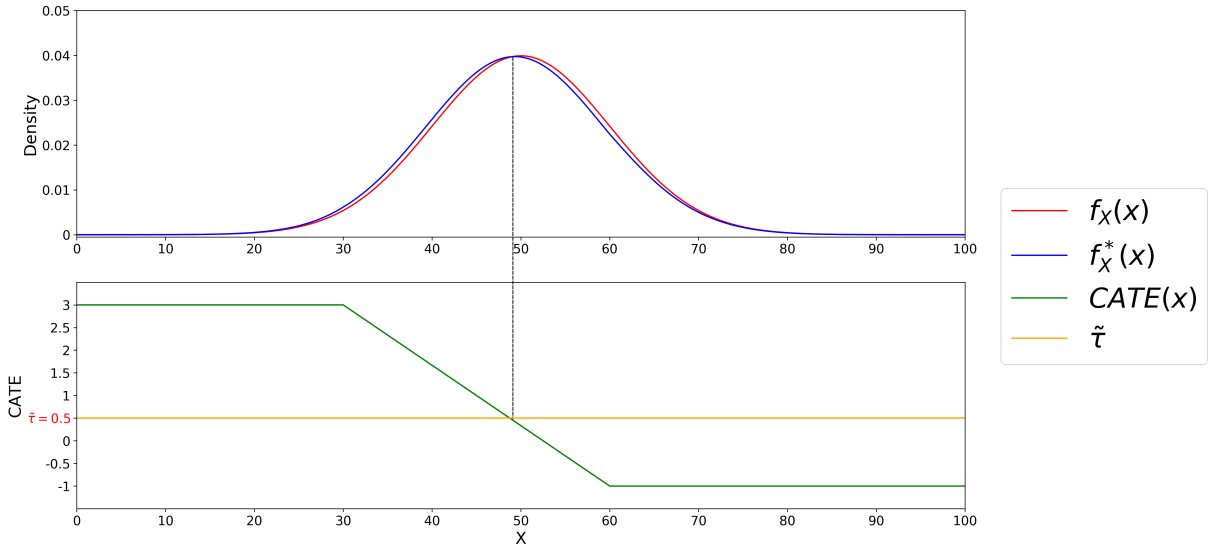


Figure 1.J.2: Piece-wise Linear CATE, experimental distribution is $\mathcal{N}(50, 10^2)$. Experimental ATE is 0.433, while $\tilde{\tau} = 0.5$. Because the experimental ATE is lower than the least favorable , F_X^* down-weights F_X on the subset of $\mathbb{R}$ where the $\tau(x)$ is greater than $\tilde{\tau}$ and up-weights it where it's lower. The blue curve is the closest curve to the red one, in KL-divergence, among the ones that satisfy $\tau \geq 0.5$.

favorable distribution may belong to a different class than the experimental distribution. Conversely, the "nonparametric" formula is always valid.

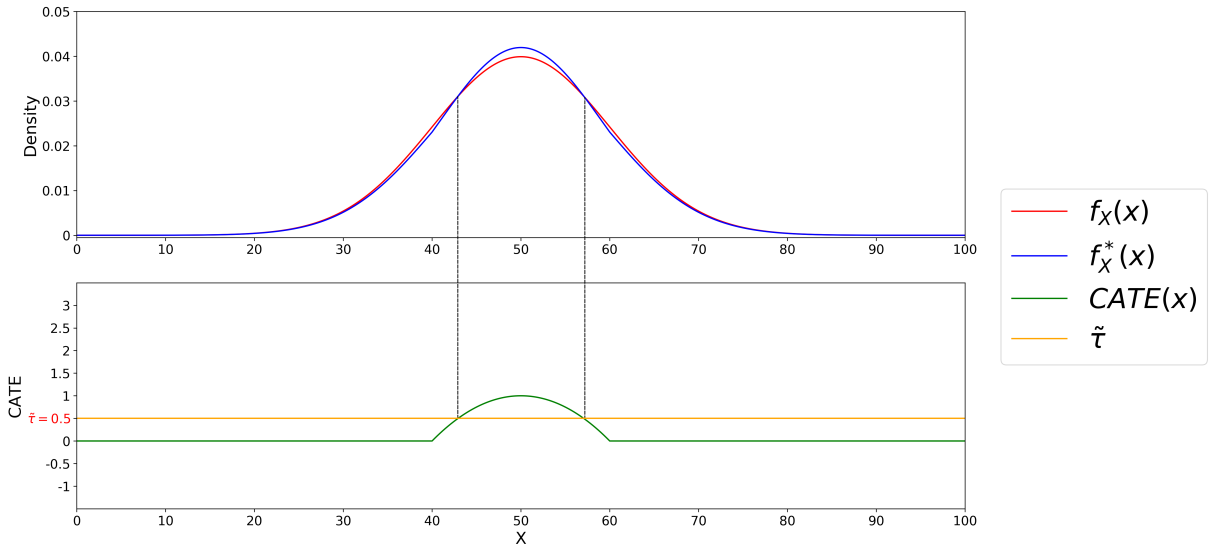


Figure 1.J.3: Piecewise Quadratic CATE, experimental distribution is $\mathcal{N}(50, 10^2)$. Experimental ATE is 0.484 while $\tilde{\tau} = 0.5$. Because the experimental ATE is lower than the least favorable , F_X^* down-weights F_X on the subset of $\mathbb{R}$ where the $\tau(x)$ is smaller than $\tilde{\tau}$ and up-weights it where it's greater. The blue curve is the closest curve to the red one, in KL-divergence, among the ones that satisfy $\tau \geq 0.5$.

Example 31. *Let's now see graphically how to construct an example for a one dimensional continuous variable example. In Figures 1.J.2 and 1.J.3 conditional treatment effects, given the 1-dimensional variable X are in green, the experimental distribution is $\mathcal{N}(50, 10^2)$ is in red. Suppose that the policy-maker's wants to maintain the claim $ATE \leq 0.5$. The experimental ATE and the "least favorable" ATE are obtained by integrating the green curve $\tau(x)$ against the red curve $dF_X(x)$ (which has density $f_X(x)$) and the blue curve $dF_X^*(x)$ (which has density $f_X^*(x)$ respectively. The blue curve is the closest distribution to the experimental distribution in red, as measured by the KL divergence, that delivers the "least favorable " ATE $\tilde{\tau} = 0.5$.*

Chapter 2

Generalized Robustness Test: Coefficient Stability across Causal Specifications

2.1 Introduction

Many applied economics studies use observational data to tease out causal relationship between policy variables of interest. Often, the identification of the main estimand requires imposing assumption on the causal mechanism that governs the data generating process. Because such assumptions involve constraints on the distribution of unobservables, they may not be fully testable. On the other hand, because of the crucial role that the assumptions play in the identification strategy, many researchers propose robustness exercises to convince their research audience that their estimate are sensible and robust.

A typical tool used in these situations is a "robustness check" which involves a comparison of estimated coefficients across several specifications. While robustness checks typically involve comparisons across multiple regression specifications, the heuristic of

this exercise can be captured with a simple example involving a single comparison. A researcher is interested in delivering estimates of the causal effect of a policy variable X on an outcome variable Y. A set of control variables $\{Z, W\}$ is available. Given the observational data, the researcher specifies an identification strategy for the causal effect of interest. The baseline specification only uses variables in Z as controls. As they anticipates some criticism regarding their baseline identification strategy, they seek to corroborate the causal estimate with a "robustness check" regression that uses both Z and W as additional controls. The most straightforward example of a robustness checks is described in Table 2.1.1.

Table 2.1.1: Robustness check table: $\beta_{X|Z}$ is the OLS regression coefficient when the set of Z controls is included in the regression, $\beta_{X|ZW}$ is the OLS regression coefficient when both the set Z and W are included in the regression

	Baseline	Robustness Check
Coefficient of interest	$\beta_{X\|Z}$	$\beta_{X\|ZW}$
	se($\beta_{X\|Z}$)	se($\beta_{X\|ZW}$)
Z controls	Yes	Yes
W controls	No	Yes

The heuristic for interpreting the table above is the following: if the baseline identification (which controls for variables in Z) is correct and if the additional covariates W in the augmented regression are valid controls, the coefficient of the augmented regression and the baseline regression should be very close. Conversely, if the estimated coefficients are sufficiently different, one should reconsider the validity of the identification strategy.[1]

In the stylized robustness exercise above, the identification strategy and the robustness exercise are based on a combination of causal (or structural) assumptions and *ad-hoc* functional form assumptions (linearity). While the causal assumptions are usually motivated by domain knowledge and/or institutional background, the functional form

[1]This particular use of the robustness check is particularly popular in the scope of difference in difference estimators where the alternative specification is triple differences. For example Muralidharan and Prakash [2017] conducts such an exercise in estimation of the effect of providing bicycles on secondary school enrollment. For additional examples see Baez et al. [2017] and Cai [2016].

assumptions are usually motivated by convenience: they usually result in a simplification of the estimation strategy or reduce the computational burden. The combination of these two types of assumptions has implications for what type of causal parameters one can hope to identify as well as for the robustness exercises associated to identification. I argue that the functional form assumption are at best superfluous and at worst can result in misleading robustness exercises when they are mis-specified. I show that the heuristic of comparing the coefficients of two (or more) regressions can be grounded exclusively in the causal (or structural) assumptions. As such, it is valid nonparametrically, enlarging the class of models and causal parameters that it can be applied to. Finally, I propose a robustness test that upgrades the heuristic of Table 2.1.1 into a formal statistical procedure. My proposed test is nonparametric and removes the limitations arising from functional form specifications.

In a recent paper, Oster [2017] has drawn attention to the practice of robustness checks in applied work. In her piece, she cautions against a naïve comparison between the coefficients of the baseline and augmented regression. Instead, she advocates for a weighted comparison that captures variation in the R^2 of the regression as a measure of the unobserved variation that the new covariate explains. While the robustness exercise in her case is geared towards sensitivity analysis, it does not constitute a robustness test *per-se*. Because it requires the user to specify a parameter, the degree of proportionality between selection on observables and unobservables, it can be though of as a complementary tool to evaluate robustness. The robustness test that I propose instead, does not require such a parametrization.

At the core of the "robustness check" in Table 2.1.1 there is a notion that, if the baseline identification with Z is valid, controlling for the additional covariates in W is redundant. Either omitting or including W would not bias the identification. Covariates that may introduce bias in the estimation of the causal effect of interest when included in the

regression have been known as "bad controls" Angrist and Pischke [2008]. Their inclusion in the regression in this context should be cautioned against since it would undermine the very nature of the robustness exercise: while the baseline causal identification may have been correctly specified, the introduction of a bad control must necessarily produce a shift in the coefficients of interest since W introduces endogeneity Chen and Pearl [2015]. Based on the heuristic, the "robustness check" would raise concern over an originally valid identification strategy because the estimated $\beta_{X|ZW}$ and $\beta_{X|Z}$ are different. But this gives the opposite answer of what the robustness check was designed to detect! Lu and White [2014] and Chen and Pearl [2015] and Cinelli et al. [2020] provide an excellent discussion on "bad controls" that the interested reader can consult. In the main body of this paper I will focus on control variables W that are valid. A discussion of the identification failures arising from the inclusion of a"bad control" is presented in the Appendix.

The authors also highlight the conditions for informative *versus* misleading robustness tests describing which covariates cannot serve the purpose of detecting violations of identification. In a similar spirit, Chen and Pearl [2015] provide a characterization of the conditions that guarantee that a robustness check is valid and informative. The authors' notions rule out robustness checks that hinge on a set of controls W which "opens a spurious path between the causal variable of interest and the outcome". This class of problematic controls includes, but is not limited to alternative outcome variables directly caused by the variable interest. When a "bad control" is added in the robustness check regression, the baseline model is correctly identified but the augmented model is not. Hence, a shift in the causal coefficient of interest cannot be regarded as informative.

Lu and White [2014] proposed a robustness test in linear models, providing an advancements relative to the "robustness checks" heuristic above. Their procedure requires imposing a combination of causal and functional form assumptions, making a rejection in their test hard to interpret. It can signal either a violation of the hypothesized causal

structure or a mis-specification of the functional form. The two types of violations have very different consequences for a researcher. The former violation requires that the researcher reassesses her identification strategy, the latter can be simply accommodated by a more flexible model. Because their test cannot distinguish between these two cases, it is not specific. A small example of this feature is given below 32.[2]

Example 32. *Consider the simple example where the data is generate according to:*

$$\begin{aligned} W &= v \\ Z &= q + u \\ X &= Z^2 + W + s \\ Y &= \beta \cdot X^2 + Z \cdot X + \gamma \cdot Z + u \end{aligned}$$

and v, u, s, q *are all independent* $\mathcal{N}(0,1)$*. Then* $AMTE(x_0)$ *defined below is nonparamterically identified using either* Z *only or both* Z *and* W*. On the other hand, the conditional mean functions are nonlinear. Even with such a mild non-linearity, the test proposed by Lu and White [2014], out of* $M = 1000$ *trials, rejects all the time .*

The good news is that the driving principle in the robustness test does not requires imposing the functional form assumption and is more generally valid in a nonparametric setting. There are two main advantages to constructing a test based on nonparametric identification. First, it exclusively tests for causal assumptions while accommodating a flexible treatments of the functional form. Second, it eliminates the need for pre-testing procedures to evaluate functional form assumptions which may affect post-estimation inference in an unclear way.

This paper is also related to the literature on testing the implications of the

[2]Lu and White [2014] suggest pre-testing for linearity of the structural function and/or linearity of the conditional mean function as separate tests from the ones aimed at detecting causal mis-specification.

conditional independence assumptions. From an identification point of view, it is possible to directly test whether the distribution of the data satisfies a conditional independence assumption for both conditioning sets of covariates Z and (Z, W). There is a large body of work on testing unconfounded-ness and conditional independence in a nonparametric setting, spanning a variety of nonparametric estimation techniques including kernel Cai et al. [2019], Huang et al. [2016], matching de Luna and Johansson [2014] and weighting Donald et al. [2014]. In practice though, an applied researcher may not be interested in testing conditional independence *per-se*. Rather, they only care about whether failures of nonparametric identification may affect the value of the causal estimand of interest. Because this latter quantity is a functional of the distribution of the data. testing for conditional independence is strictly more general than testing for equality of two nonparametrically estimated coefficients. Because it targets the particular functional of interest, I view my approach as striking a balance between generality and specificity. As such, it is complementary to testing the more general unconfounded-ness assumption in a nonparametric setting and the user may determine which, among the available procedures, suits best the needs of her specific research design.

The objective of this paper is to provide a method to disentangle the robustness test from the functional form assumption and provide a transparent procedure that is statistically sound. After defining the Average Marginal Treatment Effect ($AMTE$), the causal estimand of interest, I develop a fully nonparametric method to both compute the $AMTE$ and conduct a robustness test of causal identification. The test follows from a simple heuristic. When the model is correctly specified, there are two procedures that identify the same population quantity. Thus, asymptotically, the relative sample estimators must converge to the same quantity. Alternatively, when the causal structure is not correctly specified, at least one of the estimators will be biased and the robustness check will falsify the equivalence of the two proposed identification strategies. For simplicity I

focus on $AMTE$ though the framework is flexible enough to accommodate a wide variety of functionals of the counterfactual distribution.

The contribution of this method is three-fold. First, it delivers a transparent procedure that captures the intuitive heuristic of coefficient stability discussed above. Second, it abstracts from the problem of identification through functional form. Third, it can be easily computed and adapted to semi-parametric settings. Section 2 introduces notation, the causal parameter of interest and the main identification result. Section 3 discusses the estimation method and the testing procedure. Section 4 derives their asymptotic properties. Section 5 briefly concludes.

2.2 Set Up and Identification Results

To fix ideas, consider a researcher interested in the effect of a continuous policy variable X on an outcome of interest Y. The outcome is jointly caused by the policy variable as well as other unobservable factors U. One may think of a particular realization of $U = u$ as a state of the world that can make the policy variable X on the outcome more or less effective. Similarly, one may interpret a realization $U = u$ as an indexing of individuals such that for each unobserved u_i there is an idiosyncratic reference policy environment that individual i experiences. One can consider the rather general non-separable structural equation:

$$Y = C(X, U) \tag{2.1}$$

Here C is an unspecified structural function. For the definition of the Average Marginal Treatment Effect, which will be the main policy effect of interest, it is convenient to require $C \in \mathcal{C}^1(\mathcal{X} \times \mathcal{U})$, the space of continuously differentiable functions that take values in $\mathbb{R}$. Let the support of random variables X and U be denoted $\mathcal{X} \subseteq \mathbb{R}^{d_X}$ and $\mathcal{U} \subseteq \mathbb{R}^{d_U}$ respectively. The researcher observes a data set composed of n observations $(X_i, Y_i, Z_i, W_i)_{i=1}^n$ where

Z, W are additional (sets of) control variables. The researcher is interested in the Average Marginal Treatment Effect across the population of individuals as defined below.

Definition 33. ***Average Marginal Treatment Effect***

Let $f_u(u)$ denote the density of the absolutely continuous distribution function F_U. The Average Marginal Treatment Effect at the pre-specified point $x_0 \in \mathcal{X}$ is defined as:

$$AMTE(x_0) := \int \nabla C(x_0, u) f_U(u) = \left[\int \frac{\partial C}{\partial x^1}(x_0, u) f_u(u) du, \cdots, \int \frac{\partial C}{\partial x^{d_X}}(x_0, u) f_u(u) du \right]' \tag{2.2}$$

Equation (2.2) *defines $AMTE(x_0)$ as the effect of increasing the policy variable(s) X by one unit on the level of the outcome variable Y, starting from pre-specified level x_0, averaged across the unobserved, individual specific, realization of the policy environment $U = u$.*

Remark 34. *The notion of $AMTE(x_0)$ generalizes the notion of the slope coefficient(s) of a linear model. In fact one can immediately see that if $Y = X^T\beta + u$ then $AMTE(x_0) = \beta$ for all x_0.*

To understand why a researcher may be interested in the $AMTE(x_0)$ as a policy parameter, consider a policy-maker who is deciding whether a given level x_0 should be increased or decreased to maximize the outcome Y. For example, for policy-making in education, X may represent the years of education and the point x_0 may be the existing threshold for mandatory schooling. The policymaker may be interested in the effect of marginally raising the threshold, which was previously set at x_0, on a policy outcome like employment or incarceration. Then the value of $AMTE(x_0)$ is the meaningful parameter of interest. In many applications where the policy-maker may want to set optimal thresholds, a natural comparison for $AMTE(x_0)$ is 0, which translates to a first order condition for

optimally setting the policy threshold. Naturally, it is only meaningful for $x_0 \in \mathcal{X}$ since, for points outside of the support, there is no hope to identify the effect of a variation that is never observed. For a given $x_0 \in \mathcal{X}$, how should one identify $AMTE(x_0)$? In observational studies, individual's realizations of X_i may depend on the individual's u_i, which are unobserved. One may denote such endogenous dependence by writing $X(u)$. Returning to the education example, the choice to drop-out or to remain in school for X years is potentially correlated with other individual characteristics that are part of u. See Lochner and Moretti [2004] for a reference in this context.

More generally, whenever u jointly determines X and Y, if there are no additional variables which control for the dependence between X and Y through u, the $AMTE(x_0)$ will not be identified. The causal diagram[3] clarifies this point.

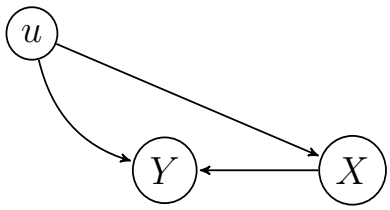


Figure 2.2.1: Non-identifiable $AMTE(x_0)$

Suppose the researcher has additional covariates available: $\{Z_i, W_i\}_{i=1}^n$ with $(Z, W) \in \mathcal{Z} \times \mathcal{W}$ and $\mathcal{Z} \times \mathcal{W} \subseteq \mathcal{R}^{d_Z} \times \mathcal{R}^{d_W}$.

She hypothesizes that the data generating process follows one of two causal models: a baseline model M_B and an augmented model M_R, whose causal diagrams are displayed in Figures 2.2.2 and 2.2.3 below.[4] According to M_B, if the researcher knew the joint

[3]Causal diagrams are a useful tool to present the identifying assumptions without resorting to functional form. They have been popularized by the work of Pearl [2000]. If there are paths connecting X and Y other than the direct arrow $X \to Y$ the identification of the causal effect of interest hinges on the possibility of blocking such paths with an appropriate selection of control variables. A path between X and Y that is unblocked results in biased identification of the causal object of interest for almost all parametrizations of the model.

[4]In general, there are multiple causal graphs that would imply the same conditional independence restrictions on the joint distribution of (X, Z, W, Y). All causal graphs in the same equivalence class are observationally equivalent and they all allow identification of $AMTE(x_0)$.

distribution of the observables (X, Z, Y), she would be able to identify $AMTE(x_0)$ by appropriately controlling for Z. Under M_B, the additional information in W is not strictly necessary for the identification of $AMTE(x_0)$. Contrast this case with the augmented model M_R. Under M_R, controlling for Z is not sufficient to identify $AMTE(x_0)$; the full set of controls (Z, W) is needed.

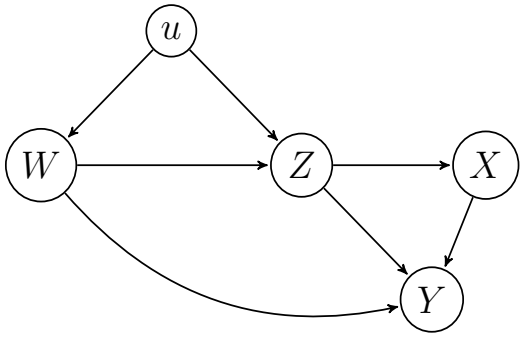


Figure 2.2.2: Example of M_B: The quantity $AMTE(x_0)$ is identified by including either Z or (Z, W) in the control set

The researcher anticipates some skepticism about the identification of $AMTE(x_0)$ using only Z as control in the baseline specification. She seeks to conduct a robustness exercise to convince the audience that using M_B correctly identifies the $AMTE(x_0)$. She includes the information carried by the variable W in the estimation of the $AMTE(x_0)$ and compares the results with the baseline model. If the identification strategy is valid when either just $\{Z\}$ or the full set $\{Z, W\}$ are included, then both procedures should identify the same population parameter. If instead including just Z would result in biased estimation of $AMTE(x_0)$, the two procedures will generally lead to different results. Importantly, this heuristic did not rely on any specific functional form. As I show below, there is no need to impose parametric restrictions on the function $C(\cdot)$ like in Lu and White [2014] because the equality of the estimated $AMTE(x_0)$ holds nonparametrically.

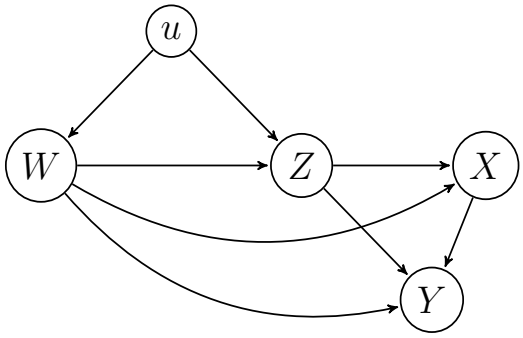


Figure 2.2.3: Example of M_R: The quantity $AMTE(x_0)$ is identified only by including (Z,W) in the control set. Simply including Z would result almost surely in a bias

2.2.1 Nonparametric Identification

In this section I formalize the researcher's heuristic by showing that, under model M_B, the $AMTE(x_0)$ is nonparametrically identified by controlling for either $\{Z\}$ or the joint set $\{Z,W\}$.

Definition 35. *(Potential Outcomes Y_x) Given $C \in \mathcal{C}^1(\mathcal{X},\mathcal{U})$ define the potential outcome Y_x as the level of Y that would be attained if X is set at x in $C(X,U)$*

$$Y_x := C(x,U) \tag{2.3}$$

Remark 36. *The definition is a matter of indexing. Naturally, $C(x,U)$ is a measurable map from $\Omega \to \mathbb{R}$ so $Y_x \circ U$, for each $x \in \mathcal{X}$ is a random variable*[5]*. The reader will find the notation closely related to the potential outcome notation of Imbens and Rubin [2015]. In the case where $\mathcal{X} = \{0,1\}$ the random variables Y_x can be written in the familiar format Y_1, Y_0. In structural work the reader may have encountered the following notation: $Y_1 = C(1,U_1), Y_0 = C(0,U_0)$. Because the dimension of U is unrestricted, the notation $Y = C(x,U)$ is without loss of generality as as one can take $U = (U_0,U_1)$.*

[5]Trivially notice the following. $U : \Omega \to \mathbb{R}$ and $C(x,\cdot) : \mathbb{R} \to \mathbb{R}$ so the composition $Y_x := C(x,\cdot) \circ U$ is a map from $\Omega \to \mathbb{R}$. $C(x,\cdot)$ is measurable because it's the restriction of C to the set $\{x\} \times \mathcal{U}$. Then Y_x is the composition of two measurable maps, hence measurable.

I now show the key identification result. The distinctive feature of the present identification framework will require that the $AMTE(x_0)$ be identified regardless of particular parametric assumption on the structural function $C(\cdot)$. That is, identification only relies on the conditional independence assumptions modeled in the causal diagrams above and some mild regularity conditions that allow $AMTE(x_0)$ to be defined. A causal quantity is nonparametrically identified if it can be expressed as a function of the conditional distributions of observable and/or estimable quantities. I explicitly consider the possibility of an estimable quantity because this includes estimators based on a control function approach or a two-step procedure. In that context, while a particular variable is not directly observable, it is still estimable. I now introduce an assumption that would be needed to prove the identification results as well as the main asymptotic results in Section 4.

Assumption 6. *The following conditions hold:*
i) (Y_i, X_i, W_i, Z_i) *are independent and identically distributed*
ii) $\mathcal{X} \times \mathcal{Z} \times \mathcal{W}$ *is compact*
iii) $m_0 := \mathbb{E}[Y|X,Z,W] \in C^p(\mathcal{X} \times \mathcal{Z} \times \mathcal{W})$ *with* $p > 3/2$[6].

Assumption 6 i, ii) contain standard assumptions on the data generating process; iii) requires sufficient smoothness of the conditional expectation function.

Proposition 37 (Identification)**.** *Let* $U \perp\!\!\!\perp X|Z,W$ *and Assumption 1 hold. Then: i) the* $AMTE(x_0)$ *is nonparametrically identified by the following formula:*

$$AMTE(x_0) = \int \frac{\partial E[Y|X,Z,W]}{\partial x}(x_0, z, w) dF_{ZW}(z,w) \tag{2.4}$$

ii) Let instead $U \perp\!\!\!\perp X|Z$ *and Assumption I hold. Then* $AMTE(x_0)$ *is nonparametrically*

[6]The reader may recognize this condition as an adapted version of the conditions in Chen [2007]

identified by the following formula:

$$AMTE(x_0) = \int \frac{\partial E[Y|X,Z]}{\partial x}(x_0, z) dF_Z(z) \tag{2.5}$$

Proof. The proof leverages the fact that $AMTE(x_0)$ is a functional of the counterfactual distribution, which can be identified through the joint distribution of the observables under the conditional independence statements of Proposition 37. See Appendix for details. □

Corollary 38. *Let $U \perp\!\!\!\perp X|Z,W$ and $U \perp\!\!\!\perp X|Z$ both hold, together with assumption 1. Then:*

$$\begin{aligned} AMTE(x_0) &= \int \frac{\partial \mathbb{E}[Y|X,Z,W]}{\partial x}(x_0, z, w) f_{zw}(z, w) dz dw \\ &= \int \frac{\partial \mathbb{E}[Y|X,Z]}{\partial x}(x_0, z) f_z(z) dz \end{aligned}$$

Proof. It follow immediately from the two conditional independence statements and Proposition 37. □

Note that, since the conditional independence statements do not imply one another, we need to require both to hold jointly to obtain the result in Corollary 38. If only one, or neither of them holds, then the expressions in Equation (2.4) and (2.5) in general identify distinct objects. The identification of $AMTE(x_0)$ hinges on specifying the relevant conditional independence assumption appropriately. In the baseline model M_B two conditional independence statements hold and the $AMTE(x_0)$ can be identified in two ways. In the augmented model M_R, controlling for $\{Z\}$ only would not eliminate all confounding factors between X and Y and therefore the weighted derivative of the conditional mean will generally not identify the causal effect of interest. In practice, the researcher has no idea of which model is correct. The key idea is that the difference in the two population objects is still informative as it raise concerns about the robustness of the

identification. The comparison of the two quantities in 38 is exactly the type of exercise that is performed by the familiar parametric robustness checks of Table 2.1.1. In the linear case, it corresponds to the equality of $\beta_{X|Z}$ and $\beta_{X|ZW}$ discussed in Lu and White [2014]. The difference here is that only nonparametric identification matters.

At a high level, the notion of coefficient stability is appealing because it is both intuitive and is applicable to nonparametric contexts. On the other hand, it is important to discuss what can be learned from a coefficient stability type of exercise: when is the comparison based on Corollary 38 informative for the researcher? If we construct a nonparametric test based on it, what would we learn from a rejection? In the discussion below I consider the more parsimonious notion of S-identifiability, which focuses on using the set of variables S to allow identification of $AMTE(x_0)$.

Definition 39. *Let S be a set of observables.*
A model M is S-identifiable if $AMTE(x_0)$ is identified by $\int \left(\frac{\partial \mathbb{E}[Y|X,S](x_0,s)}{\partial x}\right) f_S(s)ds$.

For example, Z-identifiability requires that the baseline model nonparametrically identifies the causal effect of interest by controlling for Z. Because Z is the researchers baseline specification, this is the condition that the coefficient stability robustness test aims to falsify. Consider a model that is indeed Z-identifiable. Now suppose we run a robustness test with additional control W based on the heuristic that the $AMTE(x_0)$ that is identified including both W and Z as controls must be the same as the $AMTE(x_0)$ identified using just Z as a control. The researcher observes that the test provides evidence against the null hypothesis. How should she interpret the result? Clearly, one of two situations may have arisen:

- The baseline model was correctly identified but the introduction of W has introduced bias resulting in different values for the two estimates $AMTE$s
- The baseline model was not correctly identified and the introduction of W estimates

a different effect

The first scenario may happen if W contains what is known as a "bad control". In order for the robustness check or test to be informative, it is paramount to rule out such a case, as discussed in Chen and Pearl [2015]. It suffices to requite that the underlying model satisfies:

$$M \text{ is } Z\text{-identifiable} \implies M \text{ is } ZW\text{-identifiable}. \tag{2.6}$$

This condition is the desirable property of a robustness check needed to avoid the bad control scenario above.

In practice this assumption guides the researcher to specify a set of robustness check variables that would not introduce additional bias. This is not very restrictive, since the hypothesis contained in the causal diagram should guide in the selection of controls. One can then formalize the robustness exercise in as a null hypothesis, involving two estimable quantities:

$$H_0 : \int \left(\frac{\partial \mathbb{E}\left[Y|X = x_0, Z\right]}{\partial x} \right) f_z(z)dz = \int \left(\frac{\partial \mathbb{E}\left[Y|X = x_0, Z, W\right]}{\partial x} \right) f_{zw}(z, w)dzdw \tag{2.7}$$

One can view the formulation in Equation (2.7) as a middle ground between two approaches. The first would be to test for the full equality of two (or more) conditional distributions, say $(Y|X, Z, W = w) \stackrel{d}{=} (Y|X, Z)$ for any $w \in \mathcal{W}$ but this may clearly be false even when Equation (2.7) is true. The second one would be the approach taken by Lu and White [2014] where equality of $\beta_{X|Z}$ and $\beta_{X|ZW}$ is considered. As I will show in a later section, when the data generating process is sufficiently nonlinear, $\beta_{X|Z} \neq \beta_{X|ZW}$ (and in particular, neither will be equal to $AMTE(x_0)$) even when (2.7) indeed holds. This is problematic because the rejection is entirely due to non-linearity but it will be interpreted as a failure of the identification strategy. The approach presented in this paper is indeed a middle ground

between the two described above. First it directs power to the object of interest rather than to the conditional independence assumption itself (which is helpful insofar it allows to test the identification of $AMTE(x_0)$). Second, because of the flexibility in accommodating nonlinear DGPs, it does not lead to spurious rejections.

2.3 Estimation

In this section I consider how to derive estimators for the causal estimand of interest. Under H_0, which is the testable implication derived from Corollary 38 the comparison of these two (or more) estimators will serve as the statistic for the nonparametric coefficient stability robustness test. I restrict the attention to the core example of the robustness exercise, which involves a comparison between the estimator of $AMTE(x_0)$ obtained using a small set of controls $\{Z\}$ and the one obtained using a larger set of controls $\{Z,W\}$. While in principle the nonparametric identification allows to design robustness tests based on several nonparametric and semiparametric estimation methods, in practice it may be difficult to obtain their joint asymptotic distribution which is required to carry out the test. For this reason I focus on a robustness test based on sieve estimators, for which the asymptotic theory is well-developed. Extensions of the robustness test presented in this paper to accommodate other methods could leverage the development of asymptotic theory for different classes of nonparametric estimators developed for example in Athey et al. [2016], Chetverikov et al. [2021] and Farrell et al. [2021]. This extension would provide even more flexibility on how a robustness test for coefficient stability can be carried out, which may be important in a context where high dimensional control variables are used. I leave this exciting direction for future research.

This choice of sieve estimators is motivated by the fact that the causal estimand of interest, $AMTE(x_0)$, can be expressed as a functional of the conditional mean function.

Under the baseline model, Corollary 38 says that $AMTE(x_0)$ can be obtained as a functional of either $m_{ZW}(x,z,w) := \mathbb{E}[Y|X=x,Z=z,W=w]$ or $m_Z(x,z) := \mathbb{E}[Y|X=x,Z=z]$. This suggests a plug-in approach. The techniques developed in the large body of literature on functionals of sieve estimators carry over directly, greatly simplifying the treatment.[7] Before discussing the estimation of the conditional mean function itself, we can show that the $AMTE(x_0)$ is a continuous linear functional with respect to the strong norm. Let's first introduce some notation.

Definition 40. *Let C_S^d be the space of k times continuously differentiable functions on $\mathcal{X} \times \mathcal{S}$. The "strong" norm on C_S^d is given, for a generic $m \in C_S^d$ by:*

$$\|m\|_d := \max_{\lambda \leq d} \sup_{\mathcal{X}\times\mathcal{S}} \left| \frac{\partial^{|\lambda|} m}{\partial x^{|\lambda|}} \right|$$

Definition 41. *Bounded functional*

A functional $L : \mathcal{H} \to \mathbb{R}$ is bounded with respect to the norm $\|\cdot\|_{\mathcal{H}}$ if there exists a constant M such that:

$$|L(h)| \leq M\|h\|_{\mathcal{H}}$$

for all $h \in \mathcal{H}$.

For a weighting function $\omega : \mathcal{S} \to \mathbb{R}_+$, $\int_{\mathcal{S}} \omega(s)ds = 1$, $\Gamma_S : \mathcal{C}_S^d(\mathcal{X} \times S) \to \mathbb{R}$ be the functional from the space of d-continuously differentiable functions to $\mathbb{R}$ defined below:

$$\Gamma_{S,\omega}(m) := \int \frac{\partial m}{\partial x}(x,s)|_{x=x_0}\omega(s)ds$$

This is a composition of three linear maps: the differentiation operator, the evaluation map at x_0 and the integration against the "weight" $\omega(s)$. It turns out that the under the

[7] A similar estimation method is suitable for nonparametric instrumental variables framework and for robustness tests that arise from identification through instrumental variables. The reader can find the relevant results in Chen [2007].

string norm, $\Gamma_{S,\omega}$. This is in general not true under weaker norms because the evaluation functional is not continuous with respect to the weak norm.

Lemma 42 (Bounden-ness)**.** *For any S and any ω with $\int_{\mathcal{S}} \omega(s)ds$, $\Gamma_{S,\omega}$ is bounded with respect to the "strong" norm.*

Proof. See Appendix. □

Proposition 43 (Linearity and Continuity)**.** *For any S, and any ω $\Gamma_{S,\omega}$ is a continuous linear functional with respect to the strong norm.*

Proof. To show linearity it suffices to show $\Gamma_{S,\omega}(\alpha m_1 + \beta m_2) = \alpha\Gamma_{S,\omega}(m_1) + \beta\Gamma_{S,\omega}(m_2)$ for $\alpha, \beta \in \mathbb{R}, m_1, m_2 \in \mathcal{C}^k(\mathcal{X} \times \mathcal{S})$.

$$\begin{aligned}
\Gamma_{S,\omega}(\alpha m_1 + \beta m_2) &= \int \frac{\partial \alpha m_1 + \beta m_2}{\partial x}(x,s)|_{x=x_0}\omega(s)ds \\
&= \int \left[\alpha \frac{\partial m_1}{\partial x}(x,s) + \beta \frac{\partial m_2}{\partial x}(x,s)\right]|_{x=x_0}\omega(s)ds \\
&= \int \left[\alpha \frac{\partial m_1}{\partial x}(x,s)|_{x=x_0} + \beta \frac{\partial m_2}{\partial x}(x,s)|_{x=x_0}\right]\omega(s)ds \\
&= \alpha \int \frac{\partial m_1}{\partial x}(x,z,w)|_{x=x_0}\omega(s)ds + \beta \int \frac{\partial m_2}{\partial x}(x,s)|_{x=x_0}\omega(s)ds \\
&= \alpha\Gamma_{S,\omega}(m_1) + \beta\Gamma_{S,\omega}(m_2)
\end{aligned}$$

This is straightforward: the functional of interest in the composition of the partial derivative map, the evaluation map and integration against a particular weight $\omega(s)$ all of which are linear maps. A linear functional on a Banach space is continuous if and only if it's bounded. Lemma 42 shows that $\Gamma_{S,\omega}(\cdot)$ is bounded with respect to the strong norm which finishes the proof. □

If we set $S = \{Z\}$ and $S = \{Z, W\}$ respectively, the choice of weights

$$\omega(z,w) = f_{ZW}(z,w), \omega(z) = f_Z(z)$$

respectively implies, through Proposition 37, that $U \perp\!\!\!\perp X|Z,W$ implies $AMTE(x_0) = \Gamma_{ZW}(m_{ZW})$ and $U \perp\!\!\!\perp X|Z$ implies $AMTE(x_0) = \Gamma_Z(m_Z)$.

In the context of this paper the plug-in approach boils down to using:

$$\widehat{AMTE}(x_0) = \Gamma(\widehat{m_S}) \tag{2.8}$$

We now turn to the implementation for linear sieve spaces. There are four conceptual steps in the estimation process. For the generic conditional mean function m_S with control set S:

- Obtain the nonparametric sieve estimators for the conditional mean functions based on $m_S(x,s)$.
- Compute the sieve estimator of the derivative, denoted $m'_S(x,s)$. For linear sieve spaces, this will amount to taking the linear combination of the images of the basis functions under the operator $\frac{\partial}{\partial x}$ with linear coefficients estimated for $m_S(x,s)$
- Evaluate the estimator at the desired x_0, obtaining $m'_S(x_0,s)$
- Integrate the functions $m'_S(x_0,s)$ against the empirical distribution of S to recover $AMTE(x_0)$

The approach for computing the derivative is similar to the one undertaken by Cattaneo et al. [2018] who consider estimating the density function of interest by estimating the distribution function by local polynomials and extracting the (slope) linear coefficient of the local polynomial regression. By using a linear sieve space, one obtains an approximate representation of a function $m_{0S}(x,s)$ as the linear combination of suitably chosen basis functions. The estimated derivative has the same linear coefficients but the basis functions are the images of the original basis function under the partial derivative operator. After obtaining the nonparametric estimator $\hat{m}_{ZW}(x,z,w)$ the theory for the the plug-in estimator $\Gamma(\hat{m}_{ZW})$ as well as the robustness test is relatively straightforward.

2.3.1 Notation & Sieve Estimator

In this section we discuss the estimation of the conditional means using linear sieve spaces. To lighten the notation we can focus on the estimation of m_{ZW} and consider the special case where $d_z = 1, d_w = 1$ and note that the general case follows similarly. Let $\Theta = C^d(\mathcal{X} \times \mathcal{Z} \times \mathcal{W})$. In this case our estimation target m_0 is the function: m_{ZW}. In order to approximate $m_0 \in \Theta$ it is enough to consider the approximating sieve spaces Θ_n given by

$$\Theta_{ZW,n} := \text{span}(\phi_j(x)\psi_k(z)\varphi_l(w)) \text{ for } j = 1, 2, \cdots, J_n;\ k = 1, 2, \cdots, K_n;\ l = 1, 2, \cdots, L_n$$

where the functions $\phi_j(x), \psi_k(z), \varphi_l(w)$ are a deterministic set of basis functions. We can define $\Theta_{Z,n}$ in an analogous way. I discuss the choices of an appropriate basis of functions to Section 2.3.2. Each Θ_n is the linear span of basis functions, each of which is the product of three univariate functions[8]. As noted by Chen [2007] the dimension of the sieve space is given by $\dim(\Theta_n) = L_n \cdot J_n \cdot K_n$. As such, a generic function $m \in \Theta_n$ can always be expressed as a linear combination of the basis elements in the following way:

$$m(x, z, w) = \sum_{j=1}^{J_n} \sum_{l=1}^{L_n} \sum_{k=1}^{K_n} \phi_j(x)\psi_k(z)\varphi_l(w)\beta_{j,k,l} \tag{2.9}$$

How does one guarantee that the function m_0 can be approximated by the increasingly complex functions in Θ_n? We have the following result:

Theorem 44 (Stone-Weierstrass)**.** *Let K be a compact metric space and $A[k]$ an sub-algebra of functions, (that is, a closed subset of $\mathcal{C}(K, \mathbb{R})$) which separates points and contains a constant function. The $A[k]$ is dense in $\mathcal{C}(K, \mathbb{R})$.*[9]

[8] For estimation purposes it is sufficient to consider the tensor product of simpler functions spaces, the construction is in the Appendix.

[9] Here dense is meant with respect to the topology of uniform convergence induced by the sup-norm on $\mathcal{C}(K, \mathbb{R})$. This is the reason to require the point separation property which says that for $k \neq k'$ in K there exist $a \in A[k]$ such that $a(k) \neq a(k')$.

This result says that, if the basis functions in the collection are sophisticated enough to tell apart , a large number of them can guarantee a arbitrarily good approximation over the whole domain of the function. Then by the Stone-Weierstrass, for a generic $m \in \Theta$ there exist N such that for all $n \geq N$ we have:

$$\|m_0 - m_N\|_d < \epsilon$$

in the strong norm, where $m_N \in \Theta_N$. Theorem 44 is often stated with $A[k]$ is the sub-algebra of polynomials but there are several other classes of sieve spaces that may be more useful in practice. Because the estimation strategy is not the main feature of the paper, I consider two basic cases, polynomials and splines. Polynomials are constructed as: $\Theta_n = \text{Pol}_{J_n}(\mathcal{X}) \times \text{Pol}_{K_n}(\mathcal{Z}) \times \text{Pol}_{L_n}(\mathcal{W})$. Constructing splines is a bit more involved and we discuss it in the Appendix. Because I use splines to estimate derivatives, I require that the degree of the splines sieve space is at least 2 since that would ensure that the derivative is continuously differentiable. No such requirement is needed for polynomials since they are smooth.

In order to estimate the coefficients $\beta := [\beta_{1,1,1}, \beta_{1,1,2}, \cdots, \beta_{J_n,K_n,L_n}]$ from the observations $\{Y_i, X_i, Z_i, W_i\}_{i=1}^n$ we use a nonparametric version of least squares. The construction proceeds as detailed below. Denote $\phi_i^{J_n}$ the $(J_n \times 1)$ vector:

$$\phi_i^{J_n} := (\phi_1(X_i), \phi_2(X_i), \phi_{J_n}(X_i))'$$

Now denote the $(J_n \cdots K_n \cdot L_n \times 1)$ vector Φ_i:

$$\Phi_i := \text{vec}\left(\phi_i^{J_n} \otimes \psi_i^{K_n} \otimes \varphi_i^{L_n}\right)$$

and for all $i = 1 \cdot n$ gather these into the $(J_n \cdot K_n \cdot L_n \times n)$ matrix:

$$\Phi := (\Phi_1, \Phi_2, \cdots, \Phi_n)'$$

Then the nonparametric sieve estimator of m_0 is given by:

$$\hat{m}(x, z, w) = \sum_{j=1}^{J_n} \sum_{l=1}^{K_n} \sum_{k=1}^{L_n} \phi_j(x)\psi_k(z)\varphi_l(w)\hat{\beta}_{j,l,k}$$

where $\hat{\beta}_{j,k,l}$ is the element of the vector of estimated nonparametric least-square coefficients $\hat{\beta}_{LS} = (\Phi'\Phi)^{-1}\Phi'Y$ corresponding to the (j, k, l)-product of basis functions.

2.3.2 Selection of the Sieve Spaces

In practice, different sieve spaces may be more appropriate for the construction of functions that satisfy particular desirable properties. These may include positivity, monotonicity or other shape restrictions. Other spaces may be chosen for computational convenience. In particular $Pol_{J_n}(\mathcal{X})$ is known to suffer from near-multicollinearity issues when the J_n grows. Intuitively this is due to the fact that while polynomials separate points, high powers like x^{46} behave too similarly to x^{47} for values of $x \in \mathcal{X}$ that are not very large. An interesting thing to note is that, while each approximating class $A[k]$ that satisfies the assumption of the Stone-Weierstrass theorem is sufficient to obtain a uniform approximation, in finite samples there might be an interest in using more than one approximating class simultaneously.[10] To build intuition, consider the 1-dimensional case where we are trying to approximate the function $f(x) = x \cdot \sin(x)$. If we take $Pol_{J_n}(x)$ or $TriPol_{J_n}(x)$, the class of polynomials and trigonometric polynomials respectively, $f \notin \Theta_n$ for any n. Of course this in not an issue asymptotically because each class can approximate

[10] I thank Yixiao Sun for this meaningful suggestion.

f uniformly. On the other hand, observe that $f \in Pol_1(x)Tri_1(x)$ so the tensor product of these two classes will have a better performance. Conversely, the finite sample performance of each sieve space taken separately may be unsatisfactory since a function like $f(x)$ may be poorly approximated by polynomials (because of the $\sin(x)$ component) and poorly approximated by trigonometric polynomials (because of the linear x component) when J_n is small. In some extensions of this work, I consider a richer procedure in which all basis functions are considered simultaneously and a LASSO-type procedure can be used to select redundancies. This idea is broadly inspired by the idea of over-parametrization and it would be very interesting to explore in future research. Here I simply note that one can, in general, over-parametrize the sieve spaces by considering the larger space

$$\Theta_{ZW,n} = \left\{ \sum_{q} \sum_{j=1}^{J_n} \sum_{k=1}^{K_n} \sum_{l=1}^{L_n} \phi_{jq}(x)\phi_{kq}(z)\phi_{lq}(w)\beta_{qjkl} \right\}$$
$$\text{where } j,k,l = \{1,2,\cdots J_n\}, q \in \{Pol_{J_n}, TriPol_{J_n}, Spl_{J_n}, Gauss_{J_n}, \cdots\}$$

Here q is ranging over a variety of collections of standard sieve spaces. Different sieve spaces are know to be best suited to approximate well functions with particular properties. A brief review of their properties is given in Appendix B, for a thorough treatment consider Chen [2007].

2.3.3 Estimation of the conditional mean functions $m_Z(\cdot)$ and $m_{ZW}(\cdot)$ and its derivatives

The discusssion in the previous section has highlighted the simplicity of the sieve approach. The estimated conditional mean function in Θ_n has the representation below:

$$\begin{aligned}\hat{m}(x,z,w) &= \sum_{j=1}^{J_n}\sum_{k=1}^{K_n}\sum_{l=1}^{L_n}\phi_j(x)\psi_k(z)\varphi_l(w)\hat{\beta}_{j,k,l} \\ &= \sum_{j=1}^{J_n}\sum_{k=1}^{K_n}\sum_{l=1}^{L_n}\phi_j(x)\psi_k(z)\varphi_l(w)((\Phi'\Phi)^{-1}\Phi'Y)_{j,k,l}\end{aligned}$$

Throughout this paper I will denote $\hat{\beta}^Z$ as the coefficient vector of the β_{LS} sieve estimator obtained including only the (set) of variables Z in the regression, i.e, the baseline regression. $\hat{\beta}^{ZW}$ will denote the coefficient vector of the sieve estimator for the robustness check regression.

First consider the sieve estimator for the conditional mean function:

$$\hat{m}_{ZW}(x,z,w) = \sum_{j=1}^{J_n}\sum_{k=1}^{K_n}\sum_{l=1}^{L_n}\phi_j(x)\psi_k(z)\varphi_l(w)\hat{\beta}_{j,k,l}$$

Then, an approach would be to consider the estimator for the partial derivative of m with respect to x given by:

$$\hat{m}'_{ZW}(x) = \sum_{j=1}^{J_n}\sum_{k=1}^{K_n}\sum_{l=1}^{L_n}\phi'_j(x)\psi_k(z)\varphi_l(w)\hat{\beta}_{j,k,l}$$

where $\phi'_j(\cdot)$ is the derivative of $\phi_j(\cdot)$ with respect to x.

Remark 45. *Here the key hyper-parameter is the number of terms in the series estimator, namely J_n, K_n and L_n which informs the bias-variance trade-off for the estimation of m_Z and m_{ZW} respectively. While the standard procedure is to use cross validation to select such*

a hyper-parameter, the CV-rule is built on optimizing the performance for the conditional mean. Arguably though, the true targets (in the spirit of targeted learning) are different: the average of the derivative of the conditional mean function is the actual estimator for $AMTE(x_0)$ and it is not clear that the optimal choice of J_n, K_n, L_n carries over to the optimal choice of terms for the average derivative. I leave this issue for further research.

Now that we have developed the construction of the (linear) sieve estimator above estimator and we have shown that the functional of interest is linear in the conditional mean and continuous with respect to the string norm, we can obtain the results for the joint distribution of the two estimators of $AMTE(x_0)$ under the null hypothesis that the model is both Z and ZW-identifiable. After that, constructing our robustness test is straightforward.

2.4 Asymptotic Properties

In this section I characterize the asymptotic normality of the estimator as well as the asymptotic distribution of the nonparametric robustness test. Relying on the key identification result of 38 we will construct two estimators of $AMTE(x_0)$ which under the null will converge to the same quantity. Fortunately, as shown in the previous section, we can interpret the $AMTE(x_0)$ as a linear functional of the sieve estimators $\hat{m}_{ZW}(x,z,w)$ and $\hat{m}_Z(x,z)$: it is the composition of the derivative map, evaluation at x_0 and integration against the empirical distribution. As a result the frameworks discussed in Newey [1997] and Chen [2007] are applicable. For convenience I will denote $T_n^Z = J_n \cdot K_n$ and $T_n^{ZW} = J_n \cdot K_n \cdot L_n$.

2.4.1 Main Result

Consider a generic functional a which maps the sieve estimator to $\mathbb{R}^2$. In order to obtain the main result of consistency and asymptotic normality of the $AMTE(x_0)$ estimator I introduce a key result by Newey [1997]. Mirroring the treatment in Chen [2007], define the following quantities below (with dimensions in parenthesis). Moreover, define $\epsilon_Z := Y - \mathbb{E}[Y|X, Z]$, $\epsilon_{ZW} := Y - \mathbb{E}[Y|X, Z, W]$ to be the conditional expectation residuals.

$$\underset{(T_n^{ZW}\times T_n^{ZW})}{Q^{ZW}} := \mathbb{E}[\Phi(X,Z,W)\Phi(X,Z,W)'] \tag{2.10}$$

$$\underset{(T_n^{Z}\times T_n^{Z})}{Q^{Z}} := \mathbb{E}[\Phi(X,Z)\Phi(X,Z)'] \tag{2.11}$$

$$Q := \begin{bmatrix} Q_{ZW} & 0 \\ 0 & Q_Z \end{bmatrix} \tag{2.12}$$

$$\underset{(T_n^{ZW}\times T_n^{ZW})}{\Sigma^{ZW}} := \mathbb{E}[\Phi(X,Z,W)\Phi(X,Z,W)'\mathbb{E}(\epsilon_Z^2|X,Z,W)] \tag{2.13}$$

$$\underset{(T_n^{Z}\times T_n^{Z})}{\Sigma^{Z}} := \mathbb{E}[\Phi(X,Z)\Phi(X,Z)'\mathbb{E}(\epsilon_{ZW}^2|X,Z)] \tag{2.14}$$

$$\underset{(T_n^{ZW}\times T_n^{Z})}{\Sigma_\times} := \mathbb{E}[\Phi(X,Z,W)\Phi(X,Z)'\mathbb{E}(\epsilon_{ZW}|X,Z,W)\cdot\mathbb{E}(\epsilon_Z|X,Z)] \tag{2.15}$$

$$\Sigma := \begin{bmatrix} \Sigma^Z & \Sigma_\times' \\ \Sigma_\times & \Sigma^{ZW} \end{bmatrix} \tag{2.16}$$

$$\underset{(T_n^{Z}\times 1)}{A^{Z}} := \frac{\partial a(\Phi(X,Z)'\beta)}{\partial\beta}\Big|_{\beta_n^*} \tag{2.17}$$

$$\underset{(T_n^{ZW}\times 1)}{A^{ZW}} := \frac{\partial a(\Phi(X,Z,W)'\beta)}{\partial\beta}\Big|_{\beta_n^*} \tag{2.18}$$

$$\underset{(T_n^{Z}+T_n^{ZW}\times 2)}{A} := \begin{bmatrix} (A^Z)' & 0 \\ 0 & (A^{ZW})' \end{bmatrix}' \tag{2.19}$$

$$\underset{(2\times 2)}{V_n} := A'Q^{-1}\Sigma Q^{-1}A \tag{2.20}$$

$$\zeta_0(T_n^{ZW}) := \sup_{\mathcal{X}\times\mathcal{Z}\times\mathcal{W}} \|\Phi(x,z,w)\| \tag{2.21}$$

$$\zeta_d(T_n^{ZW}) := \max_{\lambda_x+\lambda_z+\lambda_w\le d}\left(\sup_{\mathcal{X}\times\mathcal{Z}\times\mathcal{W}}\left\| \frac{\partial\Phi(x,z,w)}{\partial x^{\lambda_x}\partial z^{\lambda_z}\partial w^{\lambda_w}}\right\|\right) \tag{2.22}$$

Consider the following set of assumptions:

Assumption 7. *Suppose:*

i) $\sup_{(x,y,z)\in\mathcal{X}\times\mathcal{Z}\times\mathcal{W}}\mathbb{E}[Y-m_0(X,Z,W)|X,Z,W]^4<\infty$ *and* $Var(Y|X,Z,W)>\epsilon>0$

ii) $\inf_n \min_{T_n} \lambda(Q) > 0$

iii) there exist $\alpha_Z, \alpha_{ZW} > 0, \beta_n^*$ *such that, for* $T_n^{ZW} = J_n \cdot K_n \cdot L_n$, $T_n^Z = K_n \cdot L_n$,
$\inf_{g \in \Theta_{Z,n}} \|g - m_0\|_d = \|\Phi_{T_n^{ZW}}(x,z,w)'\beta_{ZW,n}^* - m_0\|_d = O((T_n^{ZW})^{-\alpha_{ZW}})$ *and*
$\inf_{g \in \Theta_{ZW,n}} \|g - m_0\|_d = \|\Phi_{T_n^Z}(x,z)'\beta_{Z,n}^* - m_0\|_d = O((T_n^Z)^{-\alpha_Z})$

Assumption 8. *Suppose:*
i) $\lim_{n\to\infty} \frac{T_n^{ZW}\zeta_0(T_n^{ZW})^2}{n} = 0$ *and* $a(h)$ *is linear in* h *or*
ii) $\lim_{n\to\infty} \frac{(T_n^{ZW})^2\zeta_s(T_n^{ZW})^4}{n} = 0$ *and there is a linear map* $D(h,\tilde{h})$ *such that is linear in* $h \in \Theta$ *and there are numbers* $c_1, c_2, \epsilon > 0$ *so that for any* $\tilde{h}$ *and* $\bar{h}$ *with* $\|\tilde{h} - h_0\|_d < \epsilon$, $\|\bar{h} - h_0\|_d < \epsilon$ *we have:*

$$|a(h) - a(\tilde{h}) - D(h - \tilde{h}, \tilde{h})| \leq c_1(\|h - \tilde{h}\|_d)^2$$
$$|D(h,\bar{h}) - D(h,\tilde{h})| \leq c_2\|h\|_d\|\bar{h} - \tilde{h}\|_d$$

Assumption 9. *Suppose:*
i) there is a positve constant c *such that* $|D(h,h_0)| \leq c\|h\|_d$
ii) there is an $h_n \in \Theta_{ZW,n}$ *such that* $\mathbb{E}[h_n(X,Z,W)^2] \to 0$

Assumption 1 requires three regularity conditions. Condition i) is a lower bound on the conditional variance of the outcome variable, requiring it to be nontrivial. Condition ii) requires a bound on the minimal eigenvalue of the block-diagonal matrix Q. It avoids the limiting case of near-multicollinearity whose practical considerations were described in subsection 2.3.2. Finally condition iii) specifies the rate of convergence of the approximation of the sieve spaces $\Theta_{S,n}$ as anticipated when introducing the Stone-Weierstrass theorem. This is the order of the bias incurred because, for any finite n, the target m_0, S may not be in $\Theta_{S,n}$. Assumption 2 concerns the type of functionals of sieve estimators that we

can accommodate. The condition requires that the functional is linear or it has a linear approximation function that is Lipschitz in both arguments with respect to the strong norm. Finally Assumption 3 requires continuity of the linear approximation in the strong norm.

Theorem 46 (Newey 1997)**.** *Suppose the assumptions I and 1-3 hold. For a set of variables* $S = \{Z\}$ *or* $\{ZW\}$*, let* $\hat{h}_{S,n}$ *be the sieve estimator of* $h_{0,S}$ *based on the linear sieve* $\Theta_{S,n}$*, among one of the classes considered above. If* $\lim_{n\to\infty} \sqrt{n} T_{S,n}^{-\alpha_S} \to 0$*, then:*

$$\sqrt{\frac{n}{V_{S,n}}} \left(a(\hat{h}_{S,n}) - a(h_{S,0}) \right) \xrightarrow{d} \mathcal{N}(0,1)$$

where $V_{S,n}$ *is the entry of* V_n *corresponding to* $S = \{Z\}$ *or* $\{ZW\}$*.*

Theorem 47 (Joint Asymptotic Normality)**.** *Assume* $\lim_{n\to\infty} \frac{T_n^{ZW} \zeta_d(T_n^{ZW})^2}{n} = 0$ *and* $\lim_{n\to\infty} \frac{T_n^{Z} \zeta_d(T_n^{Z})^2}{n} = 0$*. Further let Assumptions 1 hold and the model is both* Z *and* ZW*-identifiable. We have:*

$$\sqrt{n} V_n^{-\frac{1}{2}} \begin{pmatrix} \widehat{AMTE}_Z(x_0) - AMTE(x_0) \\ \widehat{AMTE}_{ZW}(x_0) - AMTE(x_0) \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, I)$$

Here $V_n^{-\frac{1}{2}}$ is the unique (by positive definiteness) matrix M such that $M^2 = V_n$. Theorem 47 says that, under the baseline model, the plug-in estimators of $AMTE(x_0)$ using either the control set Z or the control set (Z,W) will be jointly normal and, in particular, both centered around the true $AMTE(x_0)$. The proof of Theorem 47 will amount to verify conditions on the theorem by Newey [1997], which I discuss below, and to apply the Cramér-Wold device.

Proof. The terms T_n^{ZW} and T_n^{Z} satisfy the requirement of Assumption 2 i). Proposition 43 guarantees that linearity of the functionals of interest, Γ_{ZW} and Γ_Z holds, satisfying

the second part of Assumption 2 i). Similarly, Assumption 3 i) requires continuity with respect to the strong norm, again guaranteed by Proposition 43. Here we may rewrite the functional Γ_{ZW} in a more convenient format:

$$\begin{aligned}
a(\hat{m}_n) &= a\left(\sum_{j=1}^{J_n}\sum_{k=1}^{K_n}\sum_{l=1}^{L_n}\phi_j(x)\psi_k(z)\varphi_l(w)\hat{\beta}_{j,k,l}\right) \\
&= \int\left(\sum_{j=1}^{J_n}\sum_{k=1}^{K_n}\sum_{l=1}^{L_n}\phi_j(x)\psi_k(z)\varphi_l(w)\hat{\beta}_{j,k,l}\right)dF_{ZW} \\
&= \sum_{j=1}^{J_n}\phi_j'(x_0)\left(\int\sum_{k=1}^{K_n}\sum_{l=1}^{L_n}\psi_k(z)\varphi_l(w)dF_{ZW}\right)\hat{\beta}_{jkl}^{ZW} \\
&= \sum_{j=1}^{J_n}\sum_{k=1}^{K_n}\sum_{l=1}^{L_n}\phi_j'(x_0)\left(\int\psi_k(z)\varphi_l(w)dF_{ZW}\right)\hat{\beta}_{jkl}^{ZW} \\
&= \sum_{j=1}^{J_n}\sum_{k=1}^{K_n}\sum_{l=1}^{L_n}A_{jkl}\hat{\beta}^{ZW} \\
&= A'\hat{\beta}_{jkl}^{ZW}
\end{aligned}$$

A similar rewriting can be applied to the functional Γ_Z to obtain:

$$\begin{aligned}
A_{jkl}^{ZW} &:= \phi_j'(x_0)\left(\int\psi_k(z)\varphi_l(w)dF_{ZW}\right) \\
A_{jk}^{Z} &:= \phi_j'(x_0)\left(\int\psi_k(z)dF_{ZW}\right)
\end{aligned}$$

where $A_{jkl} = \phi_j'(x_0)\left(\int\psi_k(z)\varphi_l(w)dF_{ZW}\right)$ and the sum is written in terms of a simple dot product. Incidentally, because it is a linear functional, the A matrix does not depend on the particular value of the pseudo-true value[11]. Theorem 46 □

The weighted formula above can be further simplified so that A_{jkl}^{ZW}, the weight

[11] Essentially, the A^{ZW} and A^{Z} matrices depend on the choice of sieve basis functions chosen but does not depend on the pseudo-true value.

corresponding to the $\hat{\beta}_{jkl}$ term, has the format:

$$\phi_j'(x_0)\mathbb{E}[\psi_k(Z)\varphi_l(W)]$$

i.e. the weights are given by the expectation of the basis functions of the control variables multiplied by the derivative of the j^{th} basis function for the policy variable X evaluated at the point of interest x_0. Notice how the weights depend on the pre-specified point x_0.[12]

It is possible to show that the convergence rate of the $AMTE(x_0)$ can be derived the Euclidean norm of A which coincides with the norm of the linear operator. Because $V_{Z,n} \propto \|A^Z\|_E^2$ and $V_{ZW,n} \propto \|A^{ZW}\|_E^2$ we are interested in characterizing the order of $\|A\|_E^2$. We have:

$$\|A_n\|_E^2 := \sqrt{\left(\sum_{j=1}^{J_n}\sum_{k=1}^{K_n}\sum_{l=1}^{L_n}\left(\phi_j'(x_0)\mathbb{E}[\phi_k(Z)\varphi_l(W)]\right)^2\right)} \tag{2.23}$$

Following De Jong [2002] one can show that $\|A_n^Z\|_E^2 = O((T_n^Z)^3)$, $\|A_n^{ZW}\|_E^2 = O((T_n^{ZW})^3)$. Hence, the rate of convergence for $\Gamma_S(\hat{m})$ to $\Gamma_S(m_0)$ in the intrinsic norm is given by:

$$O_p\left(\frac{(T_n^S)^3}{\sqrt{n}} + (T_n^S)^{-\alpha_S}\right) \tag{2.24}$$

This rate of convergence is slower than the parametric rate. If one views this result in light of the parametric test proposed by Lu and White [2014], an interesting trade-off arises. On the one hand the nonparametric test has a slower convergence rate than the parametric test. On the other hand, the functional form imposed by the parametric test may be severely mis-specified. Hence, it is not clear in which situations the naive test

[12]The estimator $\hat{A}_{jkl}$ is $\sqrt{n}$-consistent under the weak condition that $\mathbb{E}[\phi_k(Z)\varphi_l(W)] < \infty \ \forall \ l,k$. In particular choosing r-splines as a basis amounts to bounded $(r-1)$ cross moments, a relatively easy condition to satisfy. In this paper all functions have bounded support so the existence of these moments is trivially satisfied.

performs better than its nonparametric counterpart. The quantities described in Equations (10)-(20) here are population objects. For the implementation, the sample analogs for the quantities A, Q and Σ and V are a natural choice of consistent estimators. Taking the nonparametric regression with both Z and W as an example they can be written as:

$$\hat{Q}_n^{ZW} := \frac{1}{n}\Big(\Phi(X,Z,W)\Phi(X,Z,W)'\Big) \tag{2.25}$$

$$\hat{\Sigma}^{ZW} := \frac{1}{n}\Big(\Phi(X,Z,W)\Phi(X,Z,W)'\widehat{Var}(Y|X,Z,W)\Big) \tag{2.26}$$

$$\hat{A}_{jkl} := \phi_j'(x_0)\frac{1}{n}\sum_{i=1}^{n}\psi_k(Z_i)\varphi_l(W_i) \tag{2.27}$$

$$\hat{V}_n := \hat{A}'(\hat{Q}^{ZW})^{-1}\hat{\Sigma}^{ZW}(\hat{Q}^{ZW})^{-1}\hat{A} \tag{2.28}$$

With the result of Theorem 47, we can state Corollary 48 which is based on testing the difference between the two estimators we have obtained above for the $AMTE(x_0)$ using the full control set $\{Z,W\}$ or the restricted set $\{Z\}$.

Corollary 48 (Robustness Test)**.** *Assume* $\lim_{n\to\infty}\frac{T_n^{ZW}(\zeta_d(T_n^{ZW})^2}{n}=0$, $\lim_{n\to\infty}\frac{T_n^{Z}(\zeta_d(T_n^{Z})^2}{n}=0$*, and Assumption 6. Further assume* H_0 *is true. For* $R=\begin{bmatrix}+1 & -1\end{bmatrix}$*, we have the test statistic:*

$$\hat{R}_n := n\cdot\begin{pmatrix}\widehat{AMTE}_Z(x_0)\\ \widehat{AMTE}_{ZW}(x_0)\end{pmatrix}' R'(R\hat{V}_nR')^{-1}R\begin{pmatrix}\widehat{AMTE}_Z(x_0)\\ \widehat{AMTE}_{ZW}(x_0)\end{pmatrix}\xrightarrow{d}\chi^2(1)$$

We have recovered a nonparametric test that extends the linear Lu and White [2014] robustness approach. Because this test does not rely on linearity of the conditional mean function but instead flexibly models it through the sieve approach, we can expect rejections from this test to come exclusively from a violation of the causal assumptions that allow identification of the $AMTE(x_0)$. Which information carried by the causal structure are being tested by the estimator in Corollary 48? In the spirit of causal diagrams one could

look at the robustness test as a validation exercise of model M_B against model M_R. i.e. the test is testing for the presence of the red link in the picture below.

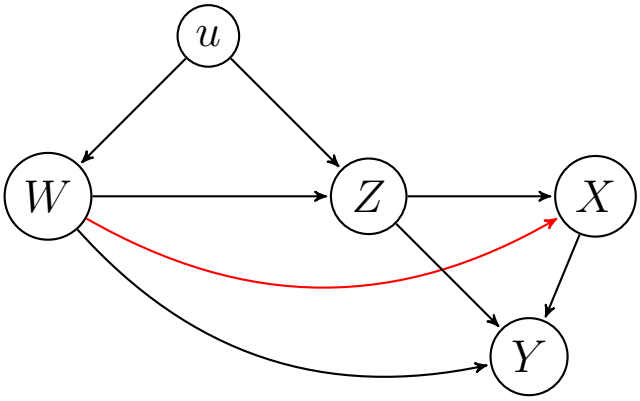


Figure 2.4.1: Falsification of M_B against M_R. If the red arrow is present, consistent estimation of $AMTE(x_0)$ requires estimating $\mathbb{E}[Y|X,Z,W]$ since $\mathbb{E}[Y|X,Z]$ has no causal interpretation.

2.5 Conclusion

In this paper I provide a theoretical framework to extend the notion of robustness test to a nonparametric setting. The procedure preserves the simple heuristic of coefficient stability: look at the result of two (or more) regression coefficients, if they are too far apart, reject your identification strategy. In contrast to available alternatives in the literature, which require restrictive assumptions like linearity, my procedure allows flexible functional forms estimation. This framework has a number of desirable properties. The proposed test is transparent: if identification is correct, the two available ways of estimating the causal object of interest must agree; if they don't, one of the models must be wrong. In the same spirit of Lu and White [2014], the proposed test upgrades the robustness check to a robustness test and characterizes its asymptotic distribution allowing for a rigorous statistical procedure. The focus on nonparametric identification circumvents the

problems of identification through functional form. Because the approach uses only the conditional independence assumptions from the causal diagram[13] the proposed test does not suffer from the generic rejection problem of its parametric counterpart. The estimation through the sieve approach accommodates a large degree of flexibility without an excessive computational burden.

There are several promising directions for future research. First, it would be important to characterize the power of the nonparametric test *versus* the parametric available versions. For certain data generating processes and certain points of evaluation, the trade off between simplicity and accuracy may favor the naive OLS test. Second, the heuristic of coefficient stability can be further extended to accommodate many other nonparametric methods though the characterization of the asymptotic distribution of the two or more estimators may be challenging in this general case.

2.6 Acknowledgements

Chapter 2, in full, is currently being prepared for submission for publication of the material. The dissertation author was the sole author of this material.

[13]The sieve estimator still restricts the class of functions to being differentiable and sufficiently smooth functions, though this class is still a nonparametric class, much richer than any particular finite dimensional parametric class of functions.

Appendix

2.A Other definitions

Definition 49. *Holder class*

Let $p \in \mathbb{R}_+$. Then write $p = \alpha + [p]$ where $[p]$ denotes the largest integer smaller than p, with $\alpha \in [0,1)$. Let $\mathcal{X}$ and $\mathcal{Y}$ be Banach spaces and $m : \mathcal{H} \to \mathcal{Y}$ be a function between them. Then m is said to be p-smooth if it is $[p]$-times continuously differentiable and for $x, x' \in \mathcal{H}$ we have:

$$\sup_{x \neq x' \in \mathcal{X}} \frac{\|m^{[p]}(x) - m^{[p]}(x')\|}{\|x - x'\|^{\alpha}} < \infty$$

Definition 50. *Holder Ball*

A Holder ball with smoothness p is defined as:

$$\Lambda_c^p(\mathcal{X}) := \left\{ m \in C^{[p]}(\mathcal{X}) : \frac{|m^{[p]}(x) - m^{[p]}(x')|}{|x - x'|^{\alpha}} \leq c;\ \|m\|_d \leq c \right\}$$

Essentially, a Holder ball is a class of functions whose derivatives are Holder continuous with exponent α. A p-smooth functions can be approximated by some basis functions with some well behaved approximation error.

Definition 51. *Tensor Product Hilbert spaces*

Let H_1 and H_2 be two Hilbert spaces with bases $\{\phi_k\}$ and $\{\psi_l\}$ respectively. The tensor product $H_1 \otimes H_2$ is a Hilbert space $\mathcal{P}$ together with a bi-linear mapping $b : (H_1, H_2) \to \mathcal{P}$

such that:

- *the closed linear span of* $b(\phi, \psi) = \mathcal{P}$
- $\langle b(\phi_1, \psi_1), b(\phi_2, \psi_2) \rangle_{\mathcal{P}} = \langle \phi_1, \phi_2 \rangle_{H_1} \cdot \langle \psi_1, \psi_2 \rangle_{H_2}$

Then $\{\phi_k \otimes \psi_l\}$ *is an orthonormal basis for the tensor product space* $H_1 \otimes H_2$.

2.B Extensions

In this section I discuss three extension of the main framework of the paper to incorporate three cases of interest beyond the standard control variable approach. The first extension discusses robustness test for an a conditionally valid instrumental variable estimator. The procedure is based on the control function approach and follows the same heuristic underlying Theorem 38. The second extension draws a connection between the type of robustness tests based on coeffcient stability and a Sargan type falsification test in the presence of multiple instruments. The third extension discusses a robustness test in the context of mediation analysis. From the perspective of a causal diagram, mediation analysis carries similar types of causal restrictions as an instrumental variable model and therefore is suitable for the same type of analysis. While not very popular among economists, mediation analysis has recently gained some interest in Imbens [2019] and Bellamare and Bloem [2019]. The extension of a robustness test in this context is straightforward.

2.B.1 Conditionally valid instrumental variables

In the main body of the paper we considered the testable implication of the coefficient stability exercise arising from a "controlling for observables" procedure. In this subsection I consider the case of instrumental variables estimation. Consider the following familiar situation in applied research: a researcher proposes identification of the effect of X on Y through an available instrument Z. The baseline identification structure hinges on,

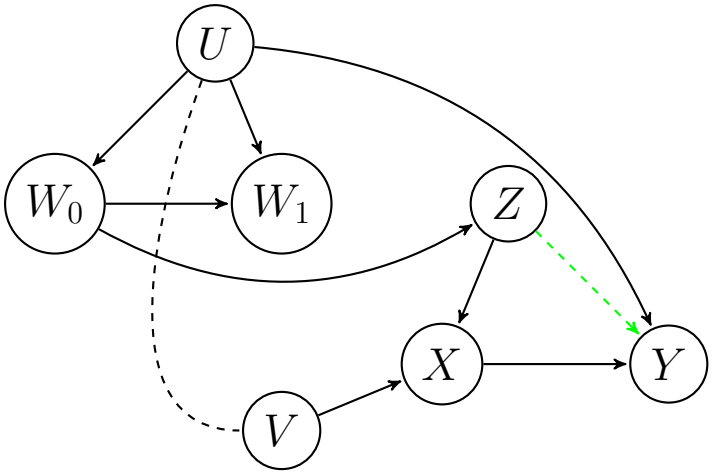


Figure 2.B.1: $AMTE(x_0)$ is not identified by including W_0, W_1 in the control set because of the path $Y \leftarrow U \rightarrow X$. On the other hand, variable Z is a potential instrument, if appropriate control variables are included. Z is a valid instrument if either W_0 or both W_0 and W_1 are included in the control set. The dashed green arrow represents the (reduced form) regression of Y on Z. The lack of solid arrows from Z to Y amounts to the exclusion restriction.

$Z \perp\!\!\!\perp U|W_0$. This is not a testable assumption, and is usually justified by domain knowledge that may reflect the institutional background of the observable variables. Suppose there is a (set of) additional controls W_1 available and that W_1 does not contain bad controls (that is, $Z \perp\!\!\!\perp U|W_0 \implies Z \perp\!\!\!\perp U|W_0, W_1$). That is, Z is a valid instrumental variable given both the control set W_0 as well as the larger control set $\{W_0, W_1\}$ like depicted in Figure 2.B.1. In this case, there is an opportunity for a coefficient stability robustness exercise to convince a reader that identification based on $Z \perp\!\!\!\perp U|W_0$ is realistic.

Note that, in the model considered in this section, the $AMTE(x_0)$ generally fails to be nonparametrically identified without additional restrictions, as discussed in Pearl [2000]. As such, I impose some sufficient conditions on the structural functions to ensure identifiability and postpone the discussion of the fully general model to a later section of

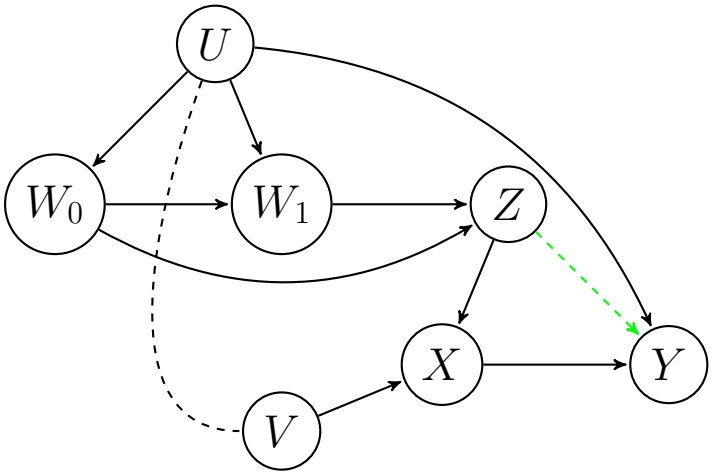


Figure 2.B.2: Here, a link between U and Z is present through two variables, W_0 and W_1. Notice again $AMTE(x_0)$ is not identified by including W_0, W_1 in the control set because of the path $Y \leftarrow U\text{---}V \rightarrow X$. On the other hand, variable Z is a potential instrument, if appropriate control variables are included. Z is a valid instrumental variable only if W_0 and W_1 are included in the control set. Simply including W_0 and using an IV approach would result almost surely in a bias. The dashed green arrow represents the (reduced form) regression of Y on Z. Notice the lack of solid arrows from Z to Y amounts to the exclusion restriction.

this Appendix. Namely, here we require separability in the structural equation for X:

$$Y = C(x, u)$$
$$X = F(z) + v$$

Here Z is a conditionally valid instrument provided that $Z \perp\!\!\!\perp U|W_0$ in the baseline model. The robustness test in this context will gauge at whether $Z \perp\!\!\!\perp U|W_0$ and $Z \perp\!\!\!\perp U|W_0, W_1$ jointly hold by comparing the two estimates of the $AMTE(x_0)$ that one can get through the control function approach. The structural function is additively separable in a function of the observable Z and unobservable v. Then one can identify the $AMTE(x_0)$ using a control function approach that follows similar steps to the main procedure discussed in Section 2.3

- Estimate $\mathbb{E}[X|Z, W_0, W_1]$ with the sieve approach
- Obtain estimates of $v = X - \mathbb{E}[X|Z, W_0, W_1]$
- Use $U \perp\!\!\!\perp X|v$ to estimate $\mathbb{E}[Y|X, v, W_0, W_1]$ using the sieve approach
- Recover the causal estimand of interest by taking: $\frac{\partial \mathbb{E}[Y|X=x, W_0, W_1, v]}{\partial x}$ and evaluate it at x_0 and integrating out the distribution of the conditional controls

One can immediately notice how the separability in X allows to tease out the residual v in the second step. The separability requirement can be slightly generalized to allow a X to be expressed as known function of $F(z)$ and v but it is hard to think about a case where this is relevant in practice.

Remark 52. *In practice, the diagram portrayed in Figure 2.B.1 and Figure 2.B.2 may be further enriched with additional covariates that are directly factored into the structural equation for Y. The control function approach may easily accommodate these changes by*

appropriately incorporating the additional control variables in the first stage, second stage, or both.

Proposition 53 (Identification)**.** *$AMTE(X_0)$ is identified by the equation:*

$$\int_{\mathcal{V}} \int_{\mathcal{W}_0 \times \mathcal{W}_1} \frac{\partial \mathbb{E}[Y|X = x, w_0, w_1, v]}{\partial x}(x_0, v) f_V(w_0, w_1, v) dw_0 dw_1 dv \tag{A2.29}$$

Proof. Observe that if V was observable, the following conditional independence holds: $u \perp\!\!\!\perp X | W_0, W_1, v$. Then, one may proceed to identify the $AMTE(x_0)$ with the control

function approach.

$$\int_U \frac{\partial C}{\partial x}(x_0,u) f_U(u) du \tag{A2.30}$$

$$= \int_U \int_V \int_{\mathcal{W}_0\times\mathcal{W}_1} \frac{\partial C}{\partial x}(x_0,u) f_U(u|v,w_0,w_1) f(v,w_0,w_1) dw_0 dw_1 dv du \tag{A2.31}$$

$$= \int_V \int_{\mathcal{W}_0\times\mathcal{W}_1} \left(\int_U \frac{\partial C}{\partial x}(x_0,u) f_U(u|v,w_0,w_1) du \right) f(v,w_0,w_1) dw_0 dw_1 dv \tag{A2.32}$$

$$= \int_V \int_{\mathcal{W}_0\times\mathcal{W}_1} \left(\int_U \lim_{\Delta x\to 0} \frac{Y_{x_0+\Delta x}(u) - Y_{x_0}(u)}{\Delta x} f_U(u|v,w_0,w_1) du \right) \tag{A2.33}$$

$$\times f(v,w_0,w_1) dw_0 dw_1 dv \tag{A2.34}$$

$$= \int_V \left(\lim_{\Delta x\to 0} \frac{\int_U Y_{x_0+\Delta x}(u) f_U(u|v,w_0,w_1) du - \int_U Y_{x_0}(u) f_U(u|v,w_0,w_1) du}{\Delta x} \right) \tag{A2.35}$$

$$\times f(v,w_0,w_1) dw_0 dw_1 dv \tag{A2.36}$$

$$= \int_V \int_{\mathcal{W}_0\times\mathcal{W}_1} \lim_{\Delta x\to 0} \frac{1}{\Delta x} \left(\int_U Y_{x_0+\Delta x}(u) f_U(u|x_0+\Delta x,w_0,w_1,v) du \right.$$

$$\left. - \int_U Y_{x_0}(u) f_U(u|x_0,w_0,w_1,v) du \right) f(v,w_0,w_1) dw_0 dw_1 dv \tag{A2.37}$$

$$= \int_V \int_{\mathcal{W}_0\times\mathcal{W}_1} \lim_{\Delta x\to 0} \frac{1}{\Delta x} \left(\mathbb{E}[Y_{x_0+\Delta x}|X,W_0,W_1,V](x_0+\Delta x,w_0,w_1,v) \right.$$

$$\left. - \mathbb{E}[Y_{x_0}|X,W_0,W_1,V](x_0,w_0,w_1,v) \right) f(v,w_0,w_1) dw_0 dw_1 dv \tag{A2.38}$$

$$= \int_V \int_{\mathcal{W}_0\times\mathcal{W}_1} \lim_{\Delta x\to 0} \frac{1}{\Delta x} \left(\mathbb{E}[Y|X,W_0,W_1,V](x_0+\Delta x,w_0,w_1,v) \right.$$

$$\left. - \mathbb{E}[Y|X,W_0,W_1,V](x_0,w_0,w_1,v) \right) f(v,w_0,w_1) dw_0 dw_1 dv \tag{A2.39}$$

$$= \int_V \int_{\mathcal{W}_0\times\mathcal{W}_1} \frac{\partial \mathbb{E}[Y|X,W_0,W_1,V]}{\partial x}(x_0,w_0,w_1,v) f(v,w_0,w_1) dw_0 dw_1 dv \tag{A2.40}$$

Equation (A2.31) follows from conditioning on v, (A2.32) follows from Fubini's theorem because $C \in \mathcal{C}^p(\mathcal{X}\times\mathcal{U})$. Equation (A2.34) is the definition of Y_x, Equation (A2.36) follows from the dominated convergence theorem. Equation (A2.37) follows from

the conditional independence of U and X given (W_0, W_1, v) and Equation (A2.38) from the definition of conditional expectation. Equation (A2.39) follows from the consistency of Y_x with observed outcome Y and Equation (A2.40) from the definition of derivative. Although v is not observable, it may be consistently estimated by a first stage nonparametric regression of X on Z, W_0, W_1.

$$\begin{aligned}
\mathbb{E}[X|Z,W_0,W_1] &= \mathbb{E}[F(Z)+v|Z,W_0,W_1] && \text{(A2.41)}\\
&= \mathbb{E}[F(Z)|Z,W_0,W_1]+\mathbb{E}[v|Z,W_0,W_1] && \text{(A2.42)}\\
&= F(Z)+\mathbb{E}[v|Z,W_0,W_1] && \text{(A2.43)}\\
&= F(Z) && \text{(A2.44)}
\end{aligned}$$

Equation (A2.41) follows from the additive separability of the structural equation, Equation (A2.42) from linearity of conditional expectations, Equation (A2.43) from $F(Z)$ being measurable with respect to $\sigma(Z, W_0, W_1)$. Finally an observation on $\mathbb{E}[v|Z, W_0, W_1] = 0$. Because X is not included in the conditioning set there is no bias from conditioning on a common outcome of Z and v. Further, because both the controls W_0 and W_1 are in the conditioning set, Z is a valid instrument and as such we have $Z \perp\!\!\!\perp \{u, v\}|W_0, W_1$. Hence one may recover $v = X - \mathbb{E}[X|Z, W_0, W_1]$ and use it to estimate $\mathbb{E}[Y|X, V]$. This finises the proof. □

Consider the conditions under which such an exercise would be revealing of identification failures. In a similar vein as discussed in the main paper, the proposed robustness test would reject if:

- $Z \perp\!\!\!\perp Y|W_0$ is false while $Z \perp\!\!\!\perp Y|W_0, W_1$ is true
- $Z \perp\!\!\!\perp Y|W_0$ is true while $Z \perp\!\!\!\perp Y|W_0, W_1$ is false

Like in the standard case, we must guarantee that rejections that depend on *ii*) are

ruled out. A sufficient condition, analogous to Equation (2.6) is that W_1 does not contain "bad controls" discussed in the introduction so that we maintain $Z \perp\!\!\!\perp Y|W_0 \implies Z \perp\!\!\!\perp Y|W_0, W_1$. There is a natural adaptation of Theorems 47 and 48 for the control function approach IV estimator. We leave the characterization of its asymptotic distribution for future work.

2.B.2 Non-additively separable models

We now explore the generalization of the above model where the structural function that disciplines the behavior of the dependent variable is not necessarily additively separable in Z. The system of structural equations is then given by

$$Y = C(x, u)$$
$$X = g(z, v)$$

One may look at the average change we can induce in Y through an exogenous change in Z and consider the reduced form relationship below:

$$\mathbb{E}\left[\frac{\partial C(X(z,v),u)}{\partial z}(x_0, U)\right] = \mathbb{E}\left[\frac{\partial C}{\partial x}(x_0, u) \cdot \frac{\partial g}{\partial z}(z, v)\right] \tag{A2.45}$$

Focusing on the left hand side:

$$
\begin{aligned}
&\int_{\mathcal{U}} \frac{\partial C(X(z,v),u)}{\partial z} f_U(u) \\
&= \int_{\mathcal{W}_0\times\mathcal{W}_1} \left(\int_{\mathcal{U}} \frac{\partial C(X(z,v),u)}{\partial z} f_U(u|w_0,w_1)du \right) f_W(w_0,w_1)dw_0dw_1 \\
&= \int_{\mathcal{W}_0\times\mathcal{W}_1} \left(\int_{\mathcal{U}} \lim_{\Delta z\to 0} \frac{Y(X_{z+\Delta z}(v),u) - Y(X_z(v),u)}{\Delta z} f_U(u|w_0,w_1)du \right) f_W(w_0,w_1)dw_0dw_1 \\
&= \int_{\mathcal{W}_0\times\mathcal{W}_1} \left(\int_{\mathcal{U}} \lim_{\Delta z\to 0} \frac{Y(X_{z+\Delta z}(v),u) - Y(X_z(v),u)}{\Delta z} f_U(u|z,w_0,w_1) \right) f_W(w_0,w_1)dw_0dw_1 \\
&= \int_{\mathcal{W}_1\times\mathcal{W}_2} \left(\lim_{\Delta z\to 0} \int_{\mathcal{U}} \frac{Y(X_{z+\Delta z}(v),u) - Y(X_z(v),u)}{\Delta z} f_U(u|z,w_0,w_1) \right) f_W(w_0,w_1)dw_0dw_1 \\
&= \int_{\mathcal{W}_0\times\mathcal{W}_1} \left(\lim_{\Delta z\to 0} \frac{\mathbb{E}[Y_{X_{z+\Delta z}} - Y_{X_z}|Z=z,W_0=w_0,W_1=w_1]}{\Delta z} \right) f_W(w_0,w_1)dw_0dw_1 \\
&= \int_{\mathcal{W}_0\times\mathcal{W}_1} \frac{\partial \mathbb{E}[Y_{X_z}|Z=z,W_0=w_0,W_1=w_1]}{\partial z} f_W(w_0,w_1)dw_0dw_1 \\
&= \int_{\mathcal{W}_0\times\mathcal{W}_1} \frac{\partial \mathbb{E}[Y|Z=z,W_0=w_0,W_1=w_1]}{\partial z} f_W(w_0,w_1)dw_0dw_1
\end{aligned}
$$

So the LHS is identified by the derivative of the conditional mean function of Y given instrument Z and control variables W_0, W_1[14]. The right hand side can be rearranged as:

$$
\begin{aligned}
&\int_{\mathcal{U}\times\mathcal{V}} \frac{\partial C}{\partial x}(x_0,u)\cdot\frac{\partial g}{\partial z}(z,v) f_{UV}(u,v)dudv \\
&= \int_{\mathcal{W}_0\times\mathcal{W}_1} \int_{\mathcal{U}\times\mathcal{V}} \left(\frac{\partial C}{\partial x}(x_0,u)\cdot\frac{\partial g}{\partial z}(z,v)\cdot f_{UV}(u,v|w_0,w_1)dudv \right) f_W(w_0,w_1)dw_0dw_1 \\
&= \int_{\mathcal{W}_0\times\mathcal{W}_1} \int_{\mathcal{U}\times\mathcal{V}} \left(\frac{\partial C}{\partial x}(x_0,u)\cdot\frac{\partial g}{\partial z}(z,v)\cdot f_{UV}(u,v|z,w_0,w_1)dudv \right) f_W(w_0,w_1)dw_0dw_1 \\
&= \int_{\mathcal{W}_0\times\mathcal{W}_1} \int_{\mathcal{U}\times\mathcal{V}} \left(\frac{\partial C}{\partial x}(x_0,u)\cdot \lim_{\Delta Z\to 0} \frac{X_{z+\Delta z}(v) - X_z(v)}{\Delta z} f_{UV}(u,v|z,w_0,w_1)dudv \right) f_W(w_0,w_1)dw_0dw_1
\end{aligned}
$$

The LHS is the nonparametric regression of Y on Z, W_0 and W_1. In a fully nonparametric, possibly non-separable model, it is hard to isolate a reduced form object in

[14]The reader will notice that this reduced form is obtained precisely as the estimator in section 2, with the exception that this statistical object does not have any causal interpretation *per-se*.

the second term. RHS is the marginal effect of increasing x by one unit in policy environment u, $\frac{\partial C}{\partial x}(x_0, u)$, weighted by the derivative of the conditional mean of X given the instrument Z and the controls W_0, W_1 for policy environment $U = u$. Ultimately, the presence of the v in the second term makes it hard to isolate a reduced form object on the right hand side. Clearly, since the two objects $\frac{\partial \mathbb{E}[Y|Z=z,W_0=w_0,W_1=w_1]}{\partial z}$ and $\frac{\mathbb{E}[X|Z=z,W_0=w_0,W_1=w_1]}{\partial z}$ depend uniquely on observed variables, they are identified. In general, the solution to (A2.45) is not unique. A sufficient condition for uniqueness is completeness of the conditional expectation of X given Z and the controls W_0, W_1.

2.B.3 A connection to over-identification tests in instrumental variables models

Perhaps not surprisingly, robustness tests based on coefficient stability can be related to the well-known Sargan-Hansen test for over-identifying restrictions in the context of instrumental variables regression. This section highlights the connection and shows that both can be seen as particular cases of falsification tests. The procedure in a standard over-identified restriction test still hinges on the general heuristic of coefficient stability discussed in this paper: if the identification structure is correct, there are two equivalent ways of identifying the causal effect of interest. Conversely, if the two procedures lead to different estimates, then the proposed identification may not hold. The example below shows that the very well studied Sargan-Hansen test may be given a coefficient stability interpretation. To keep things simple, I focus on linear instrumental variable models with just one regressor and two instruments: there is a single over-identifying restriction. The procedure can be generalized to multiple over-identified restrictions and to the context of GMM estimators. Consider the following causal diagram:

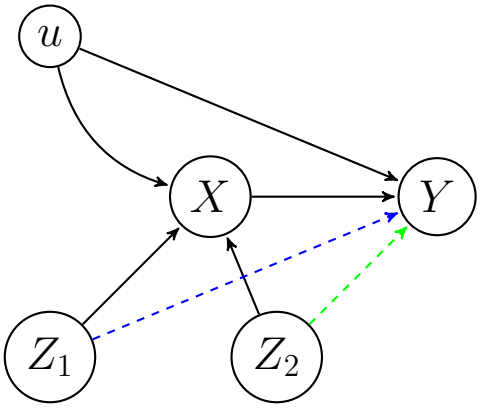


Figure 2.B.3: Over-identified instrumental variables system

To highlight the connection between robustness test and over-identification tests we can further assume that the structural functions are linear. The dashed lines represent the relationship one would obtain if they regressed Y on Z_1 and Z_2 respectively. As such, the lines purely represent reduced forms and the absence of the solid lines pointing to Y from either of the instruments Z represents the exclusion restriction. If the hypothesized causal structure is correctly specified one has two possibilities to estimate the $AMTE(x_0)$.

- Obtain the 2SLS regression of Y on X using Z_2
- Obtain the 2SLS regression of Y on X using Z_1

Under the hypothesized model both regressions lead to the same estimated causal effect. If the two estimates differ, the robustness test falsifies the model. In this light, the Sargan-Hansen J test for over-identifying restrictions reflects exactly the same heuristic of coefficient stability described in the main paper.

Proposition 54 (Coefficient Stability and over-identifying restrictions)**.** *Let* (γ_1, γ_2) *be the regression coefficient from the reduced form regression of* Y *on* $(Z_1, Z_2)'$ *and* (α_1, α_2) *be the regression of* X *on* $(Z_1, Z_2)'$*. The Sargan J test is testing:*

$$H_0 : \frac{\gamma_1}{\alpha_1} = \frac{\gamma_2}{\alpha_2}$$

Proof. Denoting $\hat{u}$ as the vector of residuals from the instrumental variable regression of Y on X using instruments $\{Z_1, Z_2\}$, in case of conditional homoskedasticity, the standard J statistic may be written as:

$$\begin{aligned}
J &= \frac{1}{\hat{\sigma}^2} \cdot \hat{u}' Z (Z'Z)^{-1} Z' \hat{u} \\
&= \frac{1}{\hat{\sigma}^2} \cdot (y - X(X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y)' Z (Z'Z)^{-1} Z' \\
&\times Z(Z'Z)^{-1}Z'(y - X(X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y) \\
&= y'Z(Z'Z)^{-1}Z' - y'Z(Z'Z)^{-1}Z'X(X'Z(Z'Z)^{-1}Z'X)^{-1}X' \\
&\times (Z'Z)^{-1}Z'y - (Z'Z)^{-1}Z'X(X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y \\
&= \left(\begin{bmatrix} \hat{\gamma_1} \\ \hat{\gamma_2} \end{bmatrix} - \begin{bmatrix} \hat{\alpha_1} \\ \hat{\alpha_2} \end{bmatrix} \hat{\beta}_{2SLS} \right)' \frac{Z'Z}{\hat{\sigma}^2} \left(\begin{bmatrix} \hat{\gamma_1} \\ \hat{\gamma_2} \end{bmatrix} - \begin{bmatrix} \hat{\alpha_1} \\ \hat{\alpha_2} \end{bmatrix} \hat{\beta}_{2SLS} \right)
\end{aligned}$$

which is a quadratic form in the coefficients of interest, testing the restriction:

$$\beta = \frac{\gamma_1}{\alpha_1} = \frac{\gamma_2}{\alpha_2}$$

which is the coefficient stability restriction. □

2.B.4 Optimal Robustness Test Selection

Lu and White [2014] implement a Feasible Optimally combined GLS estimator. In their context, a menu of valid robustness check regressions is available. Using the notation of the main body of this paper I denote the additional (sets of) covariates as $W_1, W_2, W_3, \cdots, W_J$. The natural robustness test takes advantage of the GLS structure and essentially carries out all robustness regression comparisons simultaneously. While theoretically possible, carrying out such a procedure in a nonparametric context may be computationally unfeasible.

2.C Some additional results

2.C.1 Binary Treatment

The definition of $AMTE(x_0)$ above is only meaningful when $C(\cdot, u)$ is a differentiable function for almost every $x \in \mathcal{X}$. In several applications the interest lies in the effect of a binary treatment. In such cases we need a different definition of $AMTE$.

Definition 55. *Let $\mathcal{X} = \{0,1\}$. Then we define:*

$$ATE := \int C(1,u) - C(0,u) f_U(u) du \tag{A2.46}$$

Then there is an analogous result to proposition 3 for the continuous case above.

Proposition 56. *i) Let $U \perp\!\!\!\perp X|Z,W$.*
Then $ATE_B = \int (\mathbb{E}[Y|X=1,Z,W] - \mathbb{E}[Y|X=0,Z,W]) f_{ZW}(z,w) dzdw$.
ii) Let instead $U \perp\!\!\!\perp X|Z$. Then $ATE = \int (\mathbb{E}[Y|X=1,Z] - \mathbb{E}[Y|X=0,Z]) f_Z(z) dz$

Proof.

$$\begin{aligned}
&ATE^B_{ZW} && \text{(A2.47)}\\
&:= \int \left(C(1,u) - C(0,u)\right) f_u(u) du && \text{(A2.48)}\\
&= \int \int_{\mathcal{Z}\times\mathcal{W}} \left(C(1,u|z,w) - C(0,u|z,w) f_U(u|z,w) f_{ZW}(z,w) dzdw\right) du && \text{(A2.49)}\\
&= \int_{\mathcal{Z}\times\mathcal{W}} \int \left(C(1,u|z,w) - C(0,u|z,w) f_U(u|z,w) du\right) f_{ZW}(z,w) dzdw && \text{(A2.50)}\\
&= \int_{\mathcal{Z}\times\mathcal{W}} \int \left(C(1,u|x,z,w) - C(0,u|x,z,w) f_U(u|x,z,w) du\right) f_{ZW}(z,w) dzdw && \text{(A2.51)}\\
&= \int_{\mathcal{Z}\times\mathcal{W}} \left(\mathbb{E}[Y|X=1,z,w)] - \mathbb{E}[Y|X=0,z,w)]\right) f_{ZW}(z,w) dzdw && \text{(A2.52)}
\end{aligned}$$

which is the desired result. The proof for *ii*) is identical and omitted for brevity. □

Similar to the continuous treatment case we have the immediate corollary.

Corollary 57. *$U \perp\!\!\!\perp X|Z,W$ as well as $U \perp\!\!\!\perp X|Z$. Then, $ATE_{ZW} = ATE_Z$.*

2.C.2 Asymptotic bias

Suppose now that the baseline model is causally mis-specified so that omitting W does not allow identification of the causal effect of interest. How does omitting the control variable W impact the asymptotic bias in the $AMTE(x_0)$? For simplicity here assume $\mathcal{U} = supp(f(u|x,z) = supp(f(u|x',z))$ for any $x, x' \in \mathcal{X}$. This says that, while the conditional distribution of U given Z, X may depend on the choice of x, the support of the density does not vary with the choice of x.

Proposition 58 (Asymptotic Bias)**.** *The asymptotic bias resulting from the exclusion of W is characterized by the expression below.*

$$AsyBias = \int_{\mathcal{Z}} \int_{\mathcal{W}} \frac{\partial f_{W|X,Z}}{\partial x}(w|x_0,z) \int_U C(x_0,u) f_U(u|w,z) du dw f_Z(z) dz \tag{A2.53}$$

Proof. See Appendix 2.D. □

2.C.3 What does a bad control estimate?

In the main paper we required the additional controls to be valid, in the sense that if the model is Z-identifiable then it is W-identifiable. This is required in order to rule out the possibility of introducing a bad control in a model that would otherwise be correctly identifying the $AMTE(x_0)$.

Proposition 59 (Bad Control)**.** *Let the model be Z-identifiable but not ZW-identifiable. The bias in the population for introducing a "bad" control variable W is given by the formula*

below:

$$\widetilde{AMTE}(x_0)_{bad} \tag{A2.54}$$

$$= \int_{\mathcal{W}} \int_{\mathcal{Z}} \left(\frac{\partial \mathbb{E}\left[Y|X=x,Z,W_{bad}\right]}{\partial x}(x_0,z) \right) f_{Z|W}(z|w) f_W(w) dz dw \tag{A2.55}$$

$$= \int_{\mathcal{W}} \int_{\mathcal{Z}} \left(\lim_{\Delta x \to 0} \frac{1}{\Delta x} \left(\mathbb{E}\left[Y|X=x,Z,W_{bad}\right](x_0+\Delta x,z,w) \right.\right.$$
$$\left.\left. - \mathbb{E}\left[Y|X=x,Z,W_{bad}\right](x_0,z,w) \right) \right) f_{Z|W}(z|w) f_W(w) dz dw \tag{A2.56}$$

$$= \int_{\mathcal{W}} \int_{\mathcal{Z}} \left(\lim_{\Delta x \to 0} \frac{1}{\Delta x} \left(\mathbb{E}\left[Y_{x_0+\Delta x}|X=x,Z,W_{bad}\right](x_0+\Delta x,z,w) \right.\right.$$
$$\left.\left. - \mathbb{E}\left[Y_{x_0}|X=x,Z,W_{bad}\right](x_0,z,w) \right) \right) f_{Z|W}(z|w) f_W(w) dz dw \tag{A2.57}$$

$$= \int_{\mathcal{W}} \int_{\mathcal{Z}} \left(\lim_{\Delta x \to 0} \frac{1}{\Delta x} \left(\int_{\mathcal{U}} C(x_0+\Delta,u) f_U(u|x_0+\Delta x,z,w_{bad}) \right.\right.$$
$$\left.\left. - \int_{\mathcal{U}} C(x_0,u) f_U(u|x_0,z,w_{bad}) \right) \right) f_{Z|W}(z|w) f_W(w) dz dw \tag{A2.58}$$

$$= \int_{\mathcal{W}} \int_{\mathcal{Z}} \left(\left(\int_{\mathcal{U}} \lim_{\Delta x \to 0} \frac{1}{\Delta x} (C(x_0+\Delta,u) f_U(u|x_0+\Delta x,z,w_{bad}) \right.\right.$$
$$\left.\left. - C(x_0,u)) f_U(u|x_0,z,w_{bad}) \right) \right) f_{Z|W}(z|w) f_W(w) dz dw \tag{A2.59}$$

$$= \int_{\mathcal{W}} \int_{\mathcal{Z}} \left(\left(\int_{\mathcal{U}} \frac{\partial C}{\partial x}(x_0,u) f_U(u|x_0,z,w_{bad}) \right.\right.$$
$$\left.\left. + \int_{\mathcal{U}} C(x_0,u) \frac{\partial f_{U|x,w,z}}{\partial x}(u|x_0,z,w_{bad}) \right) \right) f_{Z|W}(z|w) f_W(w) dz dw \tag{A2.60}$$

Together with proposition 58, 59 says that failures of identification may arise either from failing to control for variables that are needed for identification, or from controlling for variables that introduce endogeneity that breaks identification. In this sense, any robustness test that one may design should use variables that are in neither category. The testable restriction given by coefficient stability presumes that the researcher has hypothesised a

minimally identifying set and that the additional variables are neither needed nor harmful for identification.

2.C.4 What would OLS be estimating

The previous section has discussed the nonparametric estimation strategy for the $AMTE(x_o)$. What would happen if one simply considers the naive OLS estimator and attempts to use it to conduct the robustness exercise for causal identification. First we immediately note that the design point, x_0 bears no relevance to the OLS estimator since it is never required as an input in the estimation process. As such, OLS necessarily estimates some aggregate effect over all feasible estimation points. One may then hope that $\hat{\beta}_{OLS}$ can still be given an interpretation as a pseudo-true value, i.e some particular average of $AMTE(x_0)$ with weights coming from the empirical distribution of X. I show below that this is not the case.

Proposition 60 (Representation)**.** *The estimator β_{ZW} could be represented as:*

$$\beta_{ZW} = \frac{\int_{\mathcal{X}} \int_{\mathcal{Z}\times\mathcal{W}} m'(t,z,w)\rho(z,w,t)dF_{Z,W}dt}{\int_{\mathcal{X}} \int_{\mathcal{Z}\times\mathcal{W}} \rho(z,w,t)dF_{Z,W}dt}$$

$$\rho(z,w,t) := \left(\mathbb{E}[|X \geq t|Z,W] - \mathbb{E}[|X < t|Z,W]\right)\left(\mathbb{P}(X \geq t|Z,W])(1 - \mathbb{P}(X \geq t|Z,W])\right)$$

Proof. Following Angrist and Pischke [2008] we may recover:

$$\beta_{ZW} = \frac{\mathbb{E}[Y(X - \mathbb{E}[X|Z,W])]}{\mathbb{E}[X(X - \mathbb{E}[X|Z,W])]}$$

Looking at the numerator one has:

$$\mathbb{E}[Y(X-\mathbb{E}[X|Z,W])] \tag{A2.61}$$

$$=\mathbb{E}[\mathbb{E}[Y(X-\mathbb{E}[X|Z,W])]|X,Z,W] \tag{A2.62}$$

$$=\mathbb{E}[\mathbb{E}[Y|X,Z,W](X-\mathbb{E}[X|Z,W])] \tag{A2.63}$$

$$=\int_{\mathcal{Z}\times\mathcal{X}\times\mathcal{W}}\mathbb{E}[Y|X,Z,W](x-\mathbb{E}[X|Z=z,W=w])]dF_{X,Z,W} \tag{A2.64}$$

$$=\int_{\mathcal{Z}\times\mathcal{X}\times\mathcal{W}}m(x,z,w)(x-\mathbb{E}[X|Z=z,W=w])dF_{X,Z,W} \tag{A2.65}$$

$$=\int_{\mathcal{Z}\times\mathcal{X}\times\mathcal{W}}\left(\lim_{x\to-\infty}m(x)+\int_{-\infty}^{x}m'(t,z,w)dt\right)(x-\mathbb{E}[X|Z=z,W=w])\,dF_{X,Z,W} \tag{A2.66}$$

$$=\int_{\mathcal{Z}\times\mathcal{X}\times\mathcal{W}}\int_{-\infty}^{x}m'(t,z,w)\,(x-\mathbb{E}[X|Z=z,W=w])\,dtdF_{X,Z,W} \tag{A2.67}$$

$$=\int_{\mathcal{Z}\times\mathcal{W}}\int_{\mathcal{X}}\int_{-\infty}^{x}m'(t,z,w)\,(x-\mathbb{E}[X|Z=z,W=w])\,dtdF_{X|Z,W}dF_{Z,W} \tag{A2.68}$$

$$=\int_{\mathcal{Z}\times\mathcal{W}}\int_{\mathcal{X}}\int_{t}^{\infty}m'(t,z,w)\,(x-\mathbb{E}[X|Z=z,W=w])\,dF_{X|Z,W}dtdF_{Z,W} \tag{A2.69}$$

$$=\int_{\mathcal{Z}\times\mathcal{W}}\int_{\mathcal{X}}m'(t,z,w)\int_{t}^{\infty}(x-\mathbb{E}[X|Z=z,W=w])\,dF_{X|Z,W}dtdF_{Z,W} \tag{A2.70}$$

$$=\int_{\mathcal{Z}\times\mathcal{W}}\int_{\mathcal{X}}m'(t,z,w)\rho(z,w,t)dtdF_{Z,W} \tag{A2.71}$$

$$=\int_{\mathcal{X}}\int_{\mathcal{Z}\times\mathcal{W}}m'(t,z,w)\rho(z,w,t)dF_{Z,W}dt \tag{A2.72}$$

$$\rho(z,w,t):=(\mathbb{E}[|X\geq t|Z,W]-\mathbb{E}[|X<t|Z,W])\,(\mathbb{P}(X\geq t|Z,W])(1-\mathbb{P}(X\geq t|Z,W]) \tag{A2.73}$$

(A2.62) follows from the Law of Iterated Expectations, (A2.63) from the conditioning property since $(X-\mathbb{E}[X|Z,W]$ is a measurable function of the variables in the conditioning set. (A2.64) follows from expanding the outer expectation, (A2.65) from the fundamental theorem of calculus applied to $m(x,z,w)$. Equation (A2.66) follows from the fact that $\mathbb{E}\left[X-\mathbb{E}[X|Z=z,W=w]\right]=0$. Equations (A2.67) and (A2.68) are rearrangements based on Fubini's theorem and reversing the order of integration. Equation (A2.70) follows from

Angrist and Pischke [2008] and (A2.71) follows again from Fubini's theorem. Now for the denominator, it is entirely straightforward to characterize given the result in Angrist and Pischke [2008]. I report it here for completeness:

$$\mathbb{E}[X(X-\mathbb{E}[X|Z,W])] \tag{A2.74}$$

$$=\mathbb{E}[\mathbb{E}[X(X-\mathbb{E}[X|Z,W])]|X,Z,W] \tag{A2.75}$$

$$=\mathbb{E}[\mathbb{E}[X|X,Z,W](X-\mathbb{E}[X|Z,W])] \tag{A2.76}$$

$$=\int_{\mathcal{Z}\times\mathcal{X}\times\mathcal{W}} x(x-\mathbb{E}[X|Z=z,W=w])]dF_{X,Z,W} \tag{A2.77}$$

$$=\int_{\mathcal{Z}\times\mathcal{X}\times\mathcal{W}} \left(\int_{-\infty}^{x} dt\right)(x-\mathbb{E}[X|Z=z,W=w])\,dF_{X,Z,W} \tag{A2.78}$$

$$=\int_{\mathcal{Z}\times\mathcal{X}\times\mathcal{W}} \int_{-\infty}^{x} (x-\mathbb{E}[X|Z=z,W=w])\,dtdF_{X,Z,W} \tag{A2.79}$$

$$=\int_{\mathcal{Z}\times\mathcal{W}}\int_{\mathcal{X}} \int_{-\infty}^{x} (x-\mathbb{E}[X|Z=z,W=w])\,dtdF_{X|Z,W}dF_{Z,W} \tag{A2.80}$$

$$=\int_{\mathcal{Z}\times\mathcal{W}}\int_{\mathcal{X}} \int_{t}^{\infty} (x-\mathbb{E}[X|Z=z,W=w])\,dF_{X|Z,W}dtdF_{Z,W} \tag{A2.81}$$

$$=\int_{\mathcal{Z}\times\mathcal{W}}\int_{\mathcal{X}} \int_{t}^{\infty} (x-\mathbb{E}[X|Z=z,W=w])\,dF_{X|Z,W}dtdF_{Z,W} \tag{A2.82}$$

$$=\int_{\mathcal{Z}\times\mathcal{W}}\int_{\mathcal{X}} \rho(z,w,t)dtdF_{Z,W} \tag{A2.83}$$

$$=\int_{\mathcal{X}}\int_{\mathcal{Z}\times\mathcal{W}} \rho(z,w,t)dF_{Z,W}dt \tag{A2.84}$$

$$\rho(z,w,t) := (\mathbb{E}[|X\geq t|Z,W]-\mathbb{E}[|X<t|Z,W])\,(\mathbb{P}(X\geq t|Z,W])(1-\mathbb{P}(X\geq t|Z,W]) \tag{A2.85}$$

Observe that, if the term $\rho(z,w,t)$ did not depend on z,w then the OLS estimator

could indeed be interpreted as a weighted average of true causal effects since:

$$\begin{aligned}\frac{\int_{\mathcal{X}}\int_{\mathcal{Z}\times\mathcal{W}} m'(t,z,w)\rho(z,w,t)dF_{Z,W}dt}{\int_{\mathcal{Z}\times\mathcal{W}}\int_{\mathcal{X}}\rho(z,w,t)dtdF_{Z,W}} &= \frac{\int_{\mathcal{X}}\int_{\mathcal{Z}\times\mathcal{W}} m'(t,z,w)\rho(t)dF_{Z,W}dt}{\int_{\mathcal{Z}\times\mathcal{W}}\int_{\mathcal{X}}\rho(z,w,t)dtdF_{Z,W}} \\ &= \frac{\int_{\mathcal{X}}\rho(t)\int_{\mathcal{Z}\times\mathcal{W}} m'(t,z,w)dF_{Z,W}dt}{\int_{\mathcal{Z}\times\mathcal{W}}\int_{\mathcal{X}}\rho(z,w,t)dtdF_{Z,W}} \\ &= \frac{\int_{\mathcal{X}} AMTE(t)\rho(t)dt}{\int_{\mathcal{Z}\times\mathcal{W}}\int_{\mathcal{X}}\rho(t)dtdF_{Z,W}}\end{aligned}$$

But this is clearly not possible in any case of interest since by definition of $\rho(z,w,t)$ this would in general require $X \perp\!\!\!\perp Z,W$ which is certainly false as readily checked from the causal diagram. □

2.D Proofs

2.D.1 Proof of Proposition 37

Proposition 37 (Identification)**.** *Let $U \perp\!\!\!\perp X|Z,W$ and Assumption 1 hold. Then: i) the $AMTE(x_0)$ is nonparametrically identified by the following formula:*

$$AMTE(x_0) = \int \frac{\partial E[Y|X,Z,W]}{\partial x}(x_0,z,w)dF_{ZW}(z,w) \tag{2.4}$$

ii) Let instead $U \perp\!\!\!\perp X|Z$ and Assumption I hold. Then $AMTE(x_0)$ is nonparametrically identified by the following formula:

$$AMTE(x_0) = \int \frac{\partial E[Y|X,Z]}{\partial x}(x_0,z)dF_{Z}(z) \tag{2.5}$$

Proof. I show the first statement since the proof is identical for the second one. Here we further assume that (Z,W) are continuous random variables with joint density $f_{ZW}(z,w)$

although it is not necessary for the scope of the proof. We have:

$$AMTE(x_0) \tag{A2.86}$$

$$:= \int_{\mathcal{U}} \frac{\partial C}{\partial x}(x_0, u) f_u(u) du \tag{A2.87}$$

$$= \int_{\mathcal{U}} \lim_{\Delta x \to 0} \frac{\left[Y_{x_0+\Delta x}(u) - Y_{x_0}(u)\right]}{\Delta x} f_u(u) du \tag{A2.88}$$

$$= \int_{\mathcal{U}} \left(\int_{\mathcal{Z}\times\mathcal{W}} \lim_{\Delta x \to 0} \frac{\left[Y_{x_0+\Delta x}(u) - Y_{x_0}(u)\right]}{\Delta x} f_u(u|z,w) f_{ZW}(z,w) dz dw \right) du \tag{A2.89}$$

$$= \int_{\mathcal{Z}\times\mathcal{W}} \left(\int_{\mathcal{U}} \lim_{\Delta x \to 0} \frac{\left[Y_{x_0+\Delta x}(u) - Y_{x_0}(u)\right]}{\Delta x} f_u(u|z,w) du \right) f_{ZW}(z,w) dz dw \tag{A2.90}$$

$$= \int_{\mathcal{Z}\times\mathcal{W}} \left(\lim_{\Delta x \to 0} \int_{\mathcal{U}} \frac{\left[Y_{x_0+\Delta x}(u) - Y_{x_0}(u)\right]}{\Delta x} f_u(u|z,w) du \right) f_{ZW}(z,w) dz dw \tag{A2.91}$$

$$= \int_{\mathcal{Z}\times\mathcal{W}} \lim_{\Delta x \to 0} \frac{\int_{\mathcal{U}} Y_{x_0+\Delta x}(u) f_u(u|z,w) du - \int_{\mathcal{U}} Y_{x_0}(u) f_u(u|z,w) du}{\Delta x}$$

$$\times f_{ZW}(z,w) dz dw \tag{A2.92}$$

$$= \int_{\mathcal{Z}\times\mathcal{W}} \lim_{\Delta x \to 0} \frac{\int_{\mathcal{U}} Y_{x_0+\Delta x}(u) f_u(u|x_0+\Delta x,z,w) du - \int_{\mathcal{U}} Y_{x_0}(u) f_u(u|x_0,z,w) du}{\Delta x}$$

$$\times f_{ZW}(z,w) dz dw \tag{A2.93}$$

$$= \int_{\mathcal{Z}\times\mathcal{W}} \lim_{\Delta x \to 0} \frac{\mathbb{E}\left[Y_{x_0+\Delta x}|X,Z,W\right](x_0+\Delta x,z,w) - \mathbb{E}\left[Y_{x_0}|X,Z,W\right](x_0,z,w)}{\Delta x}$$

$$\times f_{ZW}(z,w) dz dw \tag{A2.94}$$

$$= \int_{\mathcal{Z}\times\mathcal{W}} \lim_{\Delta x \to 0} \frac{\mathbb{E}\left[Y|X,Z,W\right](x_0+\Delta x,z,w) - \mathbb{E}\left[Y|X,Z,W\right](x_0,z,w)}{\Delta x}$$

$$\times f_{ZW}(z,w) dz dw \tag{A2.95}$$

$$= \int_{\mathcal{Z}\times\mathcal{W}} \left(\frac{\partial \mathbb{E}\left[Y|X=x,Z,W\right]}{\partial x}(x_0,z,w) \right) f_{ZW}(z,w) dz dw \tag{A2.96}$$

Equation (A2.87) is the definition of $AMTE(x_0)$. Equation (A2.88) follows form the potential outcome notation and the definition of derivative, Equation (A2.89) follows from conditioning on both Z and W. Equation (A2.90) follows from Fubini's theorem. Equation

(A2.91) follows form the Dominated Convergence theorem, Equation (A2.92) from linearity and Equation (A2.93) from conditional independence $U \perp\!\!\!\perp X|Z,W$, Equation (A2.94) follows from the definition of conditional expectation. Equation (A2.95) is consistency of Y_{x_0} and $Y_{x_0+\Delta x}$ at $X = x_0$ and $X = x_0 + \Delta x$ respectively. Finally, Equation (A2.96) is the definition of derivative of the conditional expectation function at $X = x_0$. Since it depends exclusively on observed quantities, the $AMTE(x_0)$ is identified by the above population quantity. □

2.D.2 Proof of Proposition 58

Proposition 58 (Asymptotic Bias)**.** *The asymptotic bias resulting from the exclusion of W is characterized by the expression below.*

$$AsyBias = \int_{\mathcal{Z}} \int_{\mathcal{W}} \frac{\partial f_{W|X,Z}}{\partial x}(w|x_0, z) \int_U C(x_0, u) f_U(u|w, z) du dw f_Z(z) dz \tag{A2.53}$$

Proof. Using the same technique presented in Proposition 37 we have:

$$\widetilde{AMTE}(x_0) \tag{A2.97}$$

$$= \int_{\mathcal{Z}} \left(\frac{\partial \mathbb{E}\left[Y|X=x,Z\right]}{\partial x}(x_0,z) \right) f_Z(z)dz \tag{A2.98}$$

$$= \int_{\mathcal{Z}} \left(\lim_{\Delta x \to 0} \frac{\mathbb{E}\left[Y|X,Z\right](x_0+\Delta x,z) - \mathbb{E}\left[Y|X,Z\right](x_0,z)}{\Delta x} \right) f_Z(z)dz \tag{A2.99}$$

$$= \int_{\mathcal{Z}} \left(\lim_{\Delta x \to 0} \frac{1}{\Delta x} \cdot \left[\int_{\mathcal{W}} \int_{\mathcal{U}} Y_{x_0+\Delta x}(u) f_u(u|x_0+\Delta x,z,w) du f(w|x_0+\Delta x,z) dw \right.\right.$$

$$\left.\left. - \int_{\mathcal{W}} \int_{\mathcal{U}} Y_{x_0}(u) f_u(u|x_0,z,w) du f(w|x_0,z) dw \right] \right) f_Z(z)dz \tag{A2.100}$$

$$= \int_{\mathcal{Z}} \left(\lim_{\Delta x \to 0} \frac{1}{\Delta x} \cdot \left[\int_{\mathcal{W}} \int_{\mathcal{U}} Y_{x_0+\Delta x}(u) f_u(u|z,w) du f(w|x_0+\Delta x,z) dw \right.\right.$$

$$\left.\left. - \int_{\mathcal{W}} \int_{\mathcal{U}} Y_{x_0}(u) f_u(u|z,w) du f(w|x_0,z) dw \right] \right) f_Z(z)dz \tag{A2.101}$$

$$= \int_{\mathcal{Z}} \int_{\mathcal{W}} \left(\lim_{\Delta x \to 0} \frac{1}{\Delta x} \cdot \left[\int_{\mathcal{U}} Y_{x_0+\Delta x}(u) f_u(u|z,w) du f(w|x_0+\Delta x,z) dw \right.\right.$$

$$\left.\left. - \int_{\mathcal{U}} Y_{x_0}(u) f_u(u|z,w) du f(w|x_0,z) dw \right] \right) f_Z(z)dz \tag{A2.102}$$

$$= \int_{\mathcal{Z}} \int_{\mathcal{W}} \left(\lim_{\Delta x \to 0} \frac{1}{\Delta x} \cdot \left[\int_{\mathcal{U}} C(x_0+\Delta x,u) f_u(u|z,w) du f(w|x_0+\Delta x,z) dw \right.\right.$$

$$\left.\left. - \int_{\mathcal{U}} C(x_0,u) f_u(u|z,w) du f(w|x_0,z) dw \right] \right) f_Z(z)dz \tag{A2.103}$$

$$= \int_{\mathcal{Z}} \int_{\mathcal{W}} \left(\int_{\mathcal{U}} \frac{\partial C}{\partial x}(x_0,u) f_u(u|z,w) f(w|x_0,z) du \right.$$

$$\left. + \frac{\partial f_{w|x,z}}{\partial x}(w|x_0,z) \int_U C(x_0,u) f(u|w,z) du \right) dw f_Z(z)dz \tag{A2.104}$$

$$= \int_{\mathcal{Z}} \int_{\mathcal{W}} \left(\int_{\mathcal{U}} \frac{\partial C}{\partial x}(x_0,u) f_u(u|z,w) f(w|x_0,z) du dw f_Z(z)dz \right.$$

$$\left. + \int_{\mathcal{Z}} \int_{\mathcal{W}} \frac{\partial f_{w|x,z}}{\partial x}(w|x_0,z) \int_U C(x_0,u) f(u|w,z) du dw f_Z(z)dz \right. \tag{A2.105}$$

There are two terms in Equation (A2.105). The first one corresponds to the $AMTE(x_0)$ with the caveat that the integration along $\mathcal{W}$ is with respect to the conditional density $f(w|x_0, z)$ rather than the conditional density $f(w|z)$. The second term arises because an incremental change in x about x_0 changes the distribution of the control W conditional on X. If $W \perp\!\!\!\perp X|Z$ then the estimator that controls only for Z correctly identifies the causal effect of interest, as noted in Proposition 37. Equation (A2.105) explains why this is the case. When $W \perp\!\!\!\perp X|Z$, $f(w|x_0, z) = f(w|z)$ and the first term exactly equals the $AMTE(x_0)$. Moreover, $W \perp\!\!\!\perp X|Z$ also implies that $\frac{\partial f_{w|x,z}}{\partial x}(w|x_0, z)$ is the 0 function, which in turn makes the whole second term 0. Hence $\widetilde{AMTE}(x_0) = AMTE(x_0)$ In general though, if W is needed for the identification of $AMTE(x_0)$ omitting it would induce the bias formula characterized above. □

2.D.3 Proof of Proposition 62

Definition 61. *Operator norm*

Let V and W be two normed vector spaces over $\mathbb{R}$ and $T : V \to W$. Then the operator norm of T is given by:

$$\|T\|_{op} := \inf\{c \text{ such that } \|Av\|_W \leq c\|v\|_V, \text{ for all } v \in V\}$$

Lemma 62 (Operator Norms)**.** *i) Let $V_1 = \mathcal{C}^k(\mathcal{X} \times \mathcal{Z} \times \mathcal{W})$ and $V_2 = \mathcal{C}^{k-1}(\mathcal{X} \times \mathcal{Z} \times \mathcal{W})$ both endowed with the strong norm and let D be the partial differentiation operator, i.e.*

$$D : f \mapsto \frac{\partial f}{\partial x}$$

Then $\|D\|_{op} \leq 1$.

ii) Let $V_3 = \mathcal{C}^{k-1}(\mathcal{Z} \times \mathcal{W})$ endowed with the operator norm. Let the $(\cdot)_{x=x_0} : V_2 \to V_3$ be the

evaluation at x_0 map, i.e.

$$(\cdot)_{x=x_0} : f(x,z,w) \mapsto f(x_0,z,w)$$

Then $\|(\cdot)_{x=x_0}\|_{op} \leq 1$.

iii) Let $\mathcal{I} : V_3 \to \mathbb{R}$ *be the functional:*

$$\mathcal{I} : f(x_0,z,w) \mapsto \mathbb{E}[f(x_0,z,w)]$$

Then $\|\mathcal{I}\|_{op} \leq 1$.

Proof. i) Let $f \in V_1$.

$$\begin{aligned}
\|Df\|_{d,V_2} &= \max_{\lambda_x+\lambda_z+\lambda_w \leq k-1} \left(\sup_{(x,z,w)\in\mathcal{X}\times\mathcal{Z}\times\mathcal{W}} \left| \frac{\partial^\lambda (\frac{\partial f}{\partial x})}{\partial x^{\lambda_x} \partial z^{\lambda_z} \partial w^{\lambda_w}} \right| \right) \\
&\leq \max_{\lambda_x+\lambda_z+\lambda_w \leq k} \left(\sup_{(x,z,w)\in\mathcal{X}\times\mathcal{Z}\times\mathcal{W}} \left| \frac{\partial^\lambda f}{\partial x^{\lambda_x} \partial z^{\lambda_z} \partial w^{\lambda_w}} \right| \right) \\
&= \|f\|_{d,V_1}
\end{aligned}$$

So by the definition of operator norm we must have $\|D\|_{op} \leq 1$.

ii) Now take $g \in V_2$. We have:

$$\begin{aligned}
\|(\cdot)|_{x=x_0} g\|_{d,V_3} &= \max_{\lambda_z+\lambda_w \leq k} \left(\sup_{(z,w)\in\times\mathcal{Z}\times\mathcal{W}} \left| \frac{\partial^\lambda g(x_0,z,w))}{\partial z^{\lambda_z} \partial w^{\lambda_w}} \right| \right) \\
&= \max_{\lambda_x+\lambda_z+\lambda_w \leq k} \left(\sup_{(x,z,w)\in\mathcal{X}\times\mathcal{Z}\times\mathcal{W}} \left| \frac{\partial^\lambda (g)}{\partial x^{\lambda_x} \partial z^{\lambda_z} \partial w^{\lambda_w}} \right| \right) \\
&= \|g\|_{d,V_2}
\end{aligned}$$

iii) Finally take, $h \in V_3$. We have:

$$\begin{aligned}
\|(\mathcal{I}h)\| &= \mathbb{E}[|h(Z,W)|] \\
&\leq \mathbb{E}\left[\max_{\lambda_z+\lambda_w\leq k}\left(\sup_{(z,w)\in\mathcal{Z}\times\mathcal{W}}\left|\frac{\partial^\lambda h}{\partial z^{\lambda_z}\partial w^{\lambda_w}}(Z,W)\right|\right)\right] \\
&= \max_{\lambda_z+\lambda_w\leq k}\left(\sup_{(z,w)\times\mathcal{Z}\times\mathcal{W}}\left|\frac{\partial^\lambda h}{\partial z^{\lambda_z}\partial w^{\lambda_w}}\right|\right) \\
&= \|h\|_{d,V_3}
\end{aligned}$$

□

2.D.4 Proof of Lemma 42

Proposition 43 (Linearity and Continuity)**.** *For any S, and any ω $\Gamma_{S,\omega}$ is a continuous linear functional with respect to the strong norm.*

Proof. Recall that, for V_1, V_2, V_3, V_4 be normed linear spaces and $T_1 : V_1 \to V_2, T_2 : V_2 \to V_3, T_3 : V_3 \to V_4$ linear mappings between these spaces we have the following operator norm inequality:

$$\|T_1 \circ T_2 \circ T_3\|_{op} \leq \|T_1\|_{op}\|T_2\|_{op}\|T_3\|_{op}$$

By Lemma 62 and the operator norm inequality, we have:

$$\begin{aligned}
\|(\mathcal{I} \circ (\cdot)|_{x=x_0} \circ D)\|_{op} &\leq \|\mathcal{I}\|_{op} \cdot \|(\cdot)|_{x=x_0}\|_{op} \cdot \|D\|_{op} \\
&\leq 1 \cdot 1 \cdot 1 \\
&= 1
\end{aligned}$$

Therefore, by the definition of the operator norm, Γ is bounded with respect to the strong norm. But then because bounded-ness implies continuity we conclude that the functional

of interest is continuous with respect to the strong norm. □

Chapter 3

Marginal Treatment Effects with Misspecification

3.1 Introduction

Marginal treatment effects (MTEs) have unified the identification theory of several policy parameters. While the MTE framework is essentially non-parametric,[1] it is required that the recipient's participation into treatment follows a (generalized) Roy model. This is often referred to as additive separability: an "additive" comparison of costs and benefits determines selection. On the other hand, identification of the MTE is achieved via the local instrumental variable (LIV) approach (Heckman and Vytlacil [2001, 2005]). An excellent survey is provided by Mogstad and Torgovitsky [2018]. An early effort to analyze MTE under misspecification can be found in the appendix of the seminal paper by Heckman and Vytlacil [2001]. They consider a case where the additive separability in the selection equation does not hold. The most serious consequence is that the LIV approach does not identify the MTE curve.

[1] Linearity is sometimes assumed to facilitate estimation. See, e.g., Appendix B in Heckman et al. [2006]

In this paper we analyze a different type of misspecification. We model a situation in which, under additive separability, a proportion of the population does not take into account the instrumental variable when deciding whether to take up treatment or not. We refer to them as non-responders. To analyze the resulting bias, we define a pseudo-MTE curve which results from the LIV approach. Under no misspecification, the pseudo-MTE curve would coincide with the MTE curve. The resulting bias can be interpreted as a location-scale change of the MTE curve, parameterized by the proportion of non-responders and their propensity score.

We have two main results. The first one shows that the ability to recover the conditional average treatment effect (CATE) for the subpopulation of responders depends on the proportion of non-responders only through the support of the responders' propensity score. Indeed, when the support of the propensity score is the unit interval, it is possible to identify the CATE *without* having to recover the true MTE curve in the first place. In a nutshell, ignoring misspecification and integrating under the pseudo-MTE curve over the support of observed propensity score yields the correct CATE for the subpopulation of responders.

While the previous identification result for the CATE is independent of the proportion of non-responders, this is not true of the MTE curve and other parameters derived from it such as LATE and MPRTE. However, in our second result, we show how to recover the MTE curve for responders by undoing the location-scale change induced by the presence of non-responders. The correction is based on an estimate of the support of the propensity score and requires only observable data. It gives an estimator of the policy parameter of interest that is simple to implement. Cases where the propensity score is fully supported are relevant in practice. For a recent example, see the survey approach of Briggs et al. [2020] the probability of having a child is supported on the full unit interval.

Recently, Acerenza et al. [2021] and Possebom [2021] focus on the effect of mea-

surement error in treatment status on the MTE curve. We complement such results by noting that a simple change to our setup can cover the case of misclassification. In a setting where treatment status is misclassified, the observed outcome is generated with the true treatment status. In our setting of misclassification, the observed outcome can be regarded as a mixture of responders and non-responders. The proportion of non-responders is analogous to the proportion of misreporters. Indeed, our results also hold if instead of having a fraction of non-responders, we have a fraction of misreporters.

Another consequence of the presence of non-responders in the sample is that the effect of the instrumental variable on the propensity score is attenuated. Motivated by this, we model a situation where the proportion of non-responders approaches 1, analogous to the setting of weak instruments of Staiger and Stock [1997]. Thus, we can derive weak-instrument-like asymptotic distributions for the parameters derived from the MTE curve.

The rest of the paper is organized as follows: section 3.2 introduces the model; section 3.3 contains the main identification results; section 3.4 provides bounds for the case where the propensity score is not fully supported in the unit interval; section 3.5 traces the connection to the weak IV literature; and section 3.7 concludes. While this paper only deals with identification, we expect to extend our results to cover estimation and inference.

3.2 Misspecification and MTE

In this section we introduce our model for misspecification in the MTE framework (Bjorklund and Moffitt [1987], Heckman and Vytlacil [2001, 2005]). We analyze the consequences of misspecification from the identification point of view.

3.2.1 The Model

We start with a general non-separable potential outcome model

$$\begin{aligned} Y(0) &= h_0(X, U_0), \\ Y(1) &= h_1(X, U_1), \\ Y &= D^* Y(1) + (1 - D^*) Y(0), \end{aligned}$$

where D^* is the observed treatment status, X are observable covariates with support denoted by $\mathcal{X}$, and $\{Y(0), Y(1)\}, Y$ are potential and observed outcomes, respectively. The functions h_0 and h_1 are unknown.

We model misspecification as a situation where there are two types of individuals: responders and non-responders. Responders select into treatment taking into account the incentives in Z. Their selection equation is given by $D = \mathbb{1}\{\mu(X, Z) \geq V\}$. On the other hand, non-responders do not react to incentives in Z at all. Their selection equation is given by $\tilde{D} = \mathbb{1}\left\{\tilde{\mu}(X) \geq \tilde{V}\right\}$. Notice how Z is not featured in $\tilde{\mu}(\cdot)$. For the non-responders, Z fails the relevance condition of the standard MTE model.

Let S be the latent status of an individual: $S = 1$ for a responder and $S = 0$ for a non-responder. The observed treatment status D^* is given by:

$$D^* = S \cdot D + (1 - S) \cdot \tilde{D}. \tag{3.1}$$

We allow for the proportion of non-responders may vary with X. To this end, we define $\delta_X = \Pr(S = 0|X) = \Pr(D^* = \tilde{D}|X)$. Thus, for every subpopulation with characteristics $X = x$ there is a proportion $\delta_x = \Pr(S = 0|X = x) \in [0, 1)$ of non-responders. We consider values where $\sup_{x \in \mathcal{X}} \delta_x < 1$ to avoid a situation where no-one responds to the instrumental variable.

Remark 63. *We observe Y according to $Y = D^*Y(1) + (1 - D^*)Y(0)$, which is given by the actual choice D^*. If, instead, we have $Y = DY(1) + (1 - D)Y(0)$, then we can interpret D^* as a misclassified treatment status. In this case, all individuals decide according to $D = \mathbb{1}\{\mu(X, Z) \geq V\}$, but a fraction of them reports according to $\tilde{D} = \mathbb{1}\left\{\tilde{\mu}(X) \geq \tilde{V}\right\}$ See Acerenza et al. [2021] and Possebom [2021] for recent studies on MTE under misclassification.*

The econometrician observes a cross section of (Y_i, D_i^*, X_i, Z_i). When $\delta_X = 0$ almost surely, then $D^* = D$ and we are in the familiar MTE framework of Heckman and Vytlacil [2001, 2005]. Otherwise, if $\delta_X \neq 0$ almost surely, for an observation of D_i^*, we do not know whether we are observing the treatment status of a non-responder or of a responder. That is, it is unknown if we are observing D_i or $\tilde{D}_i$.

Assumption 10. ***Type Independence.*** $S \perp Z \| X$.

Assumption 10 states that once we control for X, the latent status of a individuals does not vary with the instrumental variable Z.

Assumption 11. ***Relevance and Exogeneity***

1. *$\mu(X, Z)$ is a nondegenerate random variable conditional on X.*
2. *$(U_0, U_1, V, \tilde{V})$ are independent of Z conditional on X.*

Note that, for the subpopulation of non-responders, the instrument is valid but totally irrelevant. The larger the value of δ_x, the "weaker" the instrument Z, since most participants with $X = x$ are non-responders. With the exception of the requirement that $\tilde{V} \perp Z \| X$, these are the same conditions of Heckman and Vytlacil [2001, 2005]. Our additional requirement covers the subpopulation of non-responders: neither the "cost" of treatment $\tilde{V}$ nor the "benefit" $\tilde{\mu}(X)$ depend on Z when conditioned on X.

Example 64. *To fix ideas, we can think of a two part cost of providing the incentive. A fixed cost associated to targeting a particular subpopulation with covariates $X = x$ and the*

cost of the incentive itself. If Z is a voucher, there could be administrative costs associated to making it available to subpopulation $X = x$. For non-responders who do not redeem the voucher, the cost of the incentive is zero. Such a scenario would satisfy Assumption 11.

The misclassification structure of Equation (3.1) allows to define three different propensity scores. An observed/identified one which is based on the observables (D^*, X, Z), and two latent/unobserved propensity scores: one for the responders and one for the non-responders. Formally, they are given by

$$\begin{aligned} P^*(X,Z) &:= \Pr(D^* = 1|X,Z) && \textbf{(Observed)} \\ P(X,Z) &:= \Pr(D = 1|S = 1,X,Z) && \textbf{(Responders)} \\ \tilde{P}(X) &:= \Pr(\tilde{D} = 1|S = 0,X) && \textbf{(Non-responders)} \end{aligned}$$

The next result takes (mainly) advantage of Assumption 10 to derive a useful affine relation between them.

Lemma 65. *Under Assumptions 10 and 11.2 we can relate the different propensity scores by*

$$P^*(X,Z) = (1-\delta_X)\cdot P(X,Z) + \delta_X \cdot \tilde{P}(X). \tag{3.2}$$

Proof. Starting with the model in (3.1) we can write

$$\begin{aligned} \Pr(D^* = 1|X,Z) = \Pr(S = 1|X,Z)\cdot \Pr(D = 1|S = 1,X,Z) \\ + \Pr(S = 0|X,Z)\cdot \Pr(\tilde{D} = 1|S = 0,X,Z). \end{aligned}$$

Assumption 10 simplifies the mixing probabilities to $\Pr(S = 1|X) = 1-\delta_X$ and $\Pr(S =$

$0|X) = \delta_X$. We obtain

$$\Pr(D^* = 1|X, Z) = (1 - \delta_X) \cdot \Pr(D = 1|S = 1, X, Z) + \delta_X \cdot \Pr(\tilde{D} = 1|S = 0, X, Z).$$

To see that $Pr(\tilde{D} = 1|S = 0, X, Z) = Pr(\tilde{D} = 1|S = 0, X)$, we note that By Assumptions 10 and 11.2:

$$\begin{aligned}
\Pr(\tilde{D} = 1|S = 0, X, Z) &= \Pr(\tilde{\mu}(X) \geq \tilde{V}|S = 0, X, Z) \\
&= \Pr(\tilde{\mu}(X) \geq \tilde{V}|X) \\
&= \Pr(\tilde{D} = 1|S = 0, X).
\end{aligned}$$

Therefore

$$\begin{aligned}
\Pr(D^* = 1|X, Z) &= (1 - \delta_X) \cdot \Pr(D = 1|S = 1, X, Z) + \delta_X \cdot \Pr(\tilde{D} = 1|S = 0, X) \\
&= (1 - \delta_X) \cdot P(X, Z) + \delta_X \cdot \tilde{P}(X).
\end{aligned}$$

□

For a fixed $X = x$, the result in Lemma 65 shows that the observed propensity (still random through Z) is a linear transformation of the propensity score for the responders. If, additionally, we take two different values of Z, for example z and z', we can remove the contribution of $\tilde{P}(X)$, which is invariant with respect to z and obtain[2]

$$P^*(x, z) - P^*(x, z') = (1 - \delta_x) \cdot \Big[P(x, z) - P(x, z')\Big] \tag{3.3}$$

Equation (3.3) says that the changes on the observed propensity score induced by varying Z are proportional to the changes on the true propensity score induced by varying Z. Thus,

[2]We write $P^*(x, z)$ for $\Pr(D^* = 1|X = x, Z = z)$, and $P(x, z)$ for $\Pr(D = 1|S = 1, X = x, Z = z)$.

if we knew δ_x, we could recover the change in the propensity score for the responders. When Z is continuous, we can take a limiting version of this argument, *e.g.*, as $z' \to z$, to obtain

$$\frac{\partial P^*(x,z)}{\partial z} = (1-\delta_x) \cdot \frac{\partial P(x,z)}{\partial z}. \tag{3.4}$$

Both the discrete (equation (3.3)), and the continuous (equation(3.4)) change in the propensity score play a role in the relationship between the MTE curve (defined below) and certain parameters of interest.

3.2.2 The MTE for Responders

For the subpopulation of responders, the standard MTE framework holds. This motivates us to define an MTE curve for this subpopulation. In doing so, we are implicitly assuming that this is our object of interest. The reason for this is that many times we can also control the instrumental variable Z. Thus, to asses the effects of manipulations of Z we look at the MTE curve for responders.

Let $\mathcal{P}_x$ and $\mathcal{P}_x^*$ denote the support of $P(x,Z) := \Pr(D=1|X=x,Z)$ and $P^*(x,Z) := \Pr(D^*=1|X=x,Z)$ respectively. For the subpopulation of responders, we rewrite the selection equation as $D = \mathbb{1}\{P(X,Z) \geq U_D\}$ where $U_D \sim U_{(0,1)}$.[3] Thus, we define the MTE curve for responders as

$$\text{MTE}(u,x) := \mathbb{E}\left[Y(1) - Y(0) | S=1, U_D = u, X = x\right].$$

[3]This follows from $D = \mathbb{1}\{F_{V|S,X,Z}(\mu(X,Z)|1,X,Z) \geq F_{V|S,X,Z}(V|1,X,Z)\}$. Noting that by assumptions 11.(2) and 10, we have $D = \mathbb{1}\{P(X,Z) \geq F_{V|S,X}(V|1,X)\}$. Finally, we take $U_D := F_{V|S,X}(V|1,X)$.

By the LIV approach we have the following equivalence result:[4]

$$\text{MTE}(u,x) = \frac{\partial \mathbb{E}[Y|S=1, P(X,Z)=u, X=x]}{\partial u} \text{ for } u \in \mathcal{P}_x. \tag{3.5}$$

Since we do not observe $P(X,Z)$, this is *not* an identification result in our setting. In a similar fashion, we *define* the following pseudo-MTE curve:

$$\text{MTE}^*(u,x;\delta_x) := \frac{\partial \mathbb{E}[Y|P^*(X,Z)=u, X=x]}{\partial u} \text{ for } u \in \mathcal{P}_x^*. \tag{3.6}$$

We emphasize that the pseudo-MTE curve is indexed by δ_x because it depends implicitly on the proportion of the nonresponders. From the data only, we can only compute $\text{MTE}^*(u,x;\delta_x)$, not $\text{MTE}(u,x)$. The pseudo-MTE curve is the curve that would be mistakenly taken to be the MTE curve. Indeed, in the absence of non-responders, $\text{MTE}^*(u,x;0) = \text{MTE}(u,x)$. If non-responders are present in the $X=x$ subpopulation, that is if $\delta_x > 0$, the observed $\text{MTE}^*(u,x;\delta_x)$ does not identify $\text{MTE}(u,x)$. In another words, the LIV approach is biased. We can now fully characterize the bias induced by δ_x on the MTE curve.

Lemma 66. *Under Assumptions 10 and 11, we can write*

$$MTE(v,x) = (1-\delta_x) MTE^* \left((1-\delta_x)v + \delta_x \tilde{P}(x), x; \delta_x \right) \text{ for } v \in \mathcal{P}_x. \tag{3.7}$$

Proof. Using (3.2), for $u \in \mathcal{P}_x^*$, we can write

$$\begin{aligned}
\mathbb{E}[Y|P^*(X,Z)=u, X=x] &= \mathbb{E}\left[Y|(1-\delta_x)\cdot P(X,Z) + \delta_x \cdot \tilde{P}(X) = u, X=x\right] \\
&= \mathbb{E}\left[Y \middle| P(X,Z) = \frac{u-\delta_x \tilde{P}(x)}{1-\delta_x}, X=x\right]
\end{aligned}$$

[4] See Heckman and Vytlacil [2001] for sufficient conditions.

Differentiating with respect to u, we obtain

$$\text{MTE}^*(u,x;\delta_x) = \frac{1}{1-\delta_x}\text{MTE}\left(\frac{u-\delta_x\tilde{P}(x)}{1-\delta_x},x\right) \text{ for } u \in \mathcal{P}_x^*. \tag{3.8}$$

since $\frac{u-\delta_x\tilde{P}(x)}{1-\delta} \in \mathcal{P}_x$ by (3.2). Alternatively, we can write

$$\text{MTE}(v,x) = (1-\delta_x)\text{MTE}^*\left((1-\delta_x)v+\delta_x\tilde{P}(x),x;\delta_x\right) \text{ for } v \in \mathcal{P}_x.$$

□

Lemma 66 shows that the bias is in the form of both location and scale. Equation (3.8), which is equivalent to Equation (3.7),[5] shows that MTE* is obtained by changing the location from u to $u-\delta_x\tilde{P}(x)$, and rescaling by $(1-\delta_x)^{-1}$. Thus, as in a location-scale family of densities, we can regard MTE* as a family of curves, defined over $\mathcal{P}_x^*$, which is indexed by δ_x and $\tilde{P}(x)$.

3.3 Automatic and explicit de-biasing

We now introduce our two main results. We show that, for any subpopulation $X=x$ where the instrument is strong enough to induce a propensity score supported on the full unit interval $[0,1]$, the associated $CATE(x)$ can be identified for responders. This is true even if the $MTE^*(u,x,\delta_x)$ curve is biased for $MTE(u,x)$. We note that the identified $CATE(x)$ parameters corresponds to the subpopulation of responders.

Assumption 12. ***Full Support.*** *The support of $P(x,Z)$ is $\mathcal{P}_x=[0,1]$ for every x in a subset $\mathcal{X}_B \subseteq \mathcal{X}$.*

Assumption 12 says that the incentive in the instrument Z is strong enough to induce

[5]Note the changes in the domain of integration between (3.7) and (3.8).

any individual in the $X = x$ subpopulation into or out of treatment. Perhaps surprisingly, the CATE(x), can be recovered only by resorting to the full support assumption. That is, to correctly compute the CATE(x) we do not need to recover the true MTE curve for responders.

Theorem 67. *Let Assumptions 10, 11, and 12 hold. Then, for any $x \in \mathcal{X}_B$:*

$$CATE(x) = \int_{\inf \mathcal{P}_x^*}^{\sup \mathcal{P}_x^*} MTE^*(u, x; \delta_x) du.$$

Proof. The Conditional Average Treatment Effect, CATE(x), could be computed using the true MTE curve (if it was observed) as

$$\text{CATE}(x) = \int_0^1 \text{MTE}(u, x) du.$$

Given that $\mathcal{P}_x = [0, 1]$, then $\mathcal{P}_x^* := [\underline{p_x^*}, \overline{p_x^*}]$ where $\underline{p_x^*} := \inf \mathcal{P}_x^* = \delta_x \tilde{P}(x)$ and $\overline{p_x^*} := \sup \mathcal{P}_x^*(1 - \delta_x) + \delta_x \tilde{P}(x)]$. Consider the integrating the pseudo-MTE curve over the support of the observed propensity score:

$$\int_{\delta_x \tilde{P}(x)}^{(1-\delta_x)+\delta_x \tilde{P}(x)} \text{MTE}^*(u, x; \delta_x) du.$$

Using (3.8), we have

$$\begin{aligned}
\int_{\delta_x \tilde{P}(x)}^{(1-\delta_x)+\delta_x \tilde{P}(x)} \text{MTE}^*(u, x; \delta_x) du &= \int_{\delta_x \tilde{P}(x)}^{(1-\delta_x)+\delta_x \tilde{P}(x)} \frac{1}{1-\delta_x} \text{MTE}\left(\frac{u - \delta_x \tilde{P}(x)}{1-\delta_x}, x \right) du \\
&= \int_0^1 \text{MTE}(u, x) du \\
&= \text{CATE}(x)
\end{aligned}$$

where we have done the change of variables

$$v = \frac{u - \delta_x \tilde{P}(x)}{1 - \delta_x}.$$

□

Remark 68. *The result of Theorem 67 states that by integrating the observed (and biased) marginal treatment effect curve over the support of the observed (and biased) propensity score leads to the CATE(x) provided that the propensity score for responders has full support. Thus, under the type of misspecification described in* (3.1)*, CATE(x) is robust to $\delta_x \neq 0$.*

Remark 69. *This result also hold in a setting of misclassification and was our original motivation. That is, in a setting where instead of $Y = D^*Y(1) + (1 - D^*)Y(0)$, we have $Y = DY(1) + (1 - D)Y(0)$ and we interpret D^* as a misclassified treatment status.*

Unfortunately, the automatic "de-biasing" in Theorem 67 does not hold for the other policy parameters that can be obtained via the MTE curve. On the other hand, we show that the full support assumption can be used to identify δ_x which allows an explicit "de-biasing" procedure. Given that $\mathcal{P}_x^* := [\underline{p_x^*}, \overline{p_x^*}] = [\delta_x \tilde{P}(x), (1 - \delta_x) + \delta_x \tilde{P}(x)]$ we can actually identify both δ_x and $\tilde{P}(x)$. It follows then from Lemma 66 that we can recover the MTE(u, x) curve.

Proposition 70. *Let Assumptions 10, 11, and 12 hold. Then δ_x is identified for any $x \in \mathcal{X}_B$ through:*

$$\delta_x = 1 - (\overline{p_x^*} - \underline{p_x^*})$$

Proof. According to Equation (3.2), the range of the observed propensity score is given by $\mathcal{P}_x^* = [\delta_x \tilde{P}(x), (1 - \delta_x) + \delta_x \tilde{P}(x)]$. For each x, the observed propensity score $P^*(\cdot)$ can be viewed as an affine function of $P(\cdot)$. This affine function is parameterized by δ_x and $\tilde{P}_x$.

For the endpoints $\underline{p_x}$ and $\overline{p}_x$ of the true propensity score, we have the mappings:

$$\underline{p_x} \mapsto (1-\delta_x)\underline{p_x} + \delta_x \tilde{P}(x)$$
$$\overline{p_x} \mapsto (1-\delta_x)\overline{p_x} + \delta_x \tilde{P}(x)$$

The images of this collection of mapping are observed. They are the endpoints of the observed propensity score $P^*(Z,x)$. If the original endpoints of the true $P(\cdot)$ are known to be $\underline{p_x} = 0$ and $\overline{p_x} = 1$, like stated in Assumption 12, the mapping above can be recovered by the following system of two equations in two unknowns: $\tilde{P}(x)$ and δ_x.

$$\underline{p_x^*} = \delta_x \tilde{P}(x)$$
$$\overline{p_x^*} = (1-\delta_x) + \delta_x \tilde{P}(x)$$

which implies that

$$\delta_x = 1 - (\overline{p_x^*} - \underline{p_x^*})$$
$$\tilde{P}(x) = \underline{p_x^*} \cdot \frac{1}{\delta_x}$$

□

The intuition for this result is simple. Because the original propensity score $P(Z,x)$, for any fixed x, is supported on the unit interval, the observed support $\mathcal{P}_x^* = [\underline{p_x^*}, \overline{p_x^*}]$ will contain enough information to identify δ_x. This is summarized Figure 3.3.1.

Having identified δ_x, then we use Equation (3.8) to identify the MTE curve.

Corollary 71. *Let Assumptions 10, 11, and 12 hold. Then, the MTE curve is identified:*

$$MTE(v,x) = (\overline{p_x^*} - \underline{p_x^*}) MTE^* \left((\overline{p_x^*} - \underline{p_x^*})v + \underline{p_x^*}, x; 1 - (\overline{p_x^*} - \underline{p_x^*}) \right) \text{ for } v \in \mathcal{P}_x = [0,1].$$

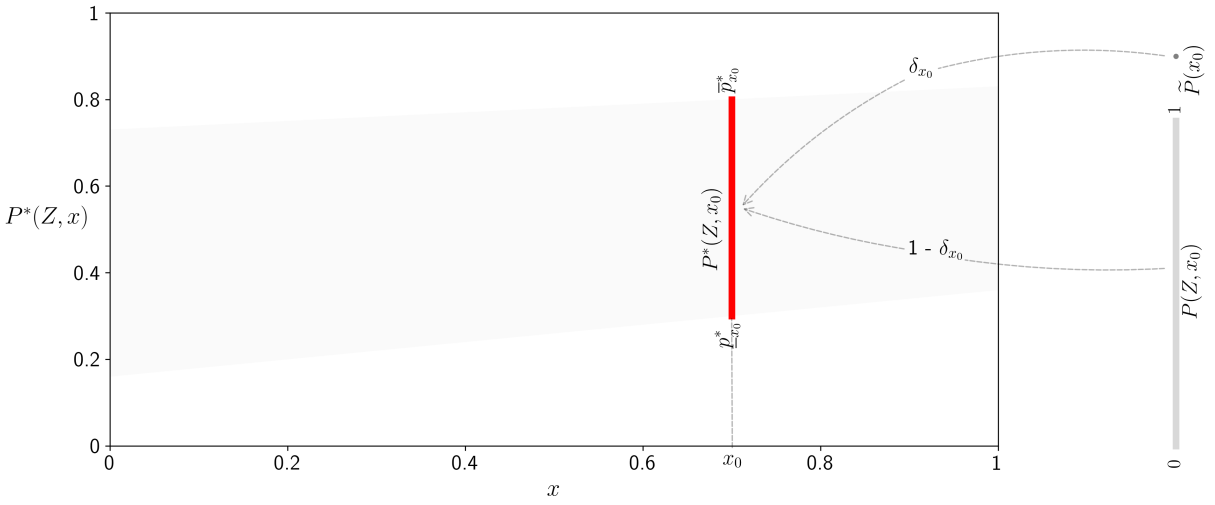


Figure 3.3.1: Identifying δ_x: The figure shows the link between the non-responders propensity score, the proportion of non-responders and the observed propensity score. Because the non-responders propensity score does not vary with the instrument Z and $supp(P(Z,x)) = [0,1]$ the δ_x can be recovered from observing the discrepancy from the observed support $P^*(Z,x)$ and $[0,1]$. The picture shows one of those points, x_0.

where $\underline{p^*_x} = \inf \mathcal{P}^*_x$ *and* $\overline{p^*_x} = \sup \mathcal{P}^*_x$.

This corollary provides the correct "de-biasing" to be performed on the observed MTE curve to match the true MTE curve. However, it is possible to recover parameters that are based on the MTE curve *without* having to recover the MTE curve in the first place. We provide two examples.

Example 72 (LATE)**.** *Consider the LATE, for* $P(x,z') < P(x,z)$ *with* $z, z' \in \mathcal{Z}$*, which can be obtained from MTE curve as*

$$LATE(x, P(x,z), P(x,z')) = \frac{1}{P(x,z) - P(x,z')} \int_{P(x,z')}^{P(x,z)} MTE(u,x)du.$$

Under misspecification, for the same $z, z' \in \mathcal{Z}$, we have

$$\begin{aligned} LATE^*(x, P^*(x,z), P^*(x,z')) &= \frac{1}{P^*(x,z) - P^*(x,z')} \int_{P^*(x,z')}^{P^*(x,z)} MTE^*(u,x;\delta_x)du \\ &= \frac{(1-\delta_x)^{-1}}{P(x,z) - P(x,z')} \int_{(1-\delta_x)P(x,z')+\delta_x\tilde{P}(x)}^{(1-\delta_x)P(x,z)+\delta_x\tilde{P}(x)} \frac{1}{1-\delta_x} \\ &\times MTE\left(\frac{u-\delta_x\tilde{P}(x)}{1-\delta_x}, x\right) du. \end{aligned}$$

Note that to go from MTE^* to MTE we used Lemma 66. We did not use Corollary 71. Defining the change of variables $\tilde{u} = \frac{u-\delta_x\tilde{P}(x)}{1-\delta_x}$, we get $(1-\delta_x)d\tilde{u} = du$. We then write

$$\begin{aligned} LATE^*(x, P^*(x,z), P^*(x,z')) &= \frac{(1-\delta_x)^{-1}}{P(x,z) - P(x,z')} \int_{(1-\delta_x)P(x,z')+\delta_x\tilde{P}(x)}^{(1-\delta_x)P(x,z)+\delta_x\tilde{P}(x)} \frac{1}{1-\delta_x} \\ &\times MTE\left(\frac{u-\delta_x\tilde{P}(x)}{1-\delta_x}, x\right) du \\ &= \frac{(1-\delta_x)^{-1}}{P(x,z) - P(x,z')} \int_{P(x,z')}^{P(x,z)} MTE(u,x)du \\ &= \frac{1}{1-\delta_x} LATE(x, P(x,z), P(x,z')). \end{aligned}$$

Now, since $\delta_x = 1 - (\overline{p_x^*} - \underline{p_x^*})$ by Proposition 70, the explicit de-biasing is achieved by

$$(\overline{p_x^*} - \underline{p_x^*}) LATE^*(x, P^*(x,z), P^*(x,z')) = LATE(x, P(x,z), P(x,z')).$$

The left hand side can be computed from the data.

Example 73 (MPRTE)**.** *The marginal policy relevant treatment effect (MPRTE) is an average of the MTE(u,x) along the margin of indifference: when $U_D = P(X,Z)$. It is given by*

$$MPRTE(x) = \int_{\mathcal{Z}} MTE(P(x,z),x) \frac{\partial P(x,z)}{\partial z} \left(E\left[\frac{\partial [P(x,Z)]}{\partial z} \right] \right)^{-1} f_{Z|X}(z|x)dz$$

Then, using Equations (3.4) *and* (3.7) *we get*

$$\begin{aligned} MPRTE^{*}(x) &= \int_{\mathcal{Z}} MTE^{*}(P^{*}(x,z),x;\delta_x)\frac{\partial P^{*}(x,z)}{\partial z}\left(E\left[\frac{\partial[P^{*}(x,Z)]}{\partial z}\right]\right)^{-1} f_{Z|X}(z|x)dz \\ &= \int_{\mathcal{Z}} \frac{1}{1-\delta_x} MTE(P(x,z),x)\frac{\partial P(x,z)}{\partial z}\left(E\left[\frac{\partial[P(X,Z)]}{\partial z}\right]\right)^{-1} f_{Z|X}(z|x)dz \\ &= \frac{1}{1-\delta_x} MPRTE(x). \end{aligned}$$

Thus, again, by Proposition 70, we obtain

$$(\overline{p_x^{*}} - \underline{p_x^{*}})MPRTE^{*}(x) = MPRTE(x).$$

In the previous examples, proceeding as if there were no misspecification, yields biased parameters. Thus, the automatic "de-biasing" in CATE is the exception rather than the rule.

3.4 Bounds under limited support

Instead of assuming full support, now we allow for limited support of the propensity score $P(x,Z)$, but we still require that it is an interval.

Assumption 13. ***Limited Support.*** *The support of* $P(x,Z)$ *is* $\mathcal{P}_x = [\underline{p_x}, \overline{p_x}] \subset [0,1]$.

Under Assumption 13, and using (3.2), we have that the observed support of $P^{*}(X,Z)$ is

$$[\underline{p_x^{*}}, \overline{p_x^{*}}] = [(1-\delta_x)\underline{p_x} + \delta_x \tilde{P}(x), (1-\delta_x)\overline{p_x} + \delta_x \tilde{P}(x)].$$

Taking the difference we obtain that $\overline{p_x^{*}} - \underline{p_x^{*}} = (1-\delta_x)(\overline{p_x} - \underline{p_x})$. Since $\overline{p_x} - \underline{p_x} \leq 1$, then $\overline{p_x^{*}} - \underline{p_x^{*}} \leq (1-\delta_x)$, so that a lower bound for δ_x is $\delta_x \geq 1 - (\overline{p_x^{*}} - \underline{p_x^{*}})$.

In general, it is not possible to provide an upper bound for δ_x. This is similar to the case of misclassification. Following that literature (see Assumption 4 in Acerenza et al. [2021], and references therein), we assume it is known that for some $\overline{\delta}_x$: $\delta_x \leq \overline{\delta}_x < 1$. Thus, we can write $1-(\overline{p_x^*}-\underline{p_x^*}) \leq \delta_x \leq \overline{\delta}_x$. The correction factor in Examples 72 and 73 is $(1-\delta_x)$. Now, it bounded by $1-\overline{\delta}_x \leq 1-\delta_x \leq \overline{p_x^*}-\underline{p_x^*}$. Thus, we can bound both LATE and MPRTE using this:

$$\begin{aligned}(1-\overline{\delta})\text{LATE}^*(x,P^*(x,z),P^*(x,z')) &\leq \text{LATE}(x,P(x,z),P(x,z')) \\ &\leq (\overline{p_x^*}-\underline{p_x^*})\text{LATE}^*(x,P^*(x,z),P^*(x,z')),\end{aligned}$$

and

$$(1-\overline{\delta})\text{MPRTE}^*(x) \leq \text{MPRTE}(x) \leq (\overline{p_x^*}-\underline{p_x^*})\text{MPRTE}^*(x).$$

Naturally, if $\overline{\delta}_x$ is not known, we can only provide upper bounds.

Again, we stress that is not necessary to bound the MTE curve in the first place. Such a bound can be complicated to obtain since, by Lemma 66, δ_x enters in three different ways in the observed MTE curve.

3.5 Misspecification as a weak instrument

We can frame our model as the triangular scheme of Staiger and Stock [1997] and consider a sequence $\{\delta_{x,n}\}_{n=1}^{\infty}$ such that $\lim_{n\to\infty}\delta_{x,n}=1$ at a certain rate as $n\to\infty$. Thus, as $n\to\infty$, the instrument becomes irrelevant in the model. A possible indicator of the presence of a large value of $\delta_{x,n}$ can be the average derivative of the observed propensity score. This equals an attenuated version of the average derivative of the true propensity

score. For a given value of $\delta_{x,n}$, by equation (3.4), we have

$$E\left[\frac{\partial P^*(x,Z)}{\partial z}\right] = (1-\delta_{x,n})E\left[\frac{\partial P(x,Z)}{\partial z}\right]$$

Thus a "small" value can be an indication that $\delta_{x,n}$ is close to 1. This is similar to a first stage regression in the linear model. We take the derivative with respect to z to get rid of the propensity score that does not respond to Z. We average, because this likely to be a non-linear expression. Thus, $(1-\delta_{x,n})$ can be thought of as the counterpart of $C/\sqrt{T}$ in the notation of Staiger and Stock [1997]. Indeed, define

$$Cov_x(Z, D^*) := E[ZD^*|X=x] - E[Z|X=x]E[D^*|X=x].$$

We have

$$\begin{aligned} E[ZD^*|X=x] &= E[ZSD|X=x] + E[Z(1-S)\tilde{D}|X=x] \\ &= E[ZSD|X=x] + E[Z|X=x]E[(1-S)\tilde{D}|X=x] \end{aligned}$$

and

$$E[D^*|X=x] = E[SD|X=x] + E[(1-S)\tilde{D}|X=x]$$

Thus,

$$\begin{aligned} Cov_x(Z, D^*) &= E[ZSD|X=x] - E[Z|X=x]E[SD|X=x] \\ &\quad + E[Z|X=x]E[(1-S)\tilde{D}|X=x] - E[Z|X=x]E[(1-S)\tilde{D}|X=x] \\ &= Cov_x(Z, SD) \end{aligned}$$

which is the covariance between the instrument and treatment status for the responders with $X = x$. To see the role of the rate at which $\delta_{x,n}$ converges to 1, suppose for a second that we know the functional form of $P^*(x, Z)$, and we estimate the average derivative using a sample mean:

$$\hat{E}\left[\frac{\partial P^*(x,Z)}{\partial z}\right] = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial P^*(x,Z_i)}{\partial z} = (1-\delta_{x,n})\frac{1}{n}\sum_{i=1}^{n}\frac{\partial P(x,Z_i)}{\partial z}$$

Then

$$\hat{E}\left[\frac{\partial P^*(x,Z)}{\partial z}\right] - E\left[\frac{\partial P^*(x,Z)}{\partial z}\right] = (1-\delta_{x,n})\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial P(x,Z_i)}{\partial z} - E\left[\frac{\partial P(x,Z)}{\partial z}\right]\right)$$

In order to investigate possible discontinuities in the limiting distributions, we follow Hahn and Kuersteiner [2002], and we let $(1-\delta_{x,n}) = n^{\nu_x}$, for $\nu_x < 0$. We obtain

$$\hat{E}\left[\frac{\partial P^*(X,Z)}{\partial z}\right] - E\left[\frac{\partial P^*(X,Z)}{\partial z}\right] = O_p(n^{\nu_x - 1/2}).$$

Then, we obtain a degenerate limit:

$$\sqrt{n}\left(\hat{E}\left[\frac{\partial P^*(X,Z)}{\partial z}\right] - E\left[\frac{\partial P^*(X,Z)}{\partial z}\right]\right) = o_p(1)$$

Now consider the MPRTE. Recall that, by Example 73, under the full support guaranteed by Assumption 12,

$$n^{\nu_x}\text{MPRTE}^*(x) = \text{MPRTE}(x).$$

Assume that, if $\delta_x = 0$, there exists $\widehat{\text{MPRTE}}(x)$, a $\sqrt{n}$-consistent estimator of $\text{MPRTE}(x)$

such that

$$\hat{\text{MPRTE}}^{*}(x) - \text{MPRTE}^{*}(x) = n^{-\nu_x}\left(\hat{\text{MPRTE}}(x) - \text{MPRTE}(x)\right).$$

Thus, if $\nu_x = -1/2$, then $\hat{\text{MPRTE}}^{*}(x)$ does not converge in probability. In future work, we will use these results to construct confidence intervals for the parameters of interest.

3.6 Simulations

Consider a linear model for the potential outcomes:

$$Y(0) = \beta_0 X + U_0,$$
$$Y(1) = \beta_1 X + U_1.$$

The selection equations are

$$D = \mathbb{1}\{X + Z \geq V\},$$
$$\tilde{D} = \mathbb{1}\left\{X \geq \tilde{V}\right\}.$$

To carry out the simulations, we assume that the vector $(U_0, U_1, V, \tilde{V})'$ is jointly normal with zero mean and variance-covariance matrix:

$$\Sigma = \begin{bmatrix} \sigma_{U_0}^2 & \sigma_{U_0,U_1} & \sigma_{U_0,V} & \sigma_{U_0,\tilde{V}} \\ \sigma_{U_0,U_1} & \sigma_{U_1}^2 & \sigma_{U_1,V} & \sigma_{U_1,\tilde{V}} \\ \sigma_{U_0,V} & \sigma_{U_1,V} & \sigma_{V}^2 & \sigma_{\tilde{V},V} \\ \sigma_{U_0,\tilde{V}} & \sigma_{U_1,\tilde{V}} & \sigma_{\tilde{V},V} & \sigma_{\tilde{V}}^2 \end{bmatrix}$$

Likewise, X and Z are jointly (bivariate) normal with zero mean and variance-covariance matrix:

$$\Xi = \begin{bmatrix} \sigma_X^2 & \sigma_{X,Z} \\ \sigma_{X,Z} & \sigma_Z^2 \end{bmatrix}$$

Finally, for the latent type, we consider: $S = \mathbb{1}\{X \geq \xi\}$, where ξ is normal with mean zero a variance σ_ξ^2, correlated with V, but independent of X and Z. The MTE curve for responders is

$$\begin{aligned} \text{MTE}(u,x) :&= \mathbb{E}[Y(1) - Y(0) | S = 1, U_D = u, X = x] \\ &= (\beta_1 - \beta_0)x + \mathbb{E}[U_1 - U_0 | S = 1, U_D = u, X = x] \end{aligned}$$

where we have used the fact that $X \perp U_0, U_1, V, \xi$. Since it is very hard to obtain a close form expression for $\mathbb{E}[U_1 - U_0 | x \geq \xi, F_V(V) = u]$ we an infeasible non-parametric estimator based on draws of (U_0, U_1, ξ, V). The estimator is based on:

$$\hat{\mathbb{E}}[U_1 - U_0 | S = 1, U_D = u, X = x] = \frac{\sum_{i:S_i=1} K_h(X_i - x) K_h(U_{Di} - u)(U_{1i} - U_{0i})}{\sum_{i:S_i=1} K_h(X_i - x) K_h(U_i - u)}$$

where $K_h(u) = 1/hK(u/h)$ for a given kernel K and bandwidth h.

3.7 Conclusion

In this paper we use the MTE framework to model a proportion of individuals who do not respond to the incentives of the instrumental variable. We show that in the special case where the observed propensity score is fully supported on the unit interval, i) the CATE is automatically identified regardless of the non-responders, and ii) we can identify the proportion of non-responders and use it to recover the MTE curve, and we can recover

any parameter associated with it. We show that for some parameters, such as LATE and MPRTE, it is even possible to bypass the recovery of the MTE curve, and directly recover these parameters. Moreover, if the propensity has limited support, we find bounds for the LATE, the MPRTE, and the MTE curve. When we let the proportion of non-responders approach 1 at a certain rate, the framework resembles that of weak instruments. In future research we hope to leverage the results in this literature to construct valid confidence intervals for the MTE curve and related parameters.

3.8 Acknowledgements

Chapter 3, in full, is currently being prepared for submission for publication of the material. The material in Chapter 3 is co-authored with Julian Martinez-Iriarte. The dissertation author was a primary author of the material.

Bibliography

Santiago Acerenza, Kyunghoon Ban, and Désiré Kedagni. Marginal treatment effects with misclassified treatment. Working Paper, 2021.

Joseph G Altonji, Todd E Elder, and Christopher R Taber. Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of political economy*, 113(1):151–184, 2005.

Isaiah Andrews, Matthew Gentzkow, and Jesse M Shapiro. Measuring the sensitivity of parameter estimates to estimation moments. *The Quarterly Journal of Economics*, 132 (4):1553–1592, 2017.

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

Bertille Antoine and Prosper Dovonon. Robust estimation with exponentially tilted hellinger distance. *Journal of Econometrics*, 2020.

Timothy B Armstrong and Michal Kolesár. Sensitivity analysis using approximate moment condition models. *Quantitative Economics*, 12(1):77–108, 2021.

Susan Athey, Guido W Imbens, and Stefan Wager. Approximate residual balancing: De-biased inference of average treatment effects in high dimensions. *arXiv preprint arXiv:1604.07125*, 2016.

Javier Baez, German Caruso, Valerie Mueller, and Chiyu Niu. Heat exposure and youth migration in central america and the caribbean. *American Economic Review*, 107(5): 446–50, 2017.

Marc F. Bellamare and Jeffrey R. Bloem. The paper of how: Estimating treatment effects using the front-door criterion. Technical report, 2019. Working paper.

Anders Bjorklund and Robert Moffitt. The Estimation of Wage Gains and Welfare Gains in Self-Selection. *The Review of Economics and Statistics*, 69(1):42–49, 1987.

Stéphane Bonhomme and Martin Weidner. Minimizing sensitivity to model misspecification. *arXiv preprint arXiv:1807.02161*, 2018.

Joseph Briggs, Andrew Caplin, Søren Leth-Petersen, Christopher Tonetti, and Gianluca Violante. Estimating marginal treatment effects with survey instruments. 2020. Working Paper.

Tamara Broderick, Ryan Giordano, and Rachael Meager. An automatic finite-sample robustness metric: Can dropping a little data change conclusions? *arXiv preprint arXiv:2011.14999*, 2020.

Jing Cai. The impact of insurance provision on household production and financial decisions. *American Economic Journal: Economic Policy*, 8(2):44–88, 2016.

Zongwu Cai, Ying Fang, Ming Lin, and Shengfang Tang. Testing unconfoundedness assumption using auxiliary variables. Technical report, University of Kansas, Department of Economics, 2019.

Nancy Cartwright and Jeremy Hardie. *Evidence-based policy: A practical guide to doing it better*. Oxford University Press, 2012.

Matias D Cattaneo, Michael Jansson, and Xinwei Ma. Simple local polynomial density estimators. *arXiv preprint arXiv:1811.11512*, 2018.

Bryant Chen and Judea Pearl. Exogeneity and robustness. Technical report, Tech. Rep, 2015.

Xiaohong Chen. Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics*, 6:5549–5632, 2007.

Victor Chernozhukov, Sokbae Lee, and Adam M Rosen. Intersection bounds: estimation and inference. *Econometrica*, 81(2):667–737, 2013.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21(1):C1–C68, 2018.

Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K. Newey, and James M. Robins. Locally robust semiparametric estimation, 2020.

Denis Chetverikov, Zhipeng Liao, and Victor Chernozhukov. On cross-validated lasso in high dimensions. *The Annals of Statistics*, 49(3):1300–1317, 2021.

Timothy Christensen and Benjamin Connault. Counterfactual sensitivity and robustness. *arXiv preprint arXiv:1904.00989*, 2019.

Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1): 39–67, 2020.

Carlos Cinelli, Andrew Forney, and Judea Pearl. A crash course in good and bad controls. *Available at SSRN 3689437*, 2020.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Imre Csiszár. Sanov property, generalized i-projection and a conditional limit theorem. *The Annals of Probability*, pages 768–793, 1984.

Robert M De Jong. A note on "convergence rates and asymptotic normality for series estimators": uniform convergence rates. *Journal of Econometrics*, 111(1):1–9, 2002.

Xavier de Luna and Per Johansson. Testing for the unconfoundedness assumption using an instrumental assumption. *Journal of Causal Inference*, 2(2):187–199, 2014.

Angus Deaton. Instruments, randomization, and learning about development. *Journal of economic literature*, 48(2):424–55, 2010.

Stephen G Donald, Yu-Chin Hsu, and Robert P Lieli. Testing the unconfoundedness assumption via inverse probability weighted estimators of (l) att. *Journal of Business & Economic Statistics*, 32(3):395–415, 2014.

Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time, i. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

A Finkelstein. Oregon health insurance experiment public use data, 2013.

Amy Finkelstein, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P Newhouse, Heidi Allen, Katherine Baicker, and Oregon Health Study Group. The oregon health insurance experiment: evidence from the first year. *The Quarterly journal of economics*, 127(3):1057–1106, 2012.

Michael Gechter. Generalizing the results from social experiments: Theory and evidence from mexico and india. *manuscript, Pennsylvania State University*, 2015.

Jinyong Hahn and Guido Kuersteiner. Discontinuities of weak instrument limiting distributions. *Economics Letters*, 75:325–331, 2002.

Erin Hartman. Generalizing experimental results. In James Druckman and Donald Green, editors, *Advances in Experimental Political Science*. Cambridge University Press, 2020.

James J. Heckman and Edward Vytlacil. Local Instrumental Variables. In C. Hsiao, K. Morimune, and J. Powell, editors, *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*, pages 1–46. Cambrigde University Press, 2001.

James J. Heckman and Edward Vytlacil. Structural Equations, Treatment Effects, and Econometric Policy Evaluation. *Econometrica*, 73(3):669–738, 2005.

James J. Heckman, Sergio Urzua, and Edward Vytlacil. Understanding instrumental variables in models with essential heterogeneity. *The Review of Economics and Statistics*, 88(3):389–432, 2006.

Paul Ho. Global robust bayesian analysis in large models. 2020.

Joel L Horowitz and Charles F Manski. Identification and robustness with contaminated and corrupted data. *Econometrica: Journal of the Econometric Society*, pages 281–302, 1995.

Yu-Chin Hsu, Tsung-Chih Lai, and Robert P Lieli. Counterfactual treatment effects: Estimation and inference. *Journal of Business & Economic Statistics*, pages 1–16, 2020.

Meng Huang, Yixiao Sun, and Halbert White. A flexible nonparametric test for conditional independence. *Econometric Theory*, 32(6):1434–1482, 2016.

Peter J Huber. A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, pages 1753–1758, 1965.

Guido W. Imbens. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. Technical report, 2019. Working paper.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Sookyo Jeong and Hongseok Namkoong. Robust causal inference under covariate shift via worst-case subpopulation treatment effects. In *Conference on Learning Theory*, pages 2079–2084. PMLR, 2020.

Edward H Kennedy, Sivaraman Balakrishnan, and Max G'Sell. Sharp instruments for classifying compliers and generalizing causal effects. *Annals of Statistics*, 48(4):2008–2030, 2020.

Amanda E Kowalski. Reconciling seemingly contradictory results from the oregon health insurance experiment and the massachusetts health reform. Technical report, National Bureau of Economic Research, 2018.

Lance Lochner and Enrico Moretti. The effect of education on crime: Evidence from prison inmates, arrests, and self-reports. *American economic review*, 94(1):155–189, 2004.

Xun Lu and Halbert White. Robustness checks and robustness tests in applied economics. *Journal of econometrics*, 178:194–206, 2014.

David G Luenberger. *Optimization by vector space methods*. John Wiley & Sons, 1997.

Matthew A Masten and Alexandre Poirier. Inference on breakdown frontiers. *Quantitative Economics*, 11(1):41–111, 2020.

Rachael Meager. Understanding the average impact of microcredit expansions: A bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics*, 11(1):57–91, 2019.

Magne Mogstad and Alexander Torgovitsky. Identification and extrapolation of causal effects with instrumental variables. *Annual Review of Economics*, 10:577–613, 2018.

Karthik Muralidharan and Nishith Prakash. Cycling to school: Increasing secondary school enrollment for girls in india. *American Economic Journal: Applied Economics*, 9(3): 321–50, 2017.

Whitney K Newey. Convergence rates and asymptotic normality for series estimators. *Journal of econometrics*, 79(1):147–168, 1997.

Whitney K Newey and Daniel McFadden. Chapter 36 large sample estimation and hypothesis testing. volume 4 of handbook of econometrics, 1994.

Emily Oster. Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, pages 1–18, 2017.

Emily Oster. Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204, 2019.

Judea Pearl. *Causality: models, reasoning and inference*, volume 29. Springer, 2000.

Edward Posner. Random coding strategies for minimum entropy. *IEEE Transactions on Information Theory*, 21(4):388–391, 1975.

Vitor Possebom. Crime and mismeasured punishment: Marginal treatment effect with misclassification. Working Paper, 2021.

Paul R Rosenbaum and Donald B Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, 79(387):516–524, 1984.

Margot Sanger-Katz. Oregon health study: The surprises in a randomized trial. *The New York Times*, 2014.

Douglas Staiger and James H. Stock. Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557–586, 1997.

John W Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, pages 448–485, 1960.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Martin J Williams. External validity and policy adaptation: From impact evaluation to policy design. *The World Bank Research Observer*, 35(2):158–191, 2020.