Robustness, Heterogeneous Treatment Effects and Covariate Shifts

Pietro Emilio Spini[1]

This draft: July 2024
First draft: May 2021

Abstract

This paper studies the robustness of estimated policy effects to changes in the distribution of covariates, a key exercise to evaluate the external validity of (quasi)-experimental results. I propose a novel scalar robustness metric. It measures the magnitude of the smallest covariate shift needed to invalidate a claim on the policy effect (say, $ATE \geq 0$) supported by the (quasi)-experimental evidence. I estimate my robustness metric using de-biased GMM, which guarantees a parametric convergence rate while allowing for machine learning-based estimators of policy effect heterogeneity (including LASSO, random forest, boosting, neural nets). I apply my procedure to study the robustness of policy effects' estimates for health-care utilization and financial strain outcomes in the Oregon Health Insurance experiment. I find that, among all outcomes, the effect of the insurance policy on outpatient visits is the most robust to shifts in the distribution of context-specific covariates.

Keywords: Robustness, Heterogeneous Treatment Effects, KL divergence, Semiparametric estimation, De-biased GMM, Oregon Health Insurance Experiment

JEL codes: C14, C18, C44, C51, C54, D81, I13

[1]Email: pietro.spini@bristol.ac.uk. University of Bristol, 12 Priory Road, BS8 1TU, UK.
I thank Yixiao Sun, Kaspar Wuthrich, James Hamilton, Sukjin Han, Sami Stouli, Xavier D'Haultfoeuille, Matt Masten, Adam Rosen, Ashesh Rambachan, Davide Viviano, Michael Pollmann and Kirill Ponomarev for their helpful comments. Participants at EGSC 2021, EWMES 2021, the Microeconometrics Class of 2022-2023 Conference at Duke and seminar participants at PSE-CREST, University of Exeter, the University of Warwick, UvA, Erasmus University Rotterdam, University of Bristol, University of Surrey, NYUAD, UCSD, University of Victoria, and the Philadelphia Federal Reserve provided valuable discussion. All remaining errors are mine.

1 Introduction

Evidence-based policy-making uses experimental and quasi-experimental studies to guide the adoption of policies in various settings. This approach relies on the premise that (quasi)-experimental findings are robust and generalizable beyond the original experimental setting. However, in practice, this is not always the case: There are several examples of policies that, when implemented in non-experimental settings, fell short of their own experimental estimates Deaton [2010], Cartwright and Hardie [2012], Williams [2020]. Researchers and policy-makers may want to complement their estimates with a tool that quantifies the robustness of their findings for policy adoption beyond the experimental setting.

In this paper, I develop a new robustness metric, given by a scalar δ^*, that quantifies how much the characteristics of the policy recipients would need 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, that routinely accompany (quasi)-experimental estimates, and can be appended to them in the same manner.

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 for recipients 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

[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], Oster [2019], Cinelli and Hazlett [2020] for confounding factors, and the break-down approaches in Horowitz and Manski [1995], Masten and Poirier [2020].

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 context prior to implementation. As a result, the procedures proposed by Hsu et al. [2020] and Hartman [2020] that re-weight sub-population effects by the new environment's entire set of covariates are generally not feasible because such covariates are missing. The 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 is not consistent with 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 the difference in their means. With an unrestricted form for the heterogeneity of policy effects, we could, 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, even though the distributional shift would change the experimental findings. This example suggests that a robustness

metric should be general enough to accommodate flexible forms of policy effect heterogeneity, whose functional form is, *ex-ante*, unknown. My robustness metric allows for arbitrary forms of observable 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.

There is a natural connection between the covariate robustness exercise in this paper and the literature on Partial Policy Effects. For example, Rothe [2012] considers the effect of a marginal or infra-marginal perturbation of the covariate distribution along a fixed direction on a functional of the unconditional outcome distribution. In contrast, in this paper, the direction of the perturbation is not specified ex-ante, in fact it maybe itself the object of interest as it represents, among all possible shifts that invalidate the policy-maker's conclusion, the hardest one to detect. This distinction reflects the different purpose and hence the complementary of the two approaches. On the one hand, a specific candidate for the covariate distribution is most useful for decomposition exercises that highlight the contribution of several variables on the features of the unconditional distribution as highlighted in the application in Rothe [2012]. On the other hand, searching within a large space of covariate distributions reflects the robustness exercise that is useful for the policy-maker when evaluating the experimental evidence for policy adoption.

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 [2023] who apply it to models defined by moment inequalities, Duchi and Namkoong [2021] for distributionally robust stochastic optimization, and Ho [2023] 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 [Qiao and Minematsu, 2010]. 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 can often be sparse: out of the rich set of covariates available in the (quasi)-experiment, just few are needed to approximate the observable variation of the policy effect. 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 metric for the policy effects on outcomes capturing recipients' heath-care utilization and financial strain. As discussed in Finkelstein et al. [2012] and Finkelstein [2013], the recipients of the Oregon lottery are predominantly older, have poorer health, and include a greater percentage of white individuals than the national average. These demographic features invite questions about the robustness of the Oregon experiment's outcomes, especially if they are used to shape policies 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 [2023]. My robustness exercise is complementary to Kowalski [2023]: I compute the smallest distributional change in some important covariates relative to the Oregon benchmark, that can eliminate the

positive effect of the lottery on recipients' health-care utilization and financial strain. 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], Gechter [2024], 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 [2023], Armstrong and Kolesár [2021], Bonhomme and Weidner [2018], and Antoine and Dovonon [2020], Adjaho and Christensen [2022] in the context of optimal policy choice. My paper complements this tool-set by giving the policy-maker an explicit measure of robustness of a policy claim 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 2 introduces the basic setting and the notion of robustness to changes in the covariate distribution. Section 3 presents the main estimator and its asymptotic properties using the de-biased GMM theory recently developed in Chernozhukov et al. [2020]. Section 4 applies the proposed robustness metric to the Oregon health insurance experiment and reports empirical findings. Section 5 briefly concludes. The main proofs are in the Appendix. The Supplementary Appendix gathers discussion and extensions and additional results.

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. For simplicity, 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 distribution of the 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 ($ATE \geq \tilde{\tau}$), the one that is closest to the distribution of covariates in the (quasi)-experiment. I label this distribution the *least favorable distribution* because it is the hardest to distinguish from the one in the (quasi)-experiment. To measure the distance between two covariate distributions I use the Kullback-Leibler (KL) divergence. The value of the KL divergence 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 ($ATE \geq \tilde{\tau}$).

2.1 Notation and Set Up

In the (quasi)-experiment, 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 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 covari-

ates. F_X denotes 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 $\pi(x) = P(D = 1|X = x)$. Following the traditional potential outcome framework, Y_d for $d = \{0, 1\}$, denotes 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, calligraphic $\mathcal{W}$ 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 and 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 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. Although strict overlap could be slightly weakened while preserving identification, this version is important for the estimation of the robustness metric in Section 3.

Here we are interested in the robustness of claims concerning the ATE with respect to changes in the distribution of the covariates. Because the ATE is obtained by

[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 sketch the approach in the Appendix but leave the details for future research.

averaging $\tau(x)$ with weights proportional to F_X we have the following map between the covariate distributions and the ATE:

$$ATE : F_X \mapsto \int_{\mathcal{X}} \tau_{F_X}(x) dF_X(x) \tag{1}$$

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 from F_X to F'_X would effect the magnitude of ATE through two channels: a direct effect thorough the weights of $F_X(x)$ and an indirect effect through changing the shape of the function $x \mapsto \tau(x)$. Of course, without further assumptions, $\tau_{F_X}(x)$ is only identified when F_X is the experimental distribution. 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 i) as analogous to a policy invariance condition 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, albeit with a different weight. 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 $\tau(x)$:

$$ATE : F_X \mapsto \int_{\mathcal{X}} \tau(x) dF_X(x) \tag{2}$$

Before presenting the general framework I give perhaps the simplest nontrivial exam-

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

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

ple 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 by their observed counterparts $\mathbb{E}[Y|D = 1, X = 0], \mathbb{E}[Y|D = 0, X = 0], \mathbb{E}[Y|D = 1, X = 1], \mathbb{E}[Y|D = 0, X = 1]$. Consequently, the average treatment effect for the sub-populations $X = 0$ and $X = 1$, $\tau(0) = \mathbb{E}[Y_1 - Y_0|X = 0]$ and $\tau(1) = \mathbb{E}[Y_1 - Y_0|X = 1]$, are also identified. Because X is Bernoulli, any distribution on $\{0, 1\}$ is fully described by $P_X(X = 1) = p_1$. Suppose that, in the experiment $ATE > 0$. 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.

In the general case, under what conditions are we 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 $\mathcal{X}$. Moreover, 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]

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

$$D_{KL}(F'_X||F_X) := \int_{\mathcal{X}} \log\left(\frac{dF'_X}{dF_X}(x)\right) \frac{dF'_X}{dF_X}(x) dF_X(x) \tag{3}$$

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 [2023] and Christensen and Connault [2023] use the KL divergence to measure the distance between probability distributions in different contexts.

2.2 The policy-maker's problem: quantifying robustness

After highlighting 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 the value of ATE for a competing policy. In Example 1, $\tilde{\tau}$ was fixed at 0, which is a natural benchmark if we are interested in a positive average effect. 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.

[5] In Appendix B, I discuss how the general procedure in this paper can be specialized to certain parametric classes of distributions. In such cases, the relevant covariate shifts coincide with mean shifts.

Formally the policy-maker wants to solve the following problem:

$$\inf_{F'_X:\ P'_X \ll P_X; P'_X(\mathcal{X})=1} D_{KL}(F'_X||F_X) \tag{4}$$

$$s.t.\ \int_{\mathcal{X}} \tau(x) dF'_X(x) \leq \tilde{\tau} \tag{5}$$

The optimization problem in Equation (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 (5) is 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 (4). Notice also that $\tau(x)$ in Equation (5) is not indexed by F'_X because of the covariate shift assumption (Assumption 2).

Remark 3. *Here, the class of probability measures for the covariates is restricted to be absolutely continuous*[6] *w.r.t to F_X but no other restriction is imposed: the class of distributions is nonparametric. Absolute continuity does restrict the distributions F'_X to be supported on $\mathcal{X}$. I view it as a reasonable requirement: the feasible distributions in Equation (5) cannot put mass on a sub-population $X = x$ that could not theoretically be observed in the (quasi)-experimental setting. The support assumption also appears in the distributional policy effect literature, see Rothe [2012], playing a similar role. Clearly, treatment effect values for sub-populations with $X = x$ that can never be observed are not identified. Absent other restrictions, they can lead to arbitrarily large average effects and making the robustness exercise not very informative.*

We are now ready to define the *least favorable distribution* and the robustness metric.

Definition 4. *i) For a given $\tilde{\tau} \in \mathbb{R}$, the the robustness metric $\delta^*(\tilde{\tau})$ is the infimum of Equation (4). ii) The (set of) least favorable distribution(s) $\{F^*_X\}$ is the (set of) minimizer(s) of Equation (4).*

I define my metric $\delta^*(\tilde{\tau})$ as the KL-distance between the experimental distribution and the *least favorable distribution*. It quantifies the robustness of the claim $ATE \geq \tilde{\tau}$. $\{F^*_X\}$ contains the closest distribution(s) of the covariates that invalidates the target claim. For certain values of $\tilde{\tau}$ it may be empty.

Observe that, if the (quasi)-experimental ATE satisfies the constraint in Equation (5), then we can always choose the *least favorable distribution* to be the (quasi)-

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

experimental one, namely $F_X^* = F_X$ since it's feasible and $D_{KL}(F_X^*||F_X) = 0$. This means that the policy-maker's claim is already invalidated in the (quasi)-experiment. The problem is non-trivial when the quasi-experimental distribution F_X satisfies $ATE(F_X) > \tilde{\tau}$. In such a case, the (quasi)-experimental distribution F_X is excluded from the feasible set of Equation (5). As a result, under some regularity conditions, the value of $D_{KL}(F_X^*||F_X)$ in Equation (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 finite set, the covariate distribution is discrete. There are many empirical applications in which covariates are either discrete or have been discretized. 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 geometric insights from the structure of the robustness problem in Equations (4) and (5), like in the example below:

Example 5. *Consider $\mathcal{X} = \{H, M, L\}$ representing income bins: high, medium and low. A distribution is a triplet (p_H, p_M, p_L). Because $p_H + p_M + p_L = 1$ the whole space of probability distributions on $\mathcal{X}$ is the 2-simplex. Suppose that CATE is decreasing in income levels: $\tau(H) = 1, \tau(M) = 2, \tau(L) = 3$. 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.*

Figure 1 depicts the level sets of the KL distance, the feasible set and the *least favorable distribution* from Example 5. The functions in Equations (4) and (5) are differentiable in p_H and p_M so the solution is characterized by the standard KKT conditions. The KL level set associated to $\delta^*(\tilde{\tau})$ is highlighted in green. The set of triplets (p_H, p_M, p_L) that are within it are guaranteed to satisfy the policy-maker's claim. This region is conservative: 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 (4) and (5).

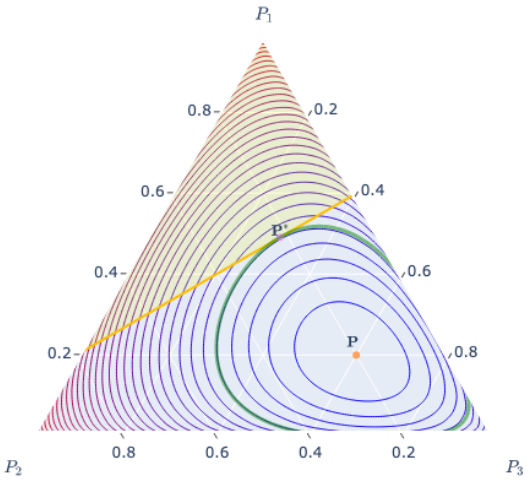

Figure 1: The triangle represents the collection of all arbitrary probability distribution triplets (p_H, p_M, p_L) on the discrete set (H, M, L) represented in barycentric coordinates. P denotes the experimental distribution, given by $(0.2, 0.2, 0.6)$. We have $CATE(H, M, L) = (1, 2, 3)$ so the conditional treatment effect is greater in the highest income group. The yellow shaded region is the feasible set: the collection of triplets (p_H, p_M, p_L) with an $ATE \leq 1.8$, which invalidate the policy-maker's claim. Importantly, the requirement $P'_X \ll P_X$ excludes distributions where one of the entries is equal to 0. The solid yellow line is the boundary of the feasible set. The contour lines 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.

When $\mathcal{X}$ is not discrete, a visualization like the one in Figure 1 may not be possible. Nonetheless, given some regularity conditions, a solution for F_X^* like the one in Figure 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.

2.3 A closed form solution for quantifying robustness

I characterize the solution for the policy-maker's robustness problem in Equations (4) and (5) in the general case under some regularity conditions.

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 $\mathcal{X}$ where $\tau(x)$ is unbounded. In principle, one can weaken Assumption 3 to only require $\mathbb{E}[\exp(\kappa|\tau(X)|)] < +\infty, \forall \kappa > 0$[7]. Bounded-ness as in Assumption 3 is not very restrictive in a micro-econometrics framework where in many contexts, all variables are bounded in the cross-section. I thus maintain Assumption 3 in its form to ease the exposition.

Consider the feasible set in Equation (5). While the set, being a half space, is guaranteed to be convex, it may still be empty. If that is the case, the value of the minimization problem in Equation (4) is $+\infty$. I avoid such cases by requiring 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. To appreciate this requirement, it is instructive to see an extreme case where such requirement fails.

Example 6 (Homogeneous treatment effects)**.** *Consider a situation of constant treatment effects. In this case $\tau(x) = c$ so $ATE(F_X) = \int_{\mathcal{X}} c \, dF_X = c$ so that the ATE is equal to c regardless of the distribution of the covariates.*

Not surprisingly, if there is no heterogeneity in treatment effects, under Assumption 2 one can freely extrapolate the claim from the (quasi)-experimental distribution to any other distribution. Constant treatment effects are a rather extreme case. A more realistic example is when the minimal desired magnitude $\tilde{\tau}$ is outside of the range of variation of the heterogeneous 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, since no probability distribution may ever integrate against $\tau(x)$ to an

[7]Bounded-ness and its weaker version correspond to A3 and A5 respectively in Komunjer and Ragusa [2016]. No weaker condition on $\tau(X)$ can be obtained, for example along the lines of Komunjer and Ragusa [2016] A7, because the KL divergence does not satisfy the minimal growth requirements (in particular, for the KL divergence, the Orlicz heart is strictly smaller than the Orlicz space).

ATE of 1. In this case, since the set of distributions in Equation (5) is empty, the infimum in Equation (4) evaluates to $+\infty$. So we see that enough heterogeneity of treatment effects is a necessary condition for robustness to be non-trivial.[8] 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 there is enough observable heterogeneity in treatment effects to find a distribution of covariates that, when integrated against $\tau(x)$, could induce $ATE = \tilde{\tau}$. Contrast this to the homogeneous treatment effect case in Example 6, 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 of $\tilde{\tau}$, one could obtain an estimate of $\tau(x)$ 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]. Homogeneous treatment effects are a very special case in which such a test might reject. More broadly, a violation of assumption 4 implies the robustness metric is infinite, indicating the policy-maker's claim can't be invalidated by covariate shifts.

Assumption 4 is also related to a constraint qualification, akin to a Slater condition, that appears often in convex problems involving KL and other φ-divergences. For an excellent overview of these conditions see Komunjer and Ragusa [2016]. In the setting of this paper, Assumption 4 is sufficient to guarantee strong duality and to obtain a characterization of the *least favorable distribution* through convex duality. In particular, Assumption 4 satisfies the quasi-relative interior condition in Borwein and Lewis [1993] (Equation (BL)).

Remark 7. *The interior condition cannot be relaxed. By Assumption 3, the image of $\mathcal{P}_X$ under L is a 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 (5) may consist of only a point mass measure. Because such a covariate measure is not absolutely continuous w.r.t. P_X,*

[8] Moreover, for estimation purposes it is convenient to consider a parameter space for the robustness metric that is a subset of $\mathbb{R}$ rather than $\mathbb{R} \cup \{+\infty\}$.

the feasible set is again empty and will necessarily result in an infinite value for the KL-divergence in Equation (4). I provide some additional results about this boundary cases in Appendix 3.

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 $ATE(p_1')$ 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 state the key result that delivers a closed form solution for the robustness metric in the general case. It says that the *least favorable distribution* set in Definition 4 ii) is non-empty and 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:

Theorem 8 (Closed form solution)**.** *Let Assumptions 1, 2, 3 and 4 hold. Then: i) The infimum in Equation (4) is achieved. Moreover F_X^*, is characterized, P_X-almost everywhere, by:*

$$\frac{dF_X^*}{dF_X}(x) = \frac{\exp(-\lambda(\tau(x) - \tilde{\tau}))}{\int_{\mathcal{X}} \exp(-\lambda(\tau(x) - \tilde{\tau}))dF_X(x)} \tag{6}$$

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:

$$\int_{\mathcal{X}} \exp(-\lambda(\tau(x) - \tilde{\tau}))(\tau(x) - \tilde{\tau})dF_X(x) = 0 \tag{7}$$

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

$$\delta^*(\tilde{\tau}) = D_{KL}(F_X^* || F_X) = -\log\left(\int_{\mathcal{X}} \exp(-\lambda(\tau(x) - \tilde{\tau}))dF_X(x)\right) \tag{8}$$

Theorem 8 greatly simplifies the computation of the robustness metric. It shows that the fully general robustness problem that searches over the nonparametric space of probability distribution is not substantially harder than the parametric cases in Examples 1 and 5. We can compare the closed form solution of Theorem 8 with the KKT solution one could derive for Example 1 and verify that the two solutions are indeed identical.

Example 9. *Return to the example of the discrete variable so* $X = \{0, 1\}$. *First notice that the dominating measure here is the counting measure on* $\{0, 1\}$. *The ratio* $\frac{p_1^*}{p_1}$ *completely characterizes the solution. Because the problem is one dimensional, the unique minimizer is the one that satisfies the constraint:*

$$\tau(1) \cdot p_1^* + \tau(0) \cdot (1 - p_1^*) = \tilde{\tau} \implies p_1^* = \frac{\tilde{\tau} - \tau(0)}{\tau(1) - \tau(0)} \tag{9}$$

Recall that in Example 1 $\tilde{\tau} = 0$. *On the other hand, from the solution provided by Theorem 8 we have:*

$$\frac{p_1^*}{p_1} = \frac{\exp(-\lambda(\tau(1) - \tilde{\tau}))}{\exp(-\lambda(\tau(1) - \tilde{\tau})) \cdot p_1 + \exp(-\lambda(\tau(0) - \tilde{\tau})) \cdot (1 - p_1)} \tag{10}$$

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

Fact 10. *Equations 10 and 9 are equivalent.*

In terms of identification, Theorem 8 completely characterizes the robustness metric in terms of the (quasi)-experimental distribution F_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 also simplify the estimation strategy, as I discuss in Section 3.

3 Estimation and Asymptotic Results

In this section I introduce a semi-parametric estimator for the robustness metric δ^* in Equation (8) and I characterize its asymptotic properties. I show that the robustness metric can be estimated using a GMM setting 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 $\tau(x)$.

3.1 An empirical estimate of the robustness metric

The closed form solution in Theorem 8 suggests a natural estimator based on empirical averages. In particular, one would like to replace Equation (8) 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 conditions is denoted by $\theta_0 = (\nu_0, \lambda_0)^T$. Then, the robustness metric is given by $\delta^* = -\log(\nu_0)$. The parameter space Θ such that $\theta \in \Theta \subseteq \mathbb{R}^2$ satisfies some constraints. First, observe that if the policy-maker's claim ($ATE \geq \tilde{\tau}$) holds with a strict inequality for F_X, then $\delta^* > 0$. This implies a restriction on $\nu_0 < 1$. Moreover, $\nu_0 > 0$ because $\exp(-\lambda(\tau(x) - \tilde{\tau}) > 0$ 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 for (ν_0, λ_0) as:

$$\mathbb{E}[g(W, \theta, \tau)] = \mathbb{E}\begin{bmatrix} \exp(-\lambda_0(\tau_0(X) - \tilde{\tau})) - \nu_0 \\ \exp(-\lambda_0(\tau_0(X) - \tilde{\tau}))(\tau_0(X) - \tilde{\tau}) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{11}$$

where $\tau_0(X)$ denotes the true value of CATE. Assumptions 1–4 guarantee that the parameters of interest (λ_0, ν_0) are (globally) identified by Equation (11). Because the true value for $\tau_0(X)$ is an unknown but estimable population quantity, I consider a feasible version of Equation (11) that uses an estimate $\hat{\tau}(X)$ in place of $\tau_0(X)$. One could define the vector $\hat{\theta} = (\hat{\lambda}, \hat{\nu})^T$ as the approximate solution to the empirical moment:

$$\mathbb{E}_n[g(W, \hat{\theta}, \hat{\tau})] = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n \exp(-\hat{\lambda}(\hat{\tau}(X_i) - \tilde{\tau})) - \hat{\nu} \\ \frac{1}{n}\sum_{i=1}^n \exp(-\hat{\lambda}(\hat{\tau}(X_i) - \tilde{\tau}))(\hat{\tau}(X_i) - \tilde{\tau}) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{12}$$

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 estimation strategies, both parametric and nonparametric. 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 (12) is that the moment conditions are not Neyman orthogonal with respect to $\hat{\tau}(X)$. 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^*. This can lead to incorrect inferences on the robustness metric, see Chernozhukov et al. [2018] for a general discussion. 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.

3.2 De-biased GMM estimator

In this section, I derive the nonparametric correction for the GMM estimator of θ based on Equation (12). I map causal quantities like $\tau(X)$ to the statistical functionals that identify them and explicitly construct the nonparametric influence function for the moment conditions. 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 $\tau_0(X) = \tau_{F_0}(X)$. 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 $\gamma_{0,F}(X) := \mathbb{E}_F[Y|X, D = 0]$. Both can be gathered in the vector $\gamma_F = (\gamma_{0,F}, \gamma_{1,F})^T$.

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. With a slight abuse of notation we can write $\mathbb{E}[g(W, \theta, \gamma_F]$ to mean $\mathbb{E}[g(W, \theta, \gamma_{1,F} - \gamma_{0,F})]$ replacing τ_F with the statistical functional identifying it. In order to de-bias the moment conditions with the approach of Chernozhukov et al. [2020] one needs to compute the nonparametric influence function with respect to γ_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:

$$\left.\frac{d\mathbb{E}[g(W,\theta,\gamma_{F_r})]}{dr}\right|_{r=0} = \int \phi(w,\gamma_{F_0},\theta,\alpha)dH(w) \tag{13}$$

Note that the Riesz represent $\phi(\cdot)$ is allowed to depend on γ_{F_0} plus additional nonparametric components, gathered in $\alpha(\cdot)$. In the next result I derive the nonparametric influence function explicitly from Equation (11).

Proposition 11. *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:

$$\psi(w,\gamma,\theta,\alpha) = g(w,\theta,\gamma) + \phi(w,\theta,\gamma,\alpha) \tag{14}$$

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 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. In Appendix A, I show Monte Carlo simulations that corroborates these results.

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 12. *Equation* (14) *satisfies Neyman orthogonality.*

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 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 (note the indexing $-k$ in the subscripts of $\gamma(\cdot)$ and $\alpha(\cdot)$). Sample splitting reduces own-observation bias and, together with the Neyman 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 preliminary 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 for θ is given by:

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \hat{\psi}(\theta, \hat{\gamma}, \hat{\alpha}) \tag{15}$$

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 4, 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 13 (Asymptotic normality of θ)**.** *Let Assumptions 1–5. For $\hat{\theta}$ defined in Equation (15):*

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

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

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

Corollary 14 (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 13.

With the results of Theorem 13 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 one-sided confidence region for their breakdown frontier rather than a confidence band.

3.3 Reporting features of the *least favorable distribution*

Theorem 8 gives an explicit formula for the *least favorable distribution* F_X^* and shows that it depends on λ_0 and $\tau_0(X)$. The researcher may be interested in F_X^* directly [9]. 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 the estimator of F_X^* can, in general, be nonparametric. This is because, under some conditions, it inherits the nonparametric rate of $\hat{\tau}(X)$. Instead, the researcher could report particular moments of interest to compare the experimental distribution F_X and the *least favorable distribution* F_X^*. Reporting moments of the covariate distribution, particularly averages, is a standard practice. For example, comparing covariate means cross treatment status, like in Rosenbaum and Rubin [1984], may be motivated motivated by an interest in internal validity. By analogy, comparing moments from F_X and F_X^* could be motivated by external validity. For this reason, I provide an estimator for an arbitrary, finite collection of moments of F_X^* that may be of interest. An extension of Theorem 13 shows the asymptotic properties of the joint estimator for θ and the additional parameters, denoted by ζ.

Theorem 15 (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 11 – 25 and $g^u(\cdot)$ and

[9] In section D.1 of the Supplementary Appendix I leverage results in information theory to offer an additional interpretation for F_X^* as a particular conditional distribution.

$\phi^u(\cdot)$, whose values are vectors in $\mathbb{R}^s$ are defined below.

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

$$(\hat{\theta}, \hat{\zeta}) := \arg \min_{(\theta,\zeta) \in \mathbb{R}^{s+2}} \hat{\psi}^u(\theta, \zeta, \hat{\gamma}, \hat{\alpha})^T \hat{\psi}^u(\theta, \zeta, \hat{\gamma}, \hat{\alpha}) + o_P(1) \tag{16}$$

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.

4 Empirical Application

In a seminal study of the Oregon Medicaid expansion lottery, Finkelstein et al. [2012] finds that being offered subsidized health insurance improves recipients' healthcare consumption and financial outcomes. These results are of great interest for policy-makers considering a Medicaid expansion in their state. However, in this situation: (1) the population of recipients from their state could differ from the Oregon recipients (2) along covariates that are predictive of treatment effect heterogeneity, (3) these covariates are not available outside of the Oregon experiment. I propose to complement the findings of Finkelstein et al. [2012] by reporting δ^* to characterize their robustness. This metric quantifies the magnitude of the covariate shift needed to eliminate the positive effects of the health insurance lottery when implemented in a new state.

Between March and September 2008, the state of Oregon conducted a series of lottery draws to grant winners 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]. Finkelstein et al. [2012] discusses a variety of robustness concerns that center on external validity. They note that scaling up the experiment can induce a supply side change in providers' behavior. They also 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 underperforms 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, which is a key feature of policy adoption in other states. I stress that, in this context, the re-weighting procedure in Hartman [2020] or Hsu et al. [2020] is not applicable because the survey-specific health data that are likely to be most predictive of treatment effect heterogeneity are not available for many other states. Instead, I propose to study the robustness of the policy by reporting, for each outcome in Finkelstein et al. [2012], my robustness metric, which can be computed by exploiting the heterogeneity in the publicly available survey data Finkelstein [2013].

4.1 Robustness in the Oregon Medicaid Experiment

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]. 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 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 $\tilde{\tau} = t_j = z_\alpha \cdot \sigma_j$ 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 indistinguishable from a non-positive or non-negative outcome (respectively). 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 other hypotheses may easily adapt the procedure by specifying a different $\tilde{\tau}$.

First, 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 regression of the outcome variable on the lottery indicator and controls (survey waves indicators, household size indicators and interaction terms between the two). 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.[10] 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

[10]With discrete covariates, the CATT can be obtained by a fully saturated regression where the lottery indicator is interacted with all possible combinations of dummies.

above.[11] All outcomes are measured on the survey data Finkelstein [2013].

Columns 2, 3 and 4 of Table 1 contain the experimental ITT for each outcome variable, the estimates for $\delta^*(0)$ and the estimates for $\delta^*(t_j)$. Here $t_j = \pm 1.645 \cdot \sigma_j$ 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 (the lottery decreases the probability of skipping a payment by 5%) 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: δ^* robustness metric for the health-care utilization and financial strain outcomes in Finkelstein et al. [2012]. The measure is evaluated at $\tilde{\tau} = 0$ and $\tilde{\tau} = t_j = \pm 1.645 \cdot \sigma_j$ 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 an ordinal comparison of the robustness across outcomes because each δ^* has the same units

[11] Comparable (survey weighted) ITT estimates can be found in column 2 labelled Reduced form, of Table 1. Discrepancies with the (unweighted) ITT effects I compute are due to survey weights.

and it is measured on the same covariate space. Second, the fourth column of Table 1 has a natural interpretation as a breakdown point: what is the smallest shift 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 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 could 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.

5 Conclusion

I propose a metric δ^* to quantify the robustness of (quasi)-experimental findings with respect to covariate shifts. I focus on claims about aggregate policy effects of the type ($ATE \geq \tilde{\tau}$). While extending this approach to linear policy parameters beyond the ATE is straightforward, addressing non-linear distributional policy parameters poses a challenge due to the absence of a closed-form solution. Estimation and inference in this context requires future research. For ATE, the closed form solution leads naturally to a debiased-GMM approach. It allows CATE to be estimated using a large collection of machine learning techniques which are being increasingly adopted by applied researchers, including LASSO, random forest, boosting, neural nets. As demonstrated in the empirical application, the researcher interested in external validity may append the robustness metric to their effect estimates to inform a discussion of their results for policy adoption.

References

C. Adjaho and T. Christensen. Externally valid treatment choice. *arXiv preprint arXiv:2205.05561*, 1, 2022.

J. G. Altonji, T. E. Elder, and C. R. Taber. Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of political economy*, 113(1):151–184, 2005.

I. Andrews, M. Gentzkow, and J. M. Shapiro. Measuring the sensitivity of parameter estimates to estimation moments. *The Quarterly Journal of Economics*, 132(4): 1553–1592, 2017.

B. Antoine and P. Dovonon. Robust estimation with exponentially tilted hellinger distance. *Journal of Econometrics*, 2020.

T. B. Armstrong and M. Kolesár. Sensitivity analysis using approximate moment condition models. *Quantitative Economics*, 12(1):77–108, 2021.

S. Athey, G. W. Imbens, and S. Wager. Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *arXiv preprint arXiv:1604.07125*, 2016.

S. Bonhomme and M. Weidner. Minimizing sensitivity to model misspecification. *arXiv preprint arXiv:1807.02161*, 2018.

J. M. Borwein and A. S. Lewis. Partially-finite programming in l_1 and the existence of maximum entropy estimates. *SIAM Journal on Optimization*, 3(2):248–267, 1993.

T. Broderick, R. Giordano, and R. Meager. An automatic finite-sample robustness metric: Can dropping a little data change conclusions? *arXiv preprint arXiv:2011.14999*, 2020.

N. Cartwright and J. Hardie. *Evidence-based policy: A practical guide to doing it better*. Oxford University Press, 2012.

V. Chernozhukov, S. Lee, and A. M. Rosen. Intersection bounds: estimation and inference. *Econometrica*, 81(2):667–737, 2013.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21(1):C1–C68, 2018.

V. Chernozhukov, J. C. Escanciano, H. Ichimura, W. K. Newey, and J. M. Robins. Locally robust semiparametric estimation, 2020.

T. Christensen and B. Connault. Counterfactual sensitivity and robustness. *Econometrica*, 91(1):263–298, 2023.

C. Cinelli and C. Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):39–67, 2020.

I. Csiszár. Sanov property, generalized i-projection and a conditional limit theorem. *The Annals of Probability*, pages 768–793, 1984.

A. Deaton. Instruments, randomization, and learning about development. *Journal of economic literature*, 48(2):424–55, 2010.

M. D. Donsker and S. S. Varadhan. Asymptotic evaluation of certain markov process expectations for large time, i. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975.

J. C. Duchi and H. Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

A. Finkelstein. Oregon health insurance experiment public use data, 2013.

A. Finkelstein, S. Taubman, B. Wright, M. Bernstein, J. Gruber, J. P. Newhouse, H. Allen, K. Baicker, and O. H. S. Group. The oregon health insurance experiment: evidence from the first year. *The Quarterly journal of economics*, 127(3):1057–1106, 2012.

M. Gechter. Generalizing the results from social experiments: Theory and evidence from mexico and india. *manuscript, Pennsylvania State University*, 2015.

M. Gechter. Generalizing the results from social experiments: Theory and evidence from india. *Journal of Business & Economic Statistics*, 42(2):801–811, 2024.

E. Hartman. Generalizing experimental results. In J. Druckman and D. Green, editors, *Advances in Experimental Political Science*. Cambridge University Press, 2020.

P. Ho. Global robust bayesian analysis in large models. *Journal of Econometrics*, 235 (2):608–642, 2023.

J. L. Horowitz and C. F. Manski. Identification and robustness with contaminated and corrupted data. *Econometrica: Journal of the Econometric Society*, pages 281–302, 1995.

Y.-C. Hsu, T.-C. Lai, and R. P. Lieli. Counterfactual treatment effects: Estimation and inference. *Journal of Business & Economic Statistics*, pages 1–16, 2020.

P. J. Huber. A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, pages 1753–1758, 1965.

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

S. Jeong and H. Namkoong. Robust causal inference under covariate shift via worst-case subpopulation treatment effects. In *Conference on Learning Theory*, pages 2079–2084. PMLR, 2020.

E. H. Kennedy, S. Balakrishnan, M. G'Sell, et al. Sharp instruments for classifying compliers and generalizing causal effects. *Annals of Statistics*, 48(4):2008–2030, 2020.

I. Komunjer and G. Ragusa. Existence and characterization of conditional density projections. *Econometric Theory*, 32(4):947–987, 2016.

A. E. Kowalski. Reconciling seemingly contradictory results from the oregon health insurance experiment and the massachusetts health reform. *Review of Economics and Statistics*, 105(3):646–664, 2023.

M. A. Masten and A. Poirier. Inference on breakdown frontiers. *Quantitative Economics*, 11(1):41–111, 2020.

R. 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.

W. K. Newey and D. McFadden. Chapter 36 large sample estimation and hypothesis testing. volume 4 of handbook of econometrics, 1994.

E. Oster. Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204, 2019.

Y. Qiao and N. Minematsu. A study on invariance of f-divergence and its application to speech recognition. *IEEE Transactions on Signal Processing*, 58(7):3884–3890, 2010.

P. R. Rosenbaum and D. B. Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, 79(387):516–524, 1984.

C. Rothe. Partial distributional policy effects. *Econometrica*, 80(5):2269–2301, 2012.

M. Sanger-Katz. Oregon health study: The surprises in a randomized trial. *The New York Times*, 2014.

J. W. Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, pages 448–485, 1960.

M. J. Williams. External validity and policy adaptation: From impact evaluation to policy design. *The World Bank Research Observer*, 35(2):158–191, 2020.

A Monte Carlo Simulations

To show how we can use de-biased GMM to estimate the robustness metric, I conduct a 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.[12] To demonstrate that heterogeneity drives robustness, I fix the same baseline ATE for the three DGPs. The shape of $\tau(x)$ is chosen to ensure

[12] We have (U_1, U_0) are uncorrelated normals with $\mu = 0, \sigma = 0.25$.

- DGP1: $Y_1 - Y_0 = \exp(X_1) + U_1 - U_0$;

the same ATE across all three DGPs, irrespective of treatment effect heterogeneity, 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 $\pi_X(x)$. Hyper-parameters are tuned to the sample size through a rule of thumb though in practice one may use *within-fold* cross-validation. 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^*. 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 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 2: Monte Carlo Simulation reports 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 2 report the results. As heterogeneity of $\tau(x)$ in the DGP increases, the population value of $\delta^*(\tilde{\tau})$ decreases. This means that the shift in the covariates required to invalidate the claim ($ATE \geq 1.3$) becomes smaller. 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 $\tau(x)$.

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

In DGP1 the heuristic choice of hyper-parameters for boosting likely results in under-fitting the data, leading to a bias one 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 due 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.

B Parametric CATE and parametric shifts

Theorem 8 gives a general closed form solution to the policy-maker's problem without restricting the functional form of CATE or imposing a parametric family for the experimental distribution F_X. Leveraging this, I show that if $\tau(x)$ is quadratic and the experimental distribution belongs to the normal family, the *least favorable distribution* will also belong to the normal family, up to a shift in the parameters.[13] After giving a definition of a class being closed with respect to taking the least favorable distribution, I state this result in Proposition 17 below.

Definition 16. *We say that a class of parametric distributions indexed by θ, denoted F_X^θ is least-favorable closed with respect to a parametric model for CATE, $\tau_\eta(x)$, indexed by $\eta \in H \subseteq \mathbb{R}^{d_\eta}$ if for any θ and η, the least favorable distribution defined in 4 ii) has the form $F_X^* = F_X^{\theta^*}$ for some $\theta^*(\eta) \in \Theta$, highlighting that θ^* will in general also depend on features of η as well.*

Proposition 17 (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 $X \sim \mathcal{N}(\mu, \Sigma)$ where*

[13] An extension of Proposition 17 could be shown to hold for the more general class of distributions in the exponential family given by $f(x|\theta) = g(\theta)h(x)\exp(\eta(\theta)^T T(x))$.

Σ is p.d. and $\tau(x) = x^T Ax + x^T\beta + c$ for $\beta \in \mathbb{R}^k$ then F_X^* is the measure induced by $X^* \sim \mathcal{N}(\mu^*, \Sigma^*)$ with $\mu^* = (\Sigma^{-1} + 2\lambda A)^{-1}(\Sigma^{-1}\mu - \lambda\beta)$ and $\Sigma^* = (\Sigma^{-1} + 2\lambda A)^{-1}$, provided that $(\Sigma^{-1} + 2\lambda A)^{-1}$ is p.d. The parameter λ is defined as in Equation (7).

Corollary 18 (Linear-Normal least favorable closed-ness)**.** *If $\tau(x) = x^T\beta$ and $X \sim \mathcal{N}(\mu, \Sigma)$ then $X^* \sim \mathcal{N}(\mu^*, \Sigma)$ where $\mu^* = \mu - \lambda\Sigma\beta$.*

Corollary 18 states that in a correctly specified linear model $\tau(x) = x^T\beta$ with joint normal covariates, the nonparametric robustness exercise for arbitrary covariate shifts reduces to examining mean shifts. Proposition 17 shows that my proposed robustness procedure generalizes the heuristic of looking at mean shifts (or providing means as summaries for covariate balance type of exercises) to cases where both $\tau(x)$ and F_X take an arbitrary form. In those cases, looking at mean shifts is not sufficient and may lead to misleading judgement about the robustness of the ATE.

Consider an example where $X \sim \mathcal{N}(\mu, \sigma^2)$. CATE is linear: $\tau(x) = \pi \cdot X$ for some $\pi \in \mathbb{R}$. As a result, ATE is only a function of F_X via the population mean μ. The policy-maker's desired claim is $ATE \geq 0$. The feasible set of distributions in Figure 2 is the half-space $\mu \leq 0$. Proposition 17 tells us that the solution to Equation (4) must be another normal distribution. 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:

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

The KKT conditions imply:

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

The *least favorable distribution* amounts to a mean shift and no change in the variance. Contrast the example above with the case where CATE can be quadratic. Proposition 17 still applies, so the solution must have the form $\mathcal{N}(\mu^*, \sigma^{2*})$. This time though, ATE is a function of both μ and σ^2 as reflected in the feasible set in yellow. The least favorable distribution amounts to a mean and a variance shift.

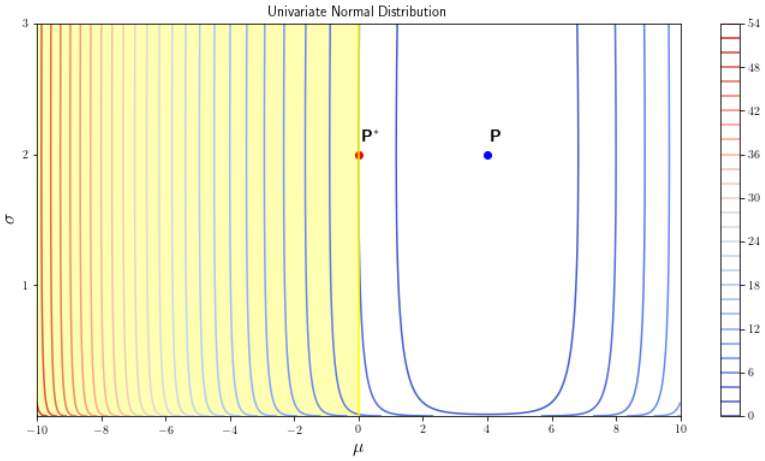

Figure 2: Univariate Normal Distribution, Linear CATE. Each point in the graph represents a normal distribution parametrized by its mean and standard deviation $\mathcal{N}(\mu, \sigma^2)$. The experimental one is $\mathrm{P} = \mathcal{N}(4, 2)$. The contour lines represent the KL divergence with respect to P. 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, ATE depends only on 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^2 parameter.

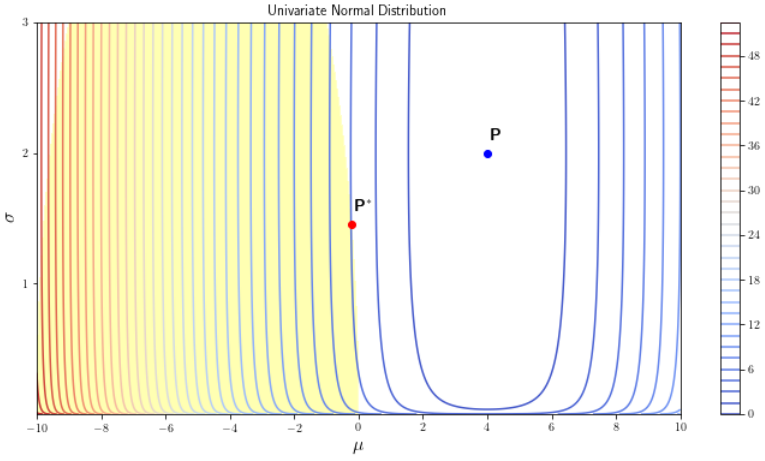

Figure 3: Univariate Normal Distribution: Quadratic CATE. The setting is identical as in Figure 2 but here is quadratic, $\tau(X) = 0.8 \cdot X^2 + 8 \cdot X$. As a result, $ATE(\mu, \sigma) = 0.8 \cdot (\mu^2 + \sigma^2) + 8\mu$: both parameters μ and σ^2 determine the ATE. The feasible set in yellow is no longer parallel to the σ axis. The *least favorable distribution* P^* features a shift in both the mean and the variance.

C Constrained Classes

A researcher may wish to restrict the class of distributions in Equation (5) by imposing additional constraints. For example, matching 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 in Ho [2023]. For a known moment function $q : \mathcal{X} \to \mathbb{R}^L$ we want:

$$\int_{\mathcal{X}} q(X) dF_X = \int_{\mathcal{X}} q(X) dF'_X \tag{17}$$

From the perspective of robustness, the value of δ^* for the constrained problem must be greater than or equal to the value for the unconstrained problem. That is:

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

If the problem contains additional constraints of the form of Equation 17, a simple characterization of the closed form solution (analogous to Theorem 8) holds, and the solution to the KL problem takes the form:

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

where each Lagrange multiplier μ_L satisfies:

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

For estimation, the additional moment restrictions result in L many additional parameters, one for each Lagrange multiplier. It is straightforward to adapt the estimation framework in Section 3 and have $\theta \in \Theta \subseteq \mathbb{R}^{L+2}$ gather the original parameters α and λ as well as the Lagrange multipliers $\mu_1, \mu_2, \cdots, \mu_L$. At the cost of a more cumbersome notation, all the asymptotic results in Section 3 apply.

[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 is preserved.

D Interpreting robustness

In this section I suggest an interpretation for robustness metric $\delta^*(\tilde{\tau})$. For every sample size n, the robustness metric $\delta^*(\tau)$ quantifies an exponential bound on the probability that, the empirical distribution $\hat{F}_{X,n}$ fails to satisfy $ATE(\hat{F}_{X,n}) \geq \tilde{\tau}$ even though $ATE(F_X) \geq \tilde{\tau}$. After introducing the method of *types* and stating Sanov's theorem, I revisit Example 5 to build the intuition in the finite-dimensional case.[15]

Suppose we collect a sample containing n i.i.d observations so we obtain a sequence of covariate values $x := (x_1, x_2, \cdots, x_n)$. Define $P_x(a) = \frac{N(a|x)}{n}$, the proportion of realizations of a appearing in x, out of n. We define the *type* P_x of x as a list of $P_x(a)$ for all possible values of a. We denote the collection of *types* as $\mathcal{P}_n$.[16] The empirical distribution $\hat{F}_{X,n}$ is a random variable, taking values in $\mathcal{P}_n$. 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. This is the set $E := \{Q \in \mathcal{P}_X : \int_{\mathcal{X}} \tau(x) dQ \leq \tilde{\tau}\}$, the constraint in Equation 5. Notice that whether $x \in E$ 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 $x \in E$? Sanov's theorem provides a link between this probability and the metric of robustness $\delta^*(\tau)$.

Theorem 19. *(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*

$$P_X^n(E \cap \mathcal{P}_n) \leq e^{-nD_{KL}(P_X^*||P_X)}$$
$$P^* := \min_{Q \in E} D_{KL}(Q||P_X)$$

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

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

Here $E := \{Q : \int_{\mathcal{X}} \tau(x) dQ \leq \tilde{\tau}\}$ is obtained through imposing a linear restriction

[15] Sanov's theorem remains true for larger classes of probability distributions, not necessarily confined to finitely supported X variables like discussed in Csiszár [1984].

[16] 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 $x = (a, a, b)$ and $x' = (a, b, a)$ are distinct: $x \neq x'$. But they have the same type: $P_{x'} = P_x$.

on Q and therefore E is convex. Note that $\delta^*(\tilde{\tau}) = D_{KL}(P^*||P)$ is precisely the metric of robustness in this paper. It provides a bound on the exponential decay of $P_X^n(E)$:

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

Importantly, the bound is non-asymptotic: it holds for any n. Since $\tilde{\tau}$ defines the constraint set E, it is natural for the above bound to depend on $\tilde{\tau}$. The bound becomes trivial when $\delta^*(\tau) = 0$ and it is monotonically decreasing in the magnitude of $\delta^*(\tau)$. When $\delta^*(\tau)$ approaches infinity, the set E will not contain any valid distributions, so it is reasonable that the upper bound converges to 0, a guarantee of greater robustness.

Example 5 (continuing from p. 13)**.** *Recall $\mathcal{X} = \{High, Medium, Low\}$ income group and the experimental distribution is $F_X = (p_1, p_2, p_3) = (0.2, 0.2, 0.6)$. We can list the types of sequences of size n that can be generated. Here, the proportion of High and Medium income individuals out of n determine a type. For $n = 3$, for example, there are 10 possible types: $(1, 0, 0)$, $(\frac{2}{3}, 0, \frac{1}{3})$, $(\frac{2}{3}, \frac{1}{3}, 0)$, $(\frac{1}{3}, \frac{2}{3}, 0)$, $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, $\cdots, (0, 0, 1)$. Therefore $|\mathcal{P}_3| = 10$. For $n = 10$, $|\mathcal{P}_{10}| = 66$. They are displayed below in barycentric coordinates as red points in the 2-simplex. The set E is in yellow.*

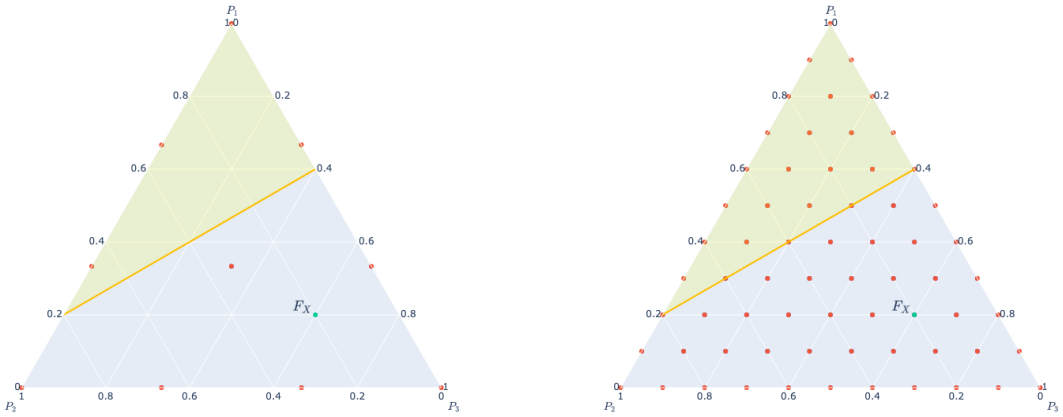

Figure 4: Possible *Types* for $n = 3$ (left panel) and $n = 10$ (right panel).

Each type P_x 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 types out of 10 that are in E, namely

$(3,0,0)$, $(2,0,1)$, $(2,1,0)$, $(1,2,0)$. *What is the probability associated with them?* $P_X^3(x_l \in E) = 0.128$. *Because here* $\delta^*(\tilde{\tau}) = 0.2492$, *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* $\lim n \to \infty$.

According to Sanov's theorem, the robustness metric is the main quantity that controls the probability of false rejections for the test of $ATE \leq \tau$ against $ATE > \tau$ using $\hat{F}_{X,n}$ in place of F_X. The quantitative interpretation of $\delta^*(\tau)$ is then precisely as the best exponent for an upper bound of this rejection probability.

D.1 An interpretation for the least favorable distribution

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 (4)? 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 20. *(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*[17] *conditional on the event that the empirical distribution* $\hat{F}_n \in E$

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

In contrast to Theorem 19 which holds for any n, Theorem 20 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 $\hat{F}_{X,n}$ does not satisfy the experimental conclusions. To visualize this, imagine drawing S-many repeated samples of n observations from P_X. Then, combining the previous results:

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

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

(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 (4) and (5) but it gives the approximate law of the data if we happen to draw a sample which does not satisfy the experimental results.

E Proofs of main results

I review a few basic results for optimization problems like the one in Equations (4-5). 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 (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 (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 (4) has a unique solution (P_X-almost everywhere) and the infimum is achieved. Theorem 8 characterizes such a solution P_X-almost everywhere.

E.1 Proof of Theorem 8

The proof is based on a result that appeared first in Donsker and Varadhan [1975]. More recently Ho [2023] has used a similar argument to characterize global sensitivity in a Bayesian setting. First note that, by the Radon-Nikodym theorem, $\frac{dF_X^*}{dF_X}$ exists

and 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:

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

Using a result from Donsker and Varadhan [1975]:

Lemma 21. *Let F^*_X satisfy $\frac{dF^*_X}{dF_X} = \frac{\exp(-\lambda(\tau(x)-\tilde{\tau}))}{\int_{\mathcal{X}} \exp(-\lambda(\tau(x)-\tilde{\tau}))dF_X}$. 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. 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:

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

□

E.2 Proof of Theorem 13

Some ancillary lemmas are needed to prove Theorem 13. Their proofs are collected in the Supplementary Appendix.

Lemma 22 (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 $(\mathbb{P}_n - \mathbb{P})(\hat{g} - g_0) = O_P\left(\frac{|\hat{g} - g_0|}{\sqrt{n}}\right)$.

Lemma 23. *For $\bar{\psi}(\theta, \gamma, \alpha) = \mathbb{E}[\psi(w, \theta, \gamma, \alpha)]$ we have:*

1. *$\bar{\psi}(\gamma, \alpha_0, \theta_0)$ is twice continuously Frechet differentiable in a neighborhood of γ_0.*
2. *If Λ is bounded then $\forall \theta \in \Theta$, $\bar{\psi}(\gamma, \alpha_0, \theta) \leq \bar{C} \|\gamma - \gamma_0\|_{L_2}^2$.*

Lemma 24 (Jacobian consistency)**.** *For Jacobian G of the debiased moment conditions:*

$$G = \mathbb{E}[D\psi(w, \theta_0, \gamma_0, \alpha_0)] = \mathbb{E}\left[\frac{\partial}{\partial \theta} \psi(w, \theta_0, \gamma_0, \alpha_0)\right] \tag{18}$$

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

Lemma 25 ($\sqrt{n}$ - consistency)**.** *Let Assumption 5 hold. Then*

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

We are ready to prove Theorem 13. Denote $\hat{G} = \frac{\hat{g}(w, \hat{\theta}, \hat{\gamma})}{\partial \theta}$. First note that by Lemma 24 we have $\|\hat{G} - G\| = o_P(1)$. Then, like in Chernozhukov et al. [2018] we

have:

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

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

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

Now by the central limit theorem and Lemma 25 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 the desired result:

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

E.3 Proof of theorem 15

The proof of Theorem 15 follows the same structure of Theorem 13 and is omitted.