A Simple Measure of Robustness for External Validity under Covariate Shifts

Pietro Emilio Spini[1]

This draft: January 2026
First draft: May 2021

Abstract

This paper studies the robustness of estimated policy effects to changes in the distribution of covariates, a key determinant of the external validity of (quasi)-experimental results. I propose a novel robustness metric δ^* which measures the smallest covariate shift needed to invalidate an empirical claim about the policy effect (e.g., $ATE > 0$). I estimate δ^* via de-biased GMM, achieving a parametric rate of convergence while accommodating machine-learning estimators of treatment-effect heterogeneity (e.g., LASSO, random forests, neural networks). I develop benchmarking and calibration exercises to interpret the magnitude of δ^*. I illustrate these tools in an application to the Oregon Health Insurance Experiment. Researchers can report δ^* alongside the point estimate and standard error as a third number gauging external validity under covariate shifts.

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, Stefan Hubner, David Pacini, Gregory Jolivet, Itzik Fadlon, Xavier D'Haultfoeuille, Matt Masten, Adam Rosen, Ashesh Rambachan, Davide Viviano, Michael Pollmann, Kirill Ponomarev, and Rami Tabri 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 Manchester, University of Surrey, NYUAD, UCSD, University of Victoria, and the Philadelphia Federal Reserve provided valuable discussion. All remaining errors are mine. A previous version of this paper circulated under the title "Robustness, Heterogeneous Treatment Effects, and Covariate Shifts."

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 (quasi)-experimental findings being robust and generalizable beyond the original experiment. 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 quantifying the robustness of their findings for policy adoption beyond the experimental setting.

In this paper, I introduce a new robustness metric, a scalar δ^*, that quantifies how much the characteristics of policy recipients would need to change to invalidate the (quasi)-experimental findings. The metric captures uncertainty arising from systematic differences in recipients' characteristics across environments.[1] This contrasts with uncertainty from sampling variation, summarized by the standard errors that routinely accompany point estimates. As such, δ^* complements standard errors and can be reported alongside them as a "third number". To make its magnitude operational for researchers, I suggest interpretation and calibration exercises, including benchmarking δ^* against covariate shifts in relevant implementation environments.

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 covariates' distribution. Small changes in the distribution of the covariates could

[1] Quantifying other sources of systematic 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 breakdown approaches in Horowitz and Manski (1995); Masten and Poirier (2020); Rambachan and Roth (2023).

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. Often, these covariates are exclusively collected in the experiment and are not all available in the new policy environment prior to implementation. As a result, the reweighting procedures in Hsu et al. (2020) and Hartman (2020) are often infeasible, since they require the full covariate set in the new environment. Moreover, the heterogeneity of policy effects can be hard to model. While domain knowledge can help select covariates that are predictive of the heterogeneity of policy effects, it typically cannot pin down its specific functional form. Because heterogeneity is the channel linking covariate shifts to 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 impose a functional form for the policy effect heterogeneity, letting it instead be flexibly estimated. 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.[2] If the way we measure a shift is misaligned with the heterogeneity model, the resulting measure of robustness may be misleading. For example, suppose that distance between two covariate distributions is measured only by their difference in means. With an unrestricted form for the heterogeneity of policy effects, one can construct a mean-preserving shift that invalidates the policy-maker's claim. If in the Oregon experiment, higher-income recipients have negative effects while lower-income recipients have positive effects, a mean-preserving spread of the income distribution could flip the aggregate effect. Yet, by construction, this shift would have a distance of zero from the experimental covariates, despite changing the experimental findings. This example motivates a robustness metric that accommodates flexible forms of policy effect heterogeneity, whose functional form is, *ex-ante*, unknown. My metric does so while remaining easy to compute 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 and the literature on Partial Policy Effects. For example, Rothe (2012) considers the effect of

[2]Many popular existing approaches to robustness and sensitivity analysis, like Altonji et al. (2005), Oster (2019) and Cinelli and Hazlett (2020), take advantage of specific functional forms.

an (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* and may itself be the object of interest as it represents, among all possible shifts invalidating the policy-maker's conclusion, the hardest one to detect. This distinction reflects the different purpose and the complementarity of the two approaches. A specific candidate for the covariate distribution is most useful for decomposition exercises that highlight the contribution of several variables on the unconditional distribution, like in the application in Rothe (2012). Conversely, searching within a large space of covariate distributions is useful for the policy-maker evaluating the experimental evidence for policy adoption.

Measuring covariate shifts requires choosing a distance between distributions. In my approach, I adopt Kullback-Leibler divergence (KL distance). It is a popular choice for sensitivity analysis, appearing recently in Christensen and Connault (2023) for moment inequality models, Duchi and Namkoong (2021) for distributionally robust stochastic optimization, and, in a Bayesian context in Ho (2023). 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 depending only on the observed covariate distribution and a functional parameter capturing the heterogeneity of policy effects.

Policy-effect heterogeneity can often be sparse: from a rich covariate set, only a few are needed to capture observable effect variation. With many covariates, it can be hard to select which ones are important *ex-ante*. Machine-learning estimators, like LASSO, random forest, or neural networks, can automatically exploit the sparsity and select the key covariates, avoiding *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 leveraging 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, letting the researcher flexibly choose among many first-step estimators of policy effect heterogeneity.

I apply my robustness procedure to study the Oregon health insurance experiment, whose findings have informed policy adoption in public health (Sanger-Katz, 2014). Focusing on health-care utilization and financial strain outcomes, I evaluate the robustness of the policy-effects estimates in Finkelstein et al. (2012) to covariate shifts. The recipients of the Oregon lottery are predominantly older, in poorer health, and with a larger proportion of White individuals than the national average (Finkelstein, 2013). 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. Differences in the magnitude and sign of the effects of the Medicaid expansions in Oregon and Massachusetts have motivated Kowalski (2023) to investigate the different populations of beneficiaries in the two states. My robustness exercise is complementary: I compute the smallest covariate shift from the Oregon experiment that eliminates the lottery's positive effects on health-care utilization and financial strain. Among the outcomes considered, outpatient visits are the most robust.

This paper is also related to the econometric and statistics literature on robustness and sensitivity analysis developed since 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, they are readily available to the policy-maker and are likely to be of first-order 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 setup and defines the robustness metric. Section 3 presents the estimator and its asymptotic properties. Section 4 applies the approach to the findings from the Oregon health insurance experiment and provides interpretations for the robustness measure. Section 5 concludes. The main proofs are in the Appendix. Additional results and proofs of the lemmas are available in Spini (2024). Throughout, all results apply to both experimental and quasi-experimental settings. I use "experiment" to refer to both interchangeably.

2 A robustness metric for covariate shifts

In this section, I link treatment-effect heterogeneity to robustness in a potential-outcomes framework, focusing on the Average Treatment Effects (ATE). A policy-maker is interested in whether a claim such as $ATE > \tilde{\tau}$ remains valid when the covariate distribution differs from the experiment. Using the Conditional Average Treatment Effect (CATE), which aggregates to the ATE, I characterize the covariate distribution that (i) violates the claim and (ii) is closest to the experimental distribution. I call it the *least-favorable distribution* because it is the hardest to distinguish from the experimental distribution. Distance is measured by the KL divergence: the resulting minimum distance, δ^*, is my robustness metric. Any covariate distribution within KL distance δ^* of the experimental distribution necessarily preserves the claim.

2.1 Set-up and Preliminaries

Let Y_d denote the potential outcome under binary treatment d. In the experiment, for each unit, one observes treatment status $D \in \{0, 1\}$, realized outcome $Y = DY_1 + (1 - D)Y_0 \in \mathcal{Y}$, and covariates $X \in \mathcal{X}$.[3] I partition the covariates as $X = (X_c, X_e)$, where X_c collects covariates with counterparts in census-type data in other states, and X_e collects experiment-specific covariates with no natural counterpart outside the experiment. In my empirical application, X_c includes indicators for race, gender, age, education, and urban area whereas X_e includes Finkelstein et al. (2012)'s proxy for health status. Though treatment effect heterogeneity will be estimated using the full vector X, its partition into (X_c, X_e) will play a role in my benchmarking exercise to gauge the magnitude of my robustness metric. Let P_X denote the probability measure for X in the experiment and let F_X be its associated distribution. I distinguish potential outcomes under F_X from those under an alternative F'_X by writing Y_d and Y'_d, respectively. The propensity score is $\pi(x) = P(D = 1 \mid X = x)$. Finally, for any random variable W, let $\mathcal{W}$ denote its support. The interior of a set S is S°.

Assumption 1. *Unconfoundedness and Overlap*

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

ii) There exists an $\epsilon > 0$ *such that all* $x \in \mathcal{X}$ *we have* $0 < \epsilon \leq \pi(x) \leq 1 - \epsilon < 1$

[3]Additional control variables W can be accommodated. I suppress them here to simplify notation.

In an RCT, under complete or covariate-based randomization of treatment assignment, Assumption 1 i) 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: weaker forms still allow identification, this version is needed for estimation in Section 3.

The policy-maker's parameter of interest is the $ATE := \mathbb{E}[Y_1 - Y_0]$. The CATE, defined by $\tau(x) := CATE(x) = \mathbb{E}[Y_1 - Y_0|X = x]$, captures how average effects change across sub-populations with covariate value $X = x$. Under Assumption 1 i), $\tau_{F_X}(x)$ is nonparametrically identified by the difference between $\gamma_1(X) := \mathbb{E}[Y|D = 1, X = x]$ and $\gamma_0(X) = \mathbb{E}[Y|D = 0, X = x]$ using data from the experiment (Imbens and Rubin, 2015).[4] ATE is obtained by averaging $\tau(x)$ with weights proportional to F_X. We can write the map sending F_X to its corresponding ATE as:

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

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 affect the magnitude of ATE through two channels: a direct effect through 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 use the covariate shift assumption[5] 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 $d \in \{0, 1\}$, 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 with respect to the distribution of covariates. Assumption 2 ii) says the

[4]If the CATE only partially identified, like in the case of non-compliance based on unobservables, it is possible to follow a bounding approach for my robustness procedure. This approach is sketched in Spini (2024) but I leave the details for future research.

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

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 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.[6] Then, the link between F_X and ATE reduces to integration against a fixed $\tau(x)$ and the map in Eq.(1) is linear in F_X:

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

Before presenting the general framework I give the simplest nontrivial example of a robustness exercise with respect to covariate shifts.

Example 1. *Consider a binary covariate $X = \{0, 1\}$. D is randomized, satisfying Assumption 1. Then $\tau(x) := \mathbb{E}[Y_1|X = x] - \mathbb{E}[Y_0|X = x]$ is identified. Because X is Bernoulli, any distribution on $\{0, 1\}$ is fully described by $P_X(X = 1) = p_1$.*

$$ATE(F_X) = ATE(p_1) = \tau(0) \cdot (1 - p_1) + \tau(1) \cdot p_1.$$

Suppose the experimental $ATE > 0$ and $\tau(1) > 0 > \tau(0)$: treatment effects have different signs. What is the closest covariate distribution invalidating the claim $ATE > 0$? Setting ATE to 0 and solving for p_1^*:

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

$|p_1^* - p_1| = |\frac{-\tau(0)}{\tau(1)-\tau(0)} - p_1|$ *is the smallest shift in p_1 that invalidates the claim $ATE > 0$.*

Under what conditions does a solution like p_1^* exist in general, and when is it unique? When $\mathcal{X}$ is not discrete, distributions on $\mathcal{X}$ are inherently infinite-dimensional unless one imposes parametric restrictions. Moreover, how should one measure the distance between p_1^* and p_1 in this case? Motivated by these questions, I adopt a nonparametric notion of distance between probability distributions.[7]

Definition 2.1 (KL-divergence)**.** *The KL-divergence between two distributions F_X and F'_X is given by:*

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

[6] This identification result follows immediately from Ass. 2. See Hsu et al. (2020), Lemma 2.1.

[7] Spini (2024) discusses how the general procedure can be specialized to certain parametric classes of distributions. In such cases, the relevant covariate shifts coincide with mean shifts.

where $\frac{dF'_X}{dF_X}$ is the Radon-Nikodym derivative of distribution F'_X w.r.t the experimental distribution F_X, provided that $P'_X \ll P_X$ for the respective probability measures.

Why choose KL? There are several ways to measure the distance between a candidate distribution F'_X and F_X, each with advantages and disadvantages. For our purposes, a first requirement is to measure discrepancies without committing to a parametric family $\mathcal{F}$. As noted in the introduction, unless the shape of $\tau(x)$ is restricted jointly with $\mathcal{F}$, one can change the ATE while keeping a parametric discrepancy arbitrarily small. This requirement rules out, for example, Mahalanobis-type measures and motivates focusing on nonparametric distances.

Two prominent classes of nonparametric distances are Integral Probability Metrics (IPMs) and ϕ-divergences. IPMs include Total Variation, kernel-based distances like MMD and, notably, Wasserstein distances (e.g., W_1, W_2) which have been increasingly adopted in recent work (e.g., Adjaho and Christensen (2022); Athey et al. (2024); Gunsilius (2023)). For our problem of quantifying robustness to covariate shifts, Wasserstein is less attractive for two reasons. First, Wasserstein is generally not invariant to invertible transformations of the covariates X. More broadly, under standard regularity conditions, requiring an IPM to satisfy this invariance property is highly restrictive and, in effect, forces the distance to be a ϕ-divergence (Qiao and Minematsu, 2010). As a consequence of this lack of invariance, transformations commonly used in empirical work, such as logs or inverse hyperbolic sine, (which are not isometries), can mechanically change the measured Wasserstein distance and the resulting robustness assessment. Second, Wasserstein-based optimization typically relies on entropic regularization to inject sufficient convexity for computational tractability.[8] In contrast, the KL formulation naturally enjoys this convexity and, in addition, delivers a sharp closed-form characterization as discussed in Section 2.3.

The second prominent class is the family of ϕ-divergences (including χ^2, Hellinger, and KL). Unlike Wasserstein, all ϕ-divergences are invariant to any invertible measurable transformation of X (with measurable inverse).[9] In addition, KL has three advantages that make it a particularly attractive choice for our problem. First, under KL the minimization problem in Eq.(4)-(5) has an exponentially tilted, unique interior solution (Theorem 3.2), which in turn enables convenient (de-biased) GMM

[8]For results on convergence rates for entropy-regularized Wasserstein distance see Tabri (2026).

[9]One can show the following result. Let $T : \mathcal{X} \to T(\mathcal{X})$ be a (deterministic) measurable transformation with measurable inverse $T^{-1}(\cdot)$. Then $D_{KL}(G_{T(X)} \| F_{T(X)}) = D_{KL}(G_X \| F_X)$.

estimation and interpretation of the minimizer (discussed in Theorem 3.2). Second, KL satisfies an exact chain-rule decomposition over conditional distributions, which I exploit in the benchmarking exercise in Section 2.4. Third, KL's natural connection to hypothesis testing and power lets us quantify "distance" in terms of statistical distinguishability: a KL magnitude can be mapped to the minimal sample size needed to detect a covariate shift. I illustrate this connection in Section 4.1.

2.2 The policy-maker's problem: quantifying robustness

Once the ATE–covariate link and a distance measure are specified, I can formalize the policy-maker's robustness problem. Consider the claim given by $ATE > \tilde{\tau}$ (the reverse inequality is analogous) where $\tilde{\tau}$ is a policy-relevant threshold capturing, for example, implementation costs or the value of a competing policy. In Example 1 $\tilde{\tau} = 0$, often a natural benchmark. The policy-maker is interested in the smallest shift from the experimental distribution, F_X, invalidating the claim $ATE > \tilde{\tau}$. Formally:

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

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

Eq.(4)–(5) define an optimization problem over the set of covariate distributions violating the claim $ATE > \tilde{\tau}$ and the objective selects the one closest to F_X in KL divergence. By Assumption 2, $\tau(x)$ in Eq.(5) is not indexed by F'_X: the constraint set is linear in F'_X. Since $D_{KL}(\cdot \| F_X)$ is a strictly convex function on a convex feasible set, the problem admits a unique (P_X–a.e.) solution, characterized in Theorem 2.1.

Remark 1. *The program in Eq.(4)–(5) is nonparametric: no structure beyond absolute continuity $P'_X \ll P_X$ is imposed on F'_X.*[10] *Absolute continuity implies that F'_X cannot assign mass outside the experimental support $\mathcal{X}$, which I view as natural: the feasible covariate distributions should not place weight on subpopulations for which the experiment provides no data. Similar support restrictions are standard in the distributional policy effects literature, e.g., Rothe (2012). Without them, $\tau(x)$ is unidentified at covariate values x that are never observed and can be chosen to generate arbitrarily large average effects, rendering the robustness exercise uninformative.*

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

Definition 2.2. *Fix $\tilde{\tau}$. The robustness metric $\delta^*(\tilde{\tau})$ is the infimum of Eq.(4). The*

[10] For example, if both P_X and P'_X are absolutely continuous with respect to a common dominating measure λ and $\text{supp}(P'_X) \subseteq \text{supp}(P_X)$, then $P'_X \ll P_X$.

(set of) least-favorable distribution(s) $\{F_X^*(\tilde{\tau})\}$ is the (set of) minimizer(s) of Eq.(4).

I define $\delta^*(\tilde{\tau})$ as the KL distance between the experimental covariate distribution F_X and a *least-favorable* distribution F_X^*, i.e., the KL-closest covariate shift that violates the claim $ATE > \tilde{\tau}$. As I show in Section 4.1, among all shifts that invalidate the claim, this least-favorable shift is the hardest to detect statistically. The set $\{F_X^*(\tilde{\tau})\}$ may be empty for some values of $\tilde{\tau}$. When the experimental ATE already violates the target (i.e., $ATE(F_X) \leq \tilde{\tau}$), the experimental distribution is feasible and achieves the minimum: one can take $F_X^* = F_X$, so $\delta^*(\tilde{\tau}) = 0$. The problem is nontrivial when $ATE(F_X) > \tilde{\tau}$, in which case F_X is infeasible for (5) and, under regularity conditions, any least-favorable solution must satisfy $\delta^*(\tilde{\tau}) > 0$. In Example 1, I imposed $ATE(p_1) > 0$ precisely to ensure this nontrivial case.

With discrete $\mathcal{X}$, F_X is a point in the probability simplex. The optimization in Eq.(4)-(5) has a simple geometry: the constraint $\int_{\mathcal{X}} \tau(x)\, dF'_X(x) \leq \tilde{\tau}$ defines the set of covariate shifts invalidating the claim, and the KL objective selects the closest one.

Example 2. *Let $\mathcal{X} = \{H, M, L\}$ index income bins and write $F'_X = (p_H, p_M, 1-p_H-p_M)$. Suppose $(\tau(H), \tau(M), \tau(L)) = (1, 2, 3)$ and set the threshold $\tilde{\tau} = 1.8$. Under the experimental distribution, $ATE(F_X) = 2.4 > \tilde{\tau}$.*

Figure 1 depicts KL level sets around F_X, the feasible region, and the *least-favorable distribution* from Example 2. Since the objective and constraint in Eq.(4)-(5) are smooth in (p_H, p_M), the solution is characterized by the KKT conditions. The green curve is the KL level set at $\delta^*(\tilde{\tau})$: any covariate distribution closer than $\delta^*(\tilde{\tau})$ is guaranteed to satisfy the policy-maker's claim. When $\mathcal{X}$ is not discrete, a visualization like Figure 1 may be difficult. Nonetheless, under mild regularity conditions, a *least-favorable distribution* exists, is unique and admits a closed-form characterization via exponential tilting, with a well-defined $\delta^*(\tilde{\tau})$ for a range of values of $\tilde{\tau}$.

2.3 A closed form solution for quantifying robustness

I characterize the solution to Eq.(4)-5 under regularity conditions.

Assumption 3 (Boundedness)**.** *The conditional average treatment effect $\tau(X)$ is bounded P_X-a.s. (i.e., there exists $M < \infty$ such that $\mathbb{P}_X(|\tau(X)| \leq M) = 1$).*

Assumption 3 also holds P'_X-a.s. for any $P'_X \ll P_X$, since absolute continuity prevents P'_X from assigning mass to P_X-null sets. Boundedness as in Assumption 3 is plausible in a cross sectional, micro-econometrics setting. One could weaken it to an exponential-moment condition, e.g., $\mathbb{E}[\exp(\kappa|\tau(X)|)] < \infty$ for all $\kappa > 0$ (see

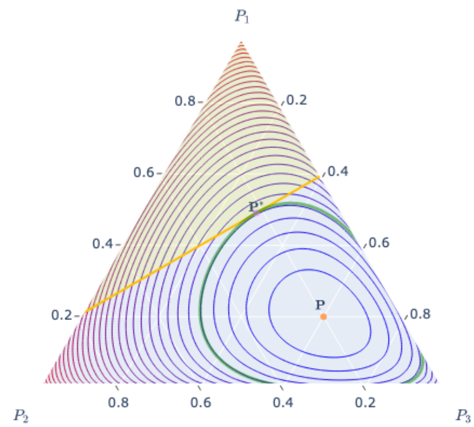

Figure 1: The triangle is the 2-simplex of distributions $P' = (p_H, p_M, p_L)$ on $\mathcal{X} = \{H, M, L\}$ (barycentric coordinates). The experimental distribution is $P = (0.2, 0.2, 0.6)$. The shaded region consists of distributions with $ATE(P') \leq 1.8$. Contours are KL level sets $D_{KL}(P' \| P)$ (darker = smaller). The point $P^* = (0.491, 0.218, 0.291)$ is the *least-favorable distribution*. The green contour is the level set at $\delta^* \approx 0.296$; any P' with $D_{KL}(P' \| P) < \delta^*$ is guaranteed to satisfy the claim.

Komunjer and Ragusa, 2016). I maintain Assumption 3 for ease of exposition.

The constraint set $\{F'_X : \int \tau \, dF'_X \leq \tilde{\tau}\}$ is a convex set. If empty, it implies a value of $+\infty$ for the program in Eq.(4)-(5). To rule this out, I assume $\tilde{\tau}$ is attainable via some covariate shift, i.e., there exists an F'_X with $ATE(F'_X) = \tilde{\tau}$. If $\tau(x)$ is not sufficiently heterogeneous, like in Example 3, ATE can never reach $\tilde{\tau}$.

Example 3 (Homogeneous treatment effects)**.** *If $\tau(x) = c$, for some $c \in \mathbb{R}$, then $ATE(F'_X) = c$ for every F'_X: no covariate shift can reach a different threshold $\tilde{\tau} \neq c$.*

Constant treatment effects are an extreme case: under Assumption 2, if $\tau(x) = c$ then $ATE(F'_X) = c$ for every F'_X, so the claim $ATE > \tilde{\tau}$ can be extrapolated from the experiment to any environment. More generally, if $\tilde{\tau}$ lies outside the range of $\tau(X)$, then no covariate distribution can satisfy $ATE = \tilde{\tau}$, the feasible set in Eq.(5) is empty (hence the value of Eq.(4) is $+\infty$). For instance, if $2 \leq \tau(X) \leq 5$ a.s., then $\tilde{\tau} = 1$ is unattainable. Non-trivial robustness requires some minimal heterogeneity of treatment effects.[11] The following assumption guarantees a nonempty feasible set:

Assumption 4. *(Non-emptiness) Recall $\mathcal{F} := \{F'_X \text{ s.t. } P'_X \ll P_X\}$. Define $L(F_X) := \int_{\mathcal{X}} \tau(x) \, dF_X(x)$. I require $\tilde{\tau} \in L^\circ(\mathcal{F})$.*

Assumption 4 is a joint condition on $\tau(\cdot)$ and $\tilde{\tau}$. It requires $\tau(\cdot)$ to be sufficiently heterogeneous so that some $F'_X \in \mathcal{F}$ attains $ATE(F'_X) = \tilde{\tau}$. In Example 1, Assumption 4 is satisfied: $\tau(0) < 0 < \tau(1)$ guaranteeing $\tilde{\tau} = 0 \in L^\circ(\mathcal{F}) = (\tau(0), \tau(1))$. In

[11] It is convenient to restrict the robustness metric to a real valued parameter rather than $\mathbb{R} \cup \{+\infty\}$.

Example 3, it fails: $L(\mathcal{F}) = \{c\}$ so $L^\circ(\mathcal{F}) = \varnothing$. More broadly, the set $L(\mathcal{F})$ captures how rich is the set of ATEs that could be attained by a covariate shift F'_X.

Assumption 4 also has testable implications. Given an estimate of $\tau(x)$, one can test whether $\tilde{\tau} \in (\inf_x \tau(x), \sup_x \tau(x))$ using the intersection-bounds procedure in Chernozhukov et al. (2013). Homogeneous (or nearly homogeneous) treatment effects are cases in which such tests may reject. Technically, Assumption 4 serves as a (Slater-like) constraint qualification guaranteeing strong duality and the existence of the *least-favorable distribution* (see Komunjer and Ragusa (2016) for an excellent overview on optimization problems involving KL and other φ-divergences).[12] After introducing Assumptions 1-4 we can state the key result below:

Theorem 2.1 (Closed form solution)**.** *Let Assumptions 1-4 hold. Then: i) Eq.(4) attains a minimum at F^*_X, uniquely 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 λ is the unique 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 robustness metric $\delta^*(\tilde{\tau})$ is finite and 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 2.1 shows that the general robustness problem admits an exponential-tilting solution, making optimization over the nonparametric class $\mathcal{F}$ no harder than the parametric Examples 1 and 2. Their solution from Theorem 2.1 coincides with the one obtained directly by the KKT conditions. The theorem also implies $\delta^*(\tilde{\tau})$ is identified as a functional of $(F_X, \tau(\cdot))$, both nonparametrically identified from the experimental data, and it motivates the estimation approach in Section 3.

$\delta^*(\tilde{\tau})$ is distinct from other popular statistics used to summarize treatment-effect heterogeneity. Relative to rich descriptive summaries like sorted effects (Chernozhukov et al., 2018b) and conditional quantile treatment effects (Koenker, 2005), $\delta^*(\tilde{\tau})$ is more parsimonious, collapsing heterogeneity into a single, easy-to-report number. Relative to other scalar summaries, $\delta^*(\tilde{\tau})$ is directly geared to the external-validity exercise:

[12]The interior condition cannot be weakened. Under Ass. 3, $L(\mathcal{F})$ is an interval. If $\tilde{\tau}$ lies on the boundary, the feasible set in Eq.(5) collapses to a degenerate measure w.r.t F_X, implying infinite KL in Eq.(4). See the quasi-relative interior condition in Borwein and Lewis (1993) Eq.(BL).

$\mathrm{Var}(\tau(X))$ is threshold-free and therefore not tied to a particular claim, though it can be locally informative about robustness when ATE_{F_X} lies near $\tilde{\tau}$. Likewise, the tail mass $P(\tau(X) \leq \tilde{\tau})$ is threshold-specific but coarse, as it ignores changes in the shape of $\tau(X)$ that do not move mass across $\tilde{\tau}$. Overall, these trade-offs reflect complementary goals for these summaries. In this sense, $\delta^*(\tilde{\tau})$ expands the researcher's toolkit with a simple way to gauge external validity under covariate shifts.

2.4 Benchmarking and interpreting robustness

After a researcher reports $\delta^*(\tilde{\tau})$, how should they interpret its magnitude? I propose to benchmark $\delta^*(\tilde{\tau})$ against the covariate shifts one could observe in other implementation environments, when such data is available. I leverage a decomposition of KL for the partition $X = (X_c, X_e)$ leading to the useful result (Cover, 1999):

Proposition 2.1. *Let $X = (X_c, X_e)$. Then for any two distributions F and G:*

$$D_{KL}(G_X \| F_X) = D_{KL}(G_{X_c} \| F_{X_c}) + \mathbb{E}_{G_{X_c}}\left(D_{KL}(G_{X_e|X_c} \| F_{X_e|X_c})\right). \tag{9}$$

If $X_c \perp\!\!\!\perp X_e$ for both F and G, then $D_{KL}(G_X \| F_X) = D_{KL}(G_{X_c} \| F_{X_c}) + D_{KL}(G_{X_e} \| F_{X_e})$.

Proposition 2.1 decomposes the left-hand side into two additive terms: (1) the KL between marginals F_{X_c} and G_{X_c} and (2) the average KL between the conditional distributions $F_{X_e|X_c}$ and $G_{X_e|X_c}$. Because the second term in Eq.(9) is the expectation of a non-negative quantity, we have the lower bound $D_{KL}(G_{X_c} \| F_{X_c}) \leq D_{KL}(G_X \| F_X)$ (with equality if and only if $G_{X_e|X_c} = F_{X_e|X_c}$ G_{X_c}-a.s.). Eq.(9) yields a simple benchmarking strategy for $\delta^*(\tilde{\tau}) = D_{KL}(F_X^* \| F_X)$. For any candidate implementation environment where X_c is observed, we can compute the first term. Because the second term is not identified when X_e is unavailable outside the experiment, a practical approach is to calibrate its magnitude as a proportion of the first term so that $D_{KL}(G_X \| F_X) = (1 + \kappa)\, D_{KL}(G_{X_c} \| F_{X_c})$. Since both are in KL units, the comparison supports cardinal statements such as *"$\delta^*(\tilde{\tau})$ is 20% larger than the KL distance based on the census covariates X_c between Texas and the Oregon experiment"*. If this κ calibration is valid, any environment with $(1+\kappa)\, D_{KL}(G_{X_c} \| F_{X_c}) < \delta^*(\tilde{\tau})$ is guaranteed to preserve the experimental claim. The special case $\kappa = 0$ corresponds to the assumption $G_{X_e|X_c} = F_{X_e|X_c}$: the conditional distribution of the missing covariates matches the experiment. Section 4.1 illustrates this benchmarking exercise for the empirical application in Finkelstein et al. (2012), where other states are the new implementation environments, X_c includes race, gender, age, education, and urban indicators (available in the state census), while X_e is a proxy for health status (available only in the experiment). In the empirical application, I consider $\kappa \in \{0, 0.2, 1\}$, corresponding to no unobserved contribution, a contribution proportional to the relative dimension of X_e and X_c, and equal observed and unobserved contributions.

3 Estimation and Asymptotic Results

This section develops semiparametric estimators and establishes the asymptotic properties of: (i) the robustness metric δ^* in Equation (8) and (ii) of user-specified moments of the least favorable distribution in Equation (6). The analysis builds on the de-biased GMM framework of Chernozhukov et al. (2020). Let $W = (Y, D, X)$ be the data. The characterization of Eq.(7)-(8)- in Theorem 2.1 as integrals w.r.t F_X implies a moment condition. As in Newey and McFadden (1994), define a parameter $\theta_0 := (\nu_0, \lambda_0)^T$ satisfying the population moment:

$$\mathbb{E}[g(W, \theta, \tau)] = \mathbb{E}\begin{bmatrix} \exp(-\lambda_0(\tau_0(X) - \tilde{\tau})) - \nu_0 \\ \exp(-\lambda_0(\tau_0(X) - \tilde{\tau}))(\tau_0(X) - \tilde{\tau}) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \tag{10}$$

where $\tau_0(\cdot)$ is the true CATE. Assumptions 1–4 guarantee that (ν_0, λ_0) are globally identified by Equation (10). From these, the robustness metric in Equation (8) is identified by $\delta^* = -\log(\nu_0)$. The parameter space $\Theta \subseteq \mathbb{R}^2$ satisfies $0 < \nu_0 \leq 1$. Positivity follows by $\exp(-\lambda_0(\tau(x) - \tilde{\tau})) > 0$ for all x. If $ATE > \tilde{\tau}$ holds under F_X, then $\delta^* > 0$ and hence $\nu_0 < 1$. Because the true $\tau_0(X)$ is unknown but identified, a feasible version of Equation (10) replaces $\tau_0(X)$ with a first step estimate $\hat{\tau}(X)$. The plug-in GMM $\hat{\theta} = (\hat{\lambda}, \hat{\nu})^T$ solves the sample equivalent of Equation 10.

Assumption 1 guarantees nonparametric identification of $\tau_0(X)$ via $\gamma_1(X) - \gamma_0(X)$, leaving the researcher free to choose among many estimation strategies, both parametric and nonparametric. Popular options include random forest (Athey et al., 2016) or doubly robust scores (Hsu et al., 2020). Because moment conditions in Equation (10) are not Neyman orthogonal w.r.t $\hat{\tau}(X)$, the first-step estimation of $\hat{\tau}(X)$ may affect the asymptotic distribution of $\hat{\delta}^* = -\log(\hat{\nu})$ through the plug-in estimator. This can lead to invalid inference unless one imposes some *ad-hoc* rate conditions on the first step, which are restrictive and difficult to verify in practice (Chernozhukov et al., 2018a). To avoid that, following Chernozhukov et al. (2020), I derive a debiased-GMM estimator that yields orthogonal moments and valid inference under a range of flexible estimators for $\tau_0(X)$. Details of this derivation are in Spini (2024). It is convenient to index any functional nuisance parameter like $\gamma_1(X)$ and $\gamma_0(X)$ with the distribution of the data, with F_0 denoting the true distribution.

Proposition 3.1. *The nonparametric influence function based on $g(\cdot)$ in Eq.(10) is:*

$$\phi(w,\theta,\gamma_0,\alpha_0) = \begin{bmatrix} \exp\left(-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})\right)\cdot(-\lambda) \\ \exp\left(-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})\right)\cdot\left(1-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})\right) \end{bmatrix}$$
$$\times\Big(\alpha_{1,F_0}(x)\,d\,(y-\gamma_{1,F_0}(x))+\alpha_{0,F_0}(x)\,(1-d)\,(y-\gamma_{0,F_0}(x))\Big)$$
$$\alpha_{F_0} := (\alpha_{1,F_0}(x),\alpha_{0,F_0}(x)) = \left(\frac{1}{\pi_{F_0}(x)},\frac{1}{1-\pi_{F_0}(x)}\right), \tag{11}$$

Proposition 3.2. *The de-biased moment functions below are Neyman orthogonal.*

$$\psi(w,\gamma,\theta,\alpha) = g(w,\theta,\gamma)+\phi(w,\theta,\gamma,\alpha). \tag{12}$$

Note that $\mathbb{E}_{F_0}[\psi(W,\theta,\gamma_0,\alpha_0)] = 0$. Under regularity conditions $\mathbb{V}[\psi(W,\theta,\gamma_0,\alpha_0)] < \infty$, $\psi(\cdot)$ is a valid influence function. The K-fold cross-fitted de-biased GMM equations from Eq.(12), and the corresponding estimator for θ are:

$$\hat{\psi}(\theta,\hat{\gamma},\hat{\alpha}) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{|I_k|}\sum_{i\in I_k}\Big(g(W_i,\theta,\hat{\gamma}_{-k})+\phi(W_i,\tilde{\theta},\hat{\gamma}_{-k},\hat{\alpha}_{-k})\Big) \tag{13}$$
$$\hat{\theta} = \arg\min_{\theta\in\Theta}\hat{\psi}(\theta,\hat{\gamma},\hat{\alpha}), \tag{14}$$

where $\tilde{\theta}$ is a consistent estimator for θ needed to evaluate $\phi(\cdot)$.[13] For each fold $k=1,\cdots,K$, $\gamma(\cdot)$ and $\alpha(\cdot)$ are estimated on the remaining $(K-1)$ folds, then the empirical moment is evaluated on fold k. Sample-splitting reduces own-observation bias and, together with the Neyman orthogonality, avoids complicated Donsker-type conditions that could fail for some estimators of $\hat{\gamma}$ and $\hat{\alpha}$, as discussed in Chernozhukov et al. (2020). To establish $\sqrt{n}$-convergence of $\hat{\theta}$, some mild conditions on the L_2 rates of convergence of the first-step estimators $\hat{\gamma}$ and $\hat{\alpha}$ are required.

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

Assumption 5 can be satisfied by many flexible nonparametric estimators of $\hat{\gamma}$ including machine learning-based estimators like lasso, random forest, boosting, and neural nets. In practice, these can be useful when the covariate space is large but the true $\tau_0(X)$ has a sparse representation. Calibrated Monte Carlo simulations in Spini (2024) illustrate this point. I derive the influence function representation for $\hat{\theta}$ under Assumptions 1-5 which implies the following asymptotic normality result.

[13] A natural candidate is the plug-in GMM, which is $o_p(1)$ but may not be $o_p(n^{-\frac{1}{2}})$ in general.

Theorem 3.1 (A.N. of $\hat{\theta}$). *Let Assumptions 1–5 hold. For $\hat{\theta}$ in Equation (14):*

$$\sqrt{n}(\hat{\theta}-\theta_0) \xrightarrow{d} \mathcal{N}(0,S) \text{ with } S := (G^{-1})\Omega(G^{-1})^T,$$
$$G := \mathbb{E}[D_\theta \psi(w,\theta_0,\gamma_0,\alpha_0)], \quad \Omega := \mathbb{E}[\psi(w,\theta_0,\gamma_0,\alpha_0)\psi(w,\theta_0,\gamma_0,\alpha_0)^T],$$

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

Corollary 1 (A. N. of $\hat{\delta}^*$). *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 3.1.

Theorem 3.1 provides point estimates and confidence intervals for δ^*. Because δ^* is defined as the minimal distance required to invalidate the policy-maker's claim, one would focus on the lower bound: overestimating δ^* overstates robustness, whereas underestimating is valid but conservative. This approach is analogous to Masten and Poirier (2020) who report a one-sided confidence region for their breakdown frontier.

3.1 Reporting features of the *least-favorable distribution*

In addition to δ^*, the researcher may be interested in the *least-favorable distribution* F_X^* itself. When $\mathcal{X}$ is large, representing F_X^* directly can be impractical. Instead, the researcher can compare F_X and F_X^* by reporting some moments. Reporting covariate means across treatment arms is standard practice in assessing internal validity (e.g. Rosenbaum and Rubin (1984)). By analogy, comparing moments from F_X and F_X^* could speak to external validity. Here I provide an estimator for a user-specified, finite collection of moments of F_X^*.

Theorem 3.2 shows the asymptotic properties of the joint estimator for θ and the additional moments, denoted by ζ.

Theorem 3.2 (De-biased estimator of *least-favorable* moments). *Let $u : \mathbb{R}^d \to \mathbb{R}^s$, with $u \in (L^\infty(\mathcal{X}, P_X))^s$. Let $\zeta_0 = \mathbb{E}_{F_X^*}[u(X)] \in \mathbb{R}^s$. Define the estimating equation for $(\hat{\theta},\hat{\zeta})$, augmenting θ by the desired moments ζ of the least-favorable distribution.*

$$\hat{\psi}^u(\theta,\zeta,\hat{\gamma},\hat{\alpha}) := \frac{1}{K}\sum_{k=1}^{K}\frac{1}{|I_k|}\sum_{i\in I_k}\begin{bmatrix} g(W_i,\theta,\hat{\gamma}_{-k}) + \phi(W_i,\theta,\hat{\gamma}_{-k},\hat{\alpha}_{-k}) \\ g^u(W_i,\theta,\zeta,\gamma_{-k}) + \phi^u(W_i,\theta,\zeta,\hat{\gamma}_{-k},\hat{\alpha}_{-k}) \end{bmatrix}$$
$$(\hat{\theta},\hat{\zeta}) := \arg\min_{(\theta,\zeta)\in\mathbb{R}^{s+2}} \hat{\psi}^u(\theta,\zeta,\hat{\gamma},\hat{\alpha})^T\hat{\psi}^u(\theta,\zeta,\hat{\gamma},\hat{\alpha}) + o_P(1)$$

where $g(\cdot),\phi(\cdot),\gamma(\cdot)$ and $\alpha(\cdot)$ are the same as in Proposition 3.1, and $g^u(\cdot)$ and $\phi^u(\cdot)$, *taking values in $\mathbb{R}^s$ and defined below.*

$$g^u(W_i, \theta, \zeta, \gamma) = u(X_i) \exp(-\lambda(\tau(X_i) - \tilde{\tau})) - \nu \cdot \zeta$$

$$\phi^u(W_i, \theta, \zeta, \gamma, \alpha) = u(X_i) \exp\left(-\lambda(\tau(X_i) - \tilde{\tau})\right)(-\lambda) \left(\frac{D_i(Y_i - \gamma_1(X_i))}{\pi(X_i)} - \frac{(1 - D_i)(Y_i - \gamma_0(X_i))}{1 - \pi(X_i)} \right)$$

Let Assumptions 1–5 hold. Then:

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

$$\sqrt{n} \left(\begin{bmatrix} \hat{\theta} - \theta_0 \\ \hat{\zeta} - \zeta_0 \end{bmatrix} \right) \xrightarrow{d} \mathcal{N}(0, S^u) \text{ with } S^u := (G^u)^{-1} \Omega^u (G^{u'})^{-1},$$

$$G^u := \mathbb{E}[D_{\theta,\zeta} \psi^u(W, \theta, \zeta, \gamma_0, \alpha_0)], \quad \Omega^u := \mathbb{E}[\psi^u(w, \theta_0, \zeta_0, \gamma_0, \alpha_0) \psi^u(w, \theta_0, \zeta_0, \gamma_0, \alpha_0)^T],$$

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

4 Empirical Application

Between March and September 2008, Oregon conducted lottery draws granting winners the option to enroll in Oregon Health Plan (OHP) Standard, a Medicaid expansion program for uninsured residents aged between 19 and 64, who have limited income and assets. In a seminal study, Finkelstein et al. (2012) find positive effects of the insurance coverage on a variety of outcomes: health-care utilization (number of prescription, inpatient, outpatient and ER visits), preventive care (cholesterol and diabetes blood test, mammogram and pap-smear test) and measures of financial strain (outstanding medical debt, denied care, borrow/skip).[14] Because not all lottery winners exercised their option to enroll, Finkelstein et al. (2012) report both an ITT and a LATE estimate. I focus on the ITT as it is likely to be the relevant parameter[15] for a policy-maker interested in offering the same intervention in another state.

For a policy-maker considering a Medicaid expansion in their state, extrapolating these effects raises an external validity concern: (i) the state's eligible population may differ from the experiment along covariates that shape treatment-effect heterogeneity, and (ii) some of these covariates will be unobserved outside the experiment. Regarding (i), Finkelstein et al. (2012) acknowledge substantial demographic differences of the experimental population relative to the US national average: a higher proportion

[14] All replicated results are from the publicly available survey data (Finkelstein, 2013).

[15] The ITT is just an ATE where the treatment D is the "the option to enroll in the health insurance": the robustness approach I discussed carries over to the ITT with only notational changes.

Health-care Utilization	Prescriptions	Out-patient	ER	In-patient
Experimental ITT	0.0259 (0.009)	0.0624 (0.008)	0.0075 (0.007)	0.0029 (0.004)
$\delta^*(0)$	0.054 (0.046)	0.424 (0.152)	0.010 (0.018)	0.003 (0.009)
Preventive care	Cholesterol	Diabetes	Mammogram	Pap test
Experimental ITT	0.0358 (0.008)	0.0282 (0.008)	0.0565 (0.012)	0.0492 (0.010)
$\delta^*(0)$	0.158 (0.073)	0.128 (0.053)	0.173 (0.060)	0.173 (0.047)
Financial Strain	Out-of-pocket	Outstanding	Borrow/Skip	Refused care
Experimental ITT	−0.0579 (0.008)	−0.0518 (0.008)	−0.0433 (0.008)	−0.0104 (0.004)
$\delta^*(0)$	0.327 (0.102)	0.310 (0.103)	0.174 (0.120)	0.030 (0.031)

Table 1: $\delta^*(0)$ robustness metric for the health-care utilization and financial strain outcomes in Finkelstein et al. (2012). All estimates use survey weights.

of Oregon's eligibles are White, older and in poorer health. If these covariates are important determinants of treatment effect heterogeneity, the experiment's results may not be robust to the covariate shifts arising when policy is adopted in other states. Regarding (ii), many survey-specific health covariates, which are likely to be predictive of treatment effect heterogeneity, are exclusively collected in the experiment and have no counterpart in other states. This makes re-weighting procedures like in Hartman (2020) and Hsu et al. (2020) not directly applicable. Instead, I propose to quantify the robustness of each outcome studied in Finkelstein et al. (2012) by reporting my robustness metric δ^*, which measures how large a covariate shift would need to be to eliminate the effects of the expansion in a new state. To operationalize δ^*, I use the same covariates examined by Finkelstein et al. (2012) in their heterogeneity analysis.

Finkelstein et al. (2012)'s exploration of treatment effect heterogeneity focuses on six main covariates: binary indicators for the recipient being White, Female, over the age of 50, having more than High School education, living in an Urban area, and being a Smoker (which the authors use as a proxy for health status). In this context, the joint distribution of the covariates is a 64×1 vector of probabilities, specifying for each combination of the above categories the relative frequency in the sample.[16] Although the authors report being underpowered to make inference on heterogeneous treatment effects for each subgroup, overall treatment-effect heterogeneity is still informative for the external validity of aggregate parameters like the ITT in other states.

Building on this heterogeneity structure, I evaluate the robustness of the aggregate ITT to covariate shifts. Following Finkelstein et al. (2012)'s notation, I consider

[16]The full joint distribution needs to specify $2^6 = 64$ probabilities, summing up to 1.

hypotheses of two forms: (i) $ITT_j > \tilde{\tau}$ for an increase in healthcare utilization or preventive care outcome j; (ii) $ITT_j < \tilde{\tau}$ for a decrease in financial strain outcome j. I report calculations for $\tilde{\tau} = 0$ to target sign robustness: $\delta^*(0)$ measures the covariate shift required to overturn the policy relevant sign of the ITT (non-negative for utilization and non-positive for financial-strain outcomes). Other thresholds can be accommodated by specifying alternative values of $\tilde{\tau}$.[17] For example, $\delta^*(0) = 0.327$ in Table 1 is the smallest KL-distance covariate shift compatible with an increase in "Out-of-pocket expenses". As a descriptive comparison, the result on "Out-of-pocket expenses" is relatively more robust than the one on "prescriptions". The next section provides a cardinal interpretation of the magnitude of $\delta^*(0)$.

4.1 Benchmarking and interpreting robustness

Should a particular value of $\delta^*(0)$ in Table 1 be considered high or low? To answer this question, I propose a benchmarking exercise based on American Community Survey (ACS) census data. I compare the magnitude of δ^* to the KL divergence between the covariate distribution of each U.S. state and the experimental distribution, accounting for lottery eligibility. Eligibility for the Oregon lottery required being currently uninsured, having an income below the poverty line, liquid assets below \$2000 and being a state resident. Using ACS data, I construct a proxy-eligibility rule mirroring these criteria and obtain a proxy-eligible sub-sample in each state. For five out of six heterogeneity covariates, I can construct the same binary indicators as in Finkelstein et al. (2012), they form X_c. The experiment's proxy for health status is not available in the ACS and thus forms X_e. Pooling years 2008-2012, I obtain each state's covariate distribution, and compute their KL-divergences from the experiment using X_c. The results are in Table 2.

Consider the *Out-of-pocket expenses* outcome in Table 1 as an example. The ITT estimate indicates that the option to enroll in OHP reduced the probability of out-of-pocket expenses by 5.79%. The robustness metric $\delta^*(0) = 0.327$ is the smallest shift in the covariate distribution away from the experimental benchmark needed to bring the ITT to zero. Hence for any target state with distribution $G_{X,s}$ if $D_{KL}(G_{X,s} \| F_X) < \delta^*(0)$, the experimental claim ($ITT < 0$) must also hold under G_X. How does $\delta^*(0) = 0.327$ compare to the empirical KL distance benchmarks in Table 2?

[17] For example, setting $\tilde{\tau} = t_j := z_{1-\alpha}\hat{\sigma}_j$ (where $\hat{\sigma}_j$ is the standard error of ITT_j and $z_{1-\alpha}$ is the quantile of the standard normal) targets statistical significance: $\delta^*(t_j)$ measures the covariate shift required for the effect to become statistically non-significant in a one-sided test at level α.

	State	$D_{KL}(G_s\|F)$		State	$D_{KL}(G_s\|F)$		State	$D_{KL}(G_s\|F)$
	Alabama	0.327		Kentucky	0.510		North Dakota	0.490
	Alaska	0.974		Louisiana	0.589	$*\dagger$	Ohio	0.170
$*\dagger$	Arizona	0.217		Maine	0.537	$*$	Oklahoma	0.288
	Arkansas	0.401		Maryland	0.596	$*\dagger\bullet$	**Oregon**	0.127
	California	0.360		Massachusetts	0.511	$*\dagger$	Pennsylvania	0.266
$*\dagger\bullet$	Colorado	0.156	$*\dagger$	Michigan	0.240		Rhode Island	0.482
	Connecticut	0.458	$*\dagger$	Minnesota	0.205		South Carolina	0.429
	Delaware	0.603		Mississippi	1.11		South Dakota	1.18
	DC	1.23	$*\dagger$	Missouri	0.209	$*\dagger$	Tennessee	0.213
$*\dagger$	Florida	0.227		Montana	0.788	$*\dagger$	Texas	0.184
	Georgia	0.517	$*\dagger$	Nebraska	0.208	$*\dagger$	Utah	0.190
	Hawaii	0.738	$*\dagger$	Nevada	0.239		Vermont	1.61
$*$	Idaho	0.320	$*$	New Hampshire	0.293	$*$	Virginia	0.276
	Illinois	0.382		New Jersey	0.654	$*\dagger\bullet$	Washington	0.137
$*\dagger$	Indiana	0.164		New Mexico	0.423		West Virginia	1.09
$*\dagger$	Iowa	0.225		New York	0.410	$*\dagger$	Wisconsin	0.260
$*\dagger$	Kansas	0.234	$*$	North Carolina	0.319		Wyoming	1.54

Table 2: KL divergence between each state distribution G_s and the experimental distribution F, computed from the observed component in Eq.(9) using X_c. I compare these values to $\delta^*(0) = 0.327$ for *Out-of-pocket expenses*. To reason about the unobserved component associated with X_e (health status), I consider hypothetical magnitudes expressed as percentages of the observed KL. Symbols indicate whether the implied total distance remains below 0.327: $*$ (0%), $\dagger$ (20%), and $\bullet$ (100%). The percentages correspond to no unobserved contribution, a contribution proportional to the relative number of covariates in X_e versus X_c, and equal observed and unobserved contributions, respectively.

As a first pass, suppose that the second term of Eq. (9) is zero, so that the reported KL distances based on the census covariates X_c coincide with KL distance over the full X. Under this benchmark, Oregon is the closest state to the experimental covariates.[18] In particular, the Oregon census population has an observable KL distance of 0.127, so $\delta^*(0)$ is about 2.5 times the Oregon-to-experiment distance.

Because health status (X_e) is not observed in the census, the reported KL distances using only X_c provide a lower bound for $D_{KL}(G_X\|F_X)$. The decomposition in Eq. (9) helps us reason about the contribution of the missing covariate X_e to KL by expressing it as a percentage of the observed KL based on X_c. In Table 2, I consider three natural benchmarks: 0% (no contribution), 20% (contribution proportional to the number of variables in X_e vs X_c), and 100% (equal contribution of observed and unobserved component).[19] Twenty-four states remain below the threshold for the 0% benchmark, nineteen states for the 20% benchmark, and Oregon, Colorado and Washington meet the very conservative 100% benchmark.

[18] The two populations are not identical ($KL > 0$) due to pooling census years 2008–2012, survey attrition, and potential self-selection into the lottery (Finkelstein et al., 2012).

[19] Using an observable benchmark to conjecture about an unobservable quantity is reminiscent of the approach in Altonji et al. (2005) Oster (2019) and discussion in (Cinelli and Hazlett, 2020; Masten and Poirier, 2022; Diegert et al., 2025).

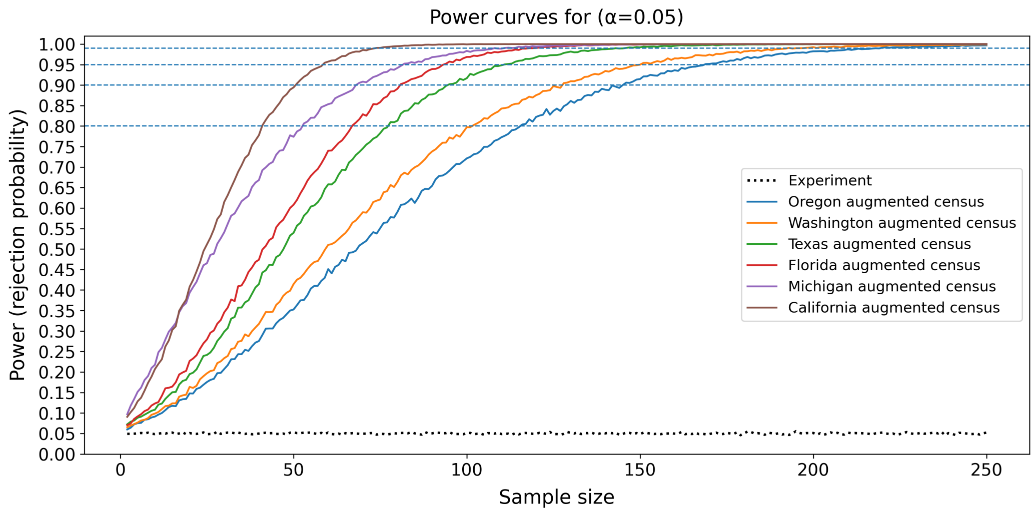

Figure 2: Power against sample size. The X-axis shows m for an i.i.d. sample drawn from G_s for a few census states. The Y-axis shows the rejection probability for $H_0 : G_s = F_X$ against $H_1 : G_s \neq F_X$ using the likelihood ratio test at $\alpha = 0.05$ (20,000 Monte Carlo replications).

4.2 A cardinal interpretation via testing

I now offer an operational calibration of $\delta^*(0)$ that corroborates in what sense it measures statistical distinguishability. Fix a target environment (state) s, and suppose a policymaker could draw an i.i.d. sample of size m of covariates from G_s. The goal is to test whether a shift from the experimental distribution F_X has occurred using a likelihood ratio (LR) test. For concreteness, I focus on one example calibrated to the empirical application where both F and G_s are multinomial vectors with 64 cells. We want to test the hypothesis of no covariate shift $H_0 : G_s = F$ against $H_1 : G_s \neq F$. To do so, we can compare each cell probability (F_j) that we would expect under F to the empirical cell probability observed in the sample ($\hat{G}_{s,j}$) and aggregate over all cells.[20] The resulting test statistic can be written as:

$$T_{s,m} := 2m \sum_{j=1}^{k} \widehat{G}_{s,j} \log\left(\frac{\widehat{G}_{s,j}}{F_j}\right) = 2m\, D_{KL}(\widehat{G}_s \| F),$$

Under H_0, $T_{s,m} \Rightarrow \chi^2_{k-1}$, so we reject at level α when $T_{s,m} > \chi^2_{k-1,1-\alpha}$. Under fixed alternative $G_s \neq F$, $T_{s,m}$ grows linearly in m, with its drift governed by the KL distance. Heuristically, larger $D_{KL}(G_s \| F)$ leads to larger $T_{s,m}$ and therefore higher power at a given sample size. To show this relationship empirically, I use the census covariate distributions. For each state s, I draw m observations from G_s, compute the

[20]Here the empirical distribution $\widehat{G}_s$ is indeed the MLE estimator for G_s.

State	D_{KL}	min m: Power			
		$\geq 80\%$	$\geq 90\%$	$\geq 95\%$	$\geq 99\%$
Oregon census	0.130	116	146	169	220
Washington census	0.141	102	128	151	193
Texas census	0.188	78	95	111	146
Florida census	0.233	67	81	94	118
Michigan census	0.246	53	69	82	108
California census	0.366	41	51	59	74

Table 3: Effective sample size calculation.

rejection frequency of the LR test at $\alpha = 0.05$ (over 20,000 Monte Carlo replications), and plot power as a function of m. Figure 2 shows power curves closely tracking the ordering of $D_{KL}(G_s \| F_X)$ across states over a wide range of sample sizes.

From the simulated power curves, I compute for each state the minimal sample size required to reject H_0 at prescribed power benchmarks (0.8, 0.9, 0.95, 0.99) and report them in Table 3. At each benchmark, states with larger $D_{KL}(G_s \| F_X)$ require fewer observations to detect a shift from F_X, consistent with the interpretation of $D_{KL}(G_s \| F_X)$ as a measure of statistical distinguishability from the experiment. This links directly to the robustness metric. Recall that $\delta^*(0)$ is the smallest KL distance from F_X among covariate shifts that invalidate the experimental claim, attained at F_X^*. Define m^* as the minimal sample size needed to detect the shift F_X to F_X^* at a given power benchmark using the LR test. Then m^* provides an operational calibration: any shift that overturns the conclusion must be at least as detectable as F_X^*, and therefore should require no more than m^* observations to reach the same power. Equivalently, a shift requiring more than m^* observations at that benchmark cannot overturn the experimental sign conclusion.

4.3 Reporting the least-favorable distribution

To complement the interpretation of δ^* in Section 4.1 and to illustrate Theorem 3.2, I report some of the features of the *least-favorable distribution* F_X^* for an example outcome. Recall that F_X^* is a 64×1 probability vector describing the joint distribution of X. I summarize it by reporting the implied marginal proportions of the corresponding indicator variables. Two remarks are in order. First, because F_X^* specifies the full joint distribution of covariates, the marginals in Table 4 provide a coarse but informative summary of the direction of the implied shift. Second, these mean shifts are closely linked to treatment-effect heterogeneity. The least-favorable

Variable	μ (Exp)	μ^* (LFD)	Mean shift
Age50Plus	0.334	0.293	-0.041
Female	0.594	0.584	-0.010
MoreThanHS	0.333	0.279	-0.054
Smoker	0.635	0.538	-0.096
Urban	0.749	0.765	+0.017
White	0.819	0.777	-0.042

Table 4: Marginal means (proportions) under the experimental distribution (μ) and the least-favorable distribution (μ^*), and their mean shifts for the outcome "Out-of-pocket expenses".

distribution shifts weight away from covariate groups (e.g., smokers, older, more educated, and White recipients) for which the conditional ITT is more negative, and toward groups for which the conditional ITT is positive. This reweighting is how the overall ITT can be driven to the threshold $\tilde{\tau} = 0$. The largest shift occurs along the Smoker indicator, which, consistent with Finkelstein et al. (2012)'s use of it as a proxy for health, is strongly predictive of health-related expenses. This is precisely where reporting $\delta^*(0)$ and the benchmarking exercise are most useful: when a heterogeneity-predictive covariate has no census counterpart and threatens external validity, we can assess the magnitude of the minimal invalidating shift relative to observed shifts in covariates that do have census counterparts in plausible implementation environments.

5 Conclusion

To measure the external validity of experimental claims of the form $(ATE > \tilde{\tau})$ under covariate shifts, I propose a robustness metric $\delta^*(\tilde{\tau})$, defined as the distance to the nearest covariate distribution that overturns the claim. Applied researchers can report $\delta^*(\tilde{\tau})$ as a "third number" alongside the point estimate and standard error, and I provide interpretation and calibration exercises that make its magnitude operational. These include benchmarking δ^* against empirically observed covariate shifts across plausible implementation environments and mapping δ^* into the sample size required to detect a shift at a given power benchmark.

Extending the framework to other linear policy parameters is straightforward: just like for the ATE, the closed-form solution yields a simple de-biased GMM procedure that accommodates estimation of the CATE via flexible ML methods (e.g., LASSO, random forests, neural networks). By contrast, for nonlinear distributional parameters, a comparable closed-form characterization and the development of estimation and inference remains an interesting open problem.

A Appendix

A.1 Proof of Theorem 2.1

The proof uses the variational characterization of Donsker and Varadhan (1975).

Lemma 1 (Donsker and Varadhan (1975))**.** *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:*

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

We can now prove Theorem 2.1. By the Radon-Nikodym theorem, $\frac{dF'_X}{dF_X}$ exists and $\operatorname{supp}\left(\frac{dF'_X}{dF_X}\right) \subset \mathcal{X}$. Then, the optimization problem in Eq.4-5 is equivalent to:

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

Proof. i) From Lemma 1 we have:

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

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

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

ii) Because $D_{KL}(F_X^*||F_X^*) = 0$ the value of the minimization problem:

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

$\square$

A.2 Auxiliary lemmas

Proving Theorem 3.1 requires some lemmas. Their derivations are in Spini (2024).

Lemma 2 (Kennedy et al. (2020))**.** *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 3. *For $\psi(\cdot)$ in Equation (12) and $\bar{\psi}(\theta, \gamma, \alpha) = \mathbb{E}[\psi(w, \theta, \gamma, \alpha)]$ we have:*

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

Lemma 4 (Jacobian consistency)**.** *For the Jacobian G defined as:*

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

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

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

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

A.3 Proof of Theorem 3.1

Denote $\hat{G} = \frac{\partial\hat{\psi}(w,\hat{\theta},\hat{\gamma})}{\partial\theta}$. First note that by Lemma 4 we have $\|\hat{G}^{-1} - G^{-1}\| = o_P(1)$. Now by the central limit theorem and Lemma 5 we have:

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

Finally a standard GMM Taylor linearization gives the desired result:

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

A.4 Proof of theorem 3.2

The proof of Theorem 3.2 follows the same structure of Theorem 3.1 and is omitted.

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. Athey, G. W. Imbens, J. Metzger, and E. Munro. Using wasserstein generative adversarial networks for the design of monte carlo simulations. *Journal of Econometrics*, 240(2):105076, 2024.

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, 2018a.

V. Chernozhukov, I. Fernández-Val, and Y. Luo. The sorted effects method: Discovering heterogeneous effects beyond their averages. *Econometrica*, 86(6):1911–1938, 2018b.

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.

T. M. Cover. *Elements of information theory*. John Wiley & Sons, 1999.

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

P. Diegert, M. A. Masten, and A. Poirier. An axiomatic approach to comparing sensitivity parameters. *arXiv preprint arXiv:2504.21106*, 2025.

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.

F. F. Gunsilius. Distributional synthetic controls. *Econometrica*, 91(3):1105–1117, 2023.

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.

R. Koenker. *Quantile Regression*. Number 38 in Econometric Society Monographs. Cambridge University Press, Cambridge, 2005. ISBN 9780521845731.

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.

M. A. Masten and A. Poirier. The effect of omitted variables on the sign of regression coefficients. *arXiv preprint arXiv:2208.00552*, 2022.

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.

A. Rambachan and J. Roth. A more credible approach to parallel trends. *Review of Economic Studies*, page rdad018, 2023.

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.

P. E. Spini. Robustness, heterogeneous treatment effects and covariate shifts. Working paper, 2024.

R. V. Tabri. A sieve m-estimator for entropic optimal transport. arXiv preprint, 2026. Working paper.

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.