An Automatic Finite-Sample Robustness Metric: Can Dropping a Little Data Change Conclusions?

Tamara Broderick, Ryan Giordano, and Rachael Meager *

December 1, 2020

Abstract

We propose a method to assess the sensitivity of econometric analyses to the removal of a small fraction of the sample. Analyzing all possible data subsets of a certain size is computationally prohibitive, so we provide a finite-sample metric to approximately compute the number (or fraction) of observations that has the greatest influence on a given result when dropped. We call our resulting metric the Approximate Maximum Influence Perturbation. Our approximation is automatically computable and works for common estimators (including OLS, IV, GMM, MLE, and variational Bayes). We provide explicit finite-sample error bounds on our approximation for linear and instrumental variables regressions. At minimal computational cost, our metric provides an exact finite-sample lower bound on sensitivity for any estimator, so any non-robustness our metric finds is conclusive. We demonstrate that the Approximate Maximum Influence Perturbation is driven by a low signal-to-noise ratio in the inference problem, is not reflected in standard errors, does not disappear asymptotically, and is not a product of misspecification. Several empirical applications show that even 2-parameter linear regression analyses of randomized trials can be highly sensitive. While we find some applications are robust, in others the sign of a treatment effect can be changed by dropping less than 1% of the sample even when standard errors are small.

*We thank Avi Feller, Jesse Shapiro, Michael Kremer, Peter Hull, Tetsuya Kaji, Heather Sarsons, Kirill Borusyak and the authors of all of our applications for their insightful comments and suggestions. All mistakes are our own. Corresponding Author: Rachael Meager, reachable at `r.meager@lse.ac.uk`.

1 Introduction

Ideally, policymakers will use economics research to inform decisions that affect people's livelihoods, health, and well-being. Yet study samples may differ from the target populations of these decisions in non-random ways, perhaps because of practical challenges in obtaining truly random samples without any missing or compromised data, or because populations generally differ across time and geographic locations. When these deviations from the ideal random sampling exercise are small, one might think that the conclusions from economic studies would still hold in the populations affected by policy. It therefore seems prudent to ask whether a small percentage of a study's sample — or a handful of data points — has been instrumental in determining its findings. In this paper we provide a finite-sample, automatically-computable metric of how dropping a small amount of data can change empirical conclusions. We show that the sensitivity to potentially non-random deviations from the target population is not captured by conventional standard errors. We show that certain empirical results from high-profile studies in economics can be reversed by removing less than 1% of the sample even when standard errors are small, and we investigate why.

There are several reasons to care about whether empirical conclusions are substantially influenced by small percentages of the finite sample. In practice, even if we can sample from the population of direct interest, small percentages of our intended random samples are often missing; either surveyors and implementers cannot find them, or they refuse to answer our questions or follow our instructions ("suggestions"), or their answers get lost or garbled during data processing. As this missingness cannot safely be assumed random, researchers might like to know how threatening this non-random data leakage could be to their analysis, i.e. whether their substantive conclusions could conceivably be overturned by that missing handful of data points. Similarly, consumers of research who are concerned about potentially non-random errors in sample construction at any stage of the analysis might be interested in this metric as a measure of the exposure of a study's conclusions to this concern. Conclusions that are highly influenced by a small handful of data points are more exposed to adverse events during data analysis, including p-hacking, even if these errors are unintentional.

Yet even if researchers could construct a perfectly random sample from a given study population, in practice we will never see that population again. The target population for our policy decisions is always different from the study population, if only because the world may change in the time between the research and the decision. And social scientists often aspire to uncover generalizable truths about the world and to make policy recommendations that would apply more broadly than to a single study population.

In this paper, then, we propose to directly measure the extent to which a small

fraction of a data sample has influenced the central claims or conclusions of a study. For a particular fraction α (e.g., $\alpha = 0.001$), we propose to find the set of no more than $100\alpha\%$ of all the observations that effects the greatest change in an estimator when those observations are removed from the sample — and to report this change. For example, suppose we were to find a statistically-significant average increase in household consumption after implementing some economic policy intervention. Further suppose that, by dropping 0.1% of the sample (often fewer than 10 data points), we instead find a statistically-significant average *decrease* in consumption. Then it would be challenging to argue that this intervention would yield consumption increases in even slightly different populations.

To quantify this sensitivity, one could consider every possible fraction $1 - \alpha$ of the data, and re-run the original analysis on all of these data subsets. But this direct implementation is computationally prohibitive.[1] Consider a small data set with $N = 100$ data points and $\alpha = 0.05$. If the original data analysis took one second to run, this proposal would take over 871 days to run; see Section 2 for more detail. The computation is dramatically more expensive in larger data sets. So instead of directly re-running the data analysis over these data subsets, we instead propose to use an approximation. Our approximation works for common estimators — including GMM, OLS, IV, and MLE. Roughly, we give each data point a weight and apply a Taylor expansion in the weights (Section 2.1 and Section 2.2). We show that our approximation is fast, automatable, and easy to use (Section 2.4). In particular, our approximation does not require running any additional data analyses beyond the original.

We use theoretical analyses, simulation studies, and applied examples to show that our approximation works. We provide exact, calculable finite sample bounds on performance for OLS and IV estimators, and we show that the approximation error is low when the percentage of the sample removed is small (Section 3.2.2). Moreover, for the cost of a single additional data analysis, we can provide an exact lower bound on the worst-case change in an analysis upon removing $100\alpha\%$ of the data (Section 3.2.1). We check that our metric flags combinations of data points that reverse empirical conclusions when removed (Section 4). For example, in the Oregon Medicaid study (Finkelstein et al., 2012), we can identify a subset containing less than 1% of the original data that controls the sign of the effects of Medicaid on certain health outcomes. In the Mexico microcredit study (Angelucci et al., 2015), we find a single observation, out of 16,500, that controls the sign of the ATE on household profit.

We investigate the source of this sensitivity when it arises, and we show that it is not captured in conventional standard errors. We find that a result's exposure to the influence of a small fraction of the sample need not reflect a model misspecification

[1] Indeed, Young (2019) finds it computationally prohibitive to re-run their analysis when leaving out every possible subset of two data points.

problem in the classical sense. Sensitivity according to our metric can arise even if the model is exactly correct and the data set large, if there is a low *signal-to-noise ratio*: that is, if the strength of the claim (signal) is small relative to the size of an estimator of the standard deviation in the asymptotic distribution of the quantity of interest (noise) (Section 3.1.1). In OLS this "noise" is large when we have data points exhibiting a combination of high leverage and large $\hat{\epsilon}_n$ (Section 3.1.6). This noise can be large even when standard errors are small, because unlike standard errors it does not disappear as N grows large (Section 3.1.4) and because it is influenced by distributional shape while standard errors reflect only distributional scale (Section 3.1.3).

We examine several applications from empirical economics papers and find that the sensitivity captured by our metric varies considerably across analyses in practice. We find that certain results across the applications we examine are robust up to 5% and even 10% removal. But we also find cases where the sign and significance of certain estimated treatment effects can be reversed by dropping less than 1% of the sample. We sometimes see this reversal even when the t-statistics are very large and inference is very precise, as in the Oregon Medicaid RCT (Finkelstein et al., 2012), in Section 4.1. In Section 4.2, we show that trimming outliers in the outcome data does not necessarily reduce sensitivity by examining the Progresa Cash Transfers RCT (Angelucci and De Giorgi, 2009). In Section 4.4, we show that the Bayesian approach does not necessarily eliminate this sensitivity by examining a Bayesian hierarchical analysis of seven Microcredit RCTs (Meager, 2020).

We recommend that researchers compute and report our metric as a complement to standard errors but also to other robustness checks. For instance, since our approximation is fundamentally local due to the Taylor expansion, practitioners may also consider global sensitivity checks such as those proposed by Leamer (1984, 1985); Sobol (2001); Saltelli (2004) or the breakdown frontiers approach of He et al. (1990); Masten and Poirier (2020). Our method is also no substitute for tailored robustness checks designed by researchers to investigate specific concerns about sensitivity of results to certain structures or assumptions. And practitioners may benefit from robustifying their analysis (Mosteller and Tukey, 1977; Hansen and Sargent, 2008; Chen et al., 2011) even if they pass our check. Our metric is also complementary to classical robustness measures, although we are able to connect our metric to these measures via the influence function. By interpreting our metric as a seminorm on the empirical influence function in Section 3.3.3, we can contrast with the gross error sensitivity, which underlies more classical robustness notions, in Section 3.3.4. We see that gross error sensitivity is set up for designing estimators and arbitrary adversarial perturbations to the population distribution, whereas our metric is set up for assessing sensitivity to dropping a small subset of the data at hand once an analysis has been performed. We do not yet recommend any specific alterations to common inferential procedures based on our metric, but we believe

this direction is promising for future research.

2 A proposed metric

Suppose we observe N data points $d_1, \ldots, d_N$. For instance, in a regression problem, the nth data point might consist of covariates x_n and response(s) y_n, with $d_n = (x_n, y_n)$. Consider a parameter $\theta \in \mathbb{R}^P$ of interest. Typically we estimate θ via some function $\hat{\theta}$ of our data. The central claim of an empirical economics paper is typically focused on some attribute of θ, such as the sign or significance of a particular effect or quantity. A frequentist analyst might be worried if removing some small fraction α of the data were to

- Change the sign of an effect.
- Change the significance of an effect.
- Generate a significant result of the opposite sign.

To capture these concerns, we define the following quantities:

Definition 1. Let the *Maximum Influence Perturbation* be the largest possible change induced in the quantity of interest by dropping no more than $100\alpha\%$ of the data.

We will often be interested in the set that achieves the Maximum Influence Perturbation, so we call it the *Most Influential Set*.

And we will be interested in the minimum data proportion $\alpha \in [0, 1]$ required to achieve a change of some size Δ in the quantity of interest, so we call that α the *Perturbation-Inducing Proportion*. We report `NA` if no such α exists.

In general, to compute the Maximum Influence Perturbation for some α, we would need to enumerate every data subset that drops no more than $100\alpha\%$ of the original data. And, for each such subset, we would need to re-run our entire data analysis. If m is the greatest integer smaller than 100α, then the number of such subsets is larger than $\binom{N}{m}$. For $N = 100$ and $m = 5$, $\binom{N}{m} = 75{,}287{,}520$. So computing the Maximum Influence Perturbation in even this simple case requires re-running our data analysis over 75 million times. If each data analysis took 1 second, computing the Maximum Influence Perturbation would take over 871 days to compute. Indeed, the Maximum Influence Perturbation, Most Influential Set, and Perturbation-Inducing Proportion may all be computationally prohibitive even for relatively small analyses.

2.1 Setup: Notation and Assumptions

To address this computational issue, we propose to use a (fast) approximation instead. We will see, for the cost of one additional data analysis, our approximation

can provide a lower bound on the exact Maximum Influence Perturbation. More generally we provide theory and experiments to support the quality of our approximation. We provide open-source code and show that our approximation is fully automatable in practice (Section 2.4).

Our approximation is akin to a Taylor expansion, so it will require certain aspects of our estimator to be differentiable. We summarize our assumptions here, and we note that many common analyses satisfy these assumptions — including, but not limited to, OLS, IV, GMM, MLE, and variational Bayes.

Assumption 1. *$\hat{\theta}$ is a* Z-estimator*; that is, $\hat{\theta}$ is the solution to the following estimating equation,*[2] *where $G(\cdot, d_n) : \mathbb{R}^P \to \mathbb{R}^P$ is a twice continuously differentiable function and 0_P is the column vector of P zeros.*

$$\sum_{n=1}^{N} G(\hat{\theta}, d_n) = 0_P. \tag{2.1}$$

Assumption 2. *$\phi : \mathbb{R}^P \to \mathbb{R}$, which we interpret as a function that takes the full parameter θ and returns the quantity of interest from θ, is continuously differentiable.*

For instance, the function that picks out the p-th effect from the vector θ, $\phi(\theta) = \theta_p$, satisfies this assumption.

To form our approximation, we introduce a vector of data weights, $\vec{w} = (w_1, \ldots, w_N)$, where w_n is the weight for the nth data point. We recover the original data set by giving every data point a weight of 1: $\vec{w} = \vec{1} = (1, \ldots, 1)$. We can denote a subset of the original data as follows: start with $\vec{w} = \vec{1}$; then, if the data point indexed by n is left out, set $w_n = 0$. We can collect weightings corresponding to all data subsets that drop no more than $100\alpha\%$ of the original data as follows:

$$W_\alpha := \{\vec{w} : \text{No more than } \lfloor \alpha N \rfloor \text{ elements of } \vec{w} \text{ are 0 and the rest are 1}\}.$$

Our main idea will be to form a Taylor expansion of our quantity of interest ϕ as a function of the weights, rather than recalculate ϕ for each data subset (i.e., for each reweighting).

To that end, we first reformulate our setup, now with the weights $\vec{w}$; note that we recover the original problem (for the full data) above by setting $\vec{w} = \vec{1}$ in what follows. Let $\hat{\theta}(\vec{w})$ be our parameter estimate at the weighted data set described by $\vec{w}$. Namely, $\hat{\theta}(\vec{w})$ is the solution to the weighted estimating equation

$$\sum_{n=1}^{N} w_n G(\hat{\theta}(\vec{w}), d_n) = 0_P. \tag{2.2}$$

[2]Sometimes Eq. 2.1 is associated with "M-estimators" that optimize a smooth objective function, since those M-estimators typically take the form of a Z-estimator. However, some Z-estimators, such as instrumental variables regression, do not optimize any particular empirical objective function, so the notion of Z-estimator is in fact more general.

We allow that the quantity of interest ϕ may depend on $\vec{w}$ not only via the estimator θ, so we write $\phi(\theta, \vec{w})$ with $\phi(\cdot, \cdot) : \mathbb{R}^P \times \mathbb{R}^N \to \mathbb{R}$ and will use the shorthand $\phi(\vec{w}) := \phi(\hat{\theta}(\vec{w}), \vec{w})$. We require that $\phi(\cdot, \cdot)$ be continuously differentiable in both its arguments. For instance, $\phi(\theta, \vec{w}) = \theta_p$ to pick out the p-th component of θ. Or, to consider questions of statistical significance, we may choose $\phi(\theta, \vec{w}) = \theta_p + 1.96\sigma_p(\theta, \vec{w})$, where $\sigma_p(\theta, \vec{w})$ is an estimate of the standard error depending smoothly on θ and $\vec{w}$; this example is our motivation for allowing the more general $\vec{w}$ dependence in $\phi(\theta, \vec{w})$.

With this notation in hand, we can restate our original goal as solving

$$\vec{w}^{**} := \underset{\vec{w} \in W_\alpha}{\arg\max} \left(\phi(\vec{w}) - \phi(\vec{1}) \right). \tag{2.3}$$

Here we focus on positive changes in ϕ since negative changes can be found by reversing the sign of ϕ and using $-\phi$ instead. In particular, the non-zero indices of $\vec{w}^{**}$ correspond to the Most Influential Set: $S_\alpha := \{n : \vec{w}_n^{**} = 0\}$. And $\Psi_\alpha = \phi(\vec{w}^{**}) - \phi(\vec{1})$ is the Maximum Influence Perturbation. The Perturbation Inducing Proportion is the smallest α that induces a change of at least size Δ: $\alpha_\Delta^* := \inf\{\alpha : \Psi_\alpha > \Delta\}$.

2.2 A Tractable Approximation

Our approximation, then, centers on a first-order Taylor expansion (and thus linear approximation) in $\phi(\vec{w})$ around $\vec{w} = \vec{1}$:

$$\phi(\vec{w}) \approx \phi^{\text{lin}}(\vec{w}) := \phi(\vec{1}) + \sum_{n=1}^{N} (w_n - 1)\psi_n, \text{ with } \psi_n := \left. \frac{\partial \phi(\vec{w})}{\partial w_n} \right|_{\vec{w} = \vec{1}}. \tag{2.4}$$

We can in turn approximate the Most Influential Set as follows.

$$\vec{w}^{**} \approx \vec{w}^{*} := \underset{\vec{w} \in W_\alpha}{\arg\max} \left(\phi^{\text{lin}}(\vec{w}) - \phi(\vec{1}) \right) \tag{2.5}$$

$$= \underset{\vec{w} \in W_\alpha}{\arg\max} \sum_{n=1}^{N} (w_n - 1)\psi_n = \underset{\vec{w} \in W_\alpha}{\arg\max} \sum_{n:\, w_n = 0} (-\psi_n). \tag{2.6}$$

To compute $\vec{w}^*$ (analogous to the $\vec{w}^{**}$ that determines the exact Most Influential Set), we compute ψ_n for each n. Then we choose $\vec{w}^*$ to have entries equal to zero at the $\lfloor \alpha N \rfloor$ indices n where ψ_n is most negative (and to have entries equal to one elsewhere). Analogous to the Perturbation Inducing Proportion, we can find the minimum data proportion α required to achieve a change of some size Δ: i.e., $\phi^{\text{lin}}(\vec{w}) - \phi(\vec{1}) > \Delta$. In particular, we iteratively remove the most negative ψ_n (and the index n) until the Δ change is achieved; if the number of removed points is M, the proportion we report is $\alpha = M/N$. Recall that finding the exact Maximum Influence Perturbation, Most Influential Set, and Perturbation-Inducing Proportion required running a data analysis more than $\binom{M}{\lfloor \alpha N \rfloor}$ times. By contrast, our approx-

imation requires running just the single original data analysis, N additional fast calculations to compute each ψ_n, and finally a sort on the ψ_n values.

We define our approximate quantities, as detailed immediately above, as follows.

Definition 2. The *Approximate Most Influential Set* is the set $\hat{S}_\alpha$ of at most $100\alpha\%$ data indices that, when left out, induce the biggest approximate change $\phi^{\text{lin}}(\vec{w}) - \phi(\vec{1})$; i.e., it is the set of data indices left out by $\vec{w}^*$: $\hat{S}_\alpha := \{n : \vec{w}^*_n = 0\}$.

The *Approximate Maximum Influence Perturbation* $\hat{\Psi}_\alpha$ is the approximate change observed at $\vec{w}^*$: $\hat{\Psi}_\alpha := \phi^{\text{lin}}(\vec{w}^*) - \phi(\vec{1})$.

The *Approximate Perturbation Inducing Proportion* $\hat{\alpha}^*_\Delta$ is the smallest α needed to cause the approximate change $\phi^{\text{lin}}(\vec{w}) - \phi(\vec{1})$ to be greater than Δ. That is, $\hat{\alpha}^*_\Delta := \inf\{\alpha : \hat{\Psi}_\alpha > \Delta\}$. We report `NA` if no $\alpha \in [0, 1]$ can effect this change.

Moreover, for the cost of a single additional data analysis, we can compute $\phi(\vec{w}^*)$ exactly — and therefore we can compute $\phi(\vec{w}^*) - \phi(\vec{1})$, which forms a lower bound on the exact Maximum Influence Perturbation.

2.2.1 Calculating the influence scores

To finish describing our approximation, it remains to detail how to compute $\psi_n = \frac{\partial\phi(\vec{w})}{\partial w_n}\Big|_{\vec{w}=\vec{1}}$ from Eq. 2.4. We will refer to the quantity $\frac{\partial\phi(\vec{w})}{\partial w_n}\Big|_{\vec{w}}$ as the *influence score* of data point n for ϕ at $\vec{w}$, since it is the *empirical influence function* evaluated at the datapoint d_n. We further discuss the connection with influence functions in Section 3.3 and here detail how to compute the ψ_n as part of our approximation. First, the chain rule gives

$$\frac{\partial\phi(\vec{w})}{\partial w_n}\Big|_{\vec{w}} = \frac{\partial\phi(\theta, \vec{w})}{\partial\theta^T}\Big|_{\hat{\theta}(\vec{w}),\vec{w}} \frac{\partial\hat{\theta}(\vec{w})}{\partial w_n}\Big|_{\vec{w}} + \frac{\partial\phi(\theta, \vec{w})}{\partial w_n}\Big|_{\hat{\theta}(\vec{w}),\vec{w}}. \tag{2.7}$$

The derivatives of $\phi(\cdot, \cdot)$ can be calculated using automatic differentiation software such as Python's `autograd` library (Maclaurin et al., 2015; Baydin et al., 2017). And once we have $\hat{\theta}(\vec{1})$ from running the original data analysis, we can evaluate these derivatives at $\vec{w} = \vec{1}$: e.g., $\frac{\partial\phi(\theta,\vec{w})}{\partial\theta^T}\Big|_{\hat{\theta}(\vec{1}),\vec{w}=\vec{1}}$.

The term $\frac{\partial\hat{\theta}(\vec{w})}{\partial w_n}\Big|_{\vec{w}=\vec{1}}$ requires slightly more work since $\hat{\theta}(\vec{w})$ is defined implicitly. We follow standard arguments from the statistics and mathematics literatures (Krantz and Parks, 2012; Hampel, 1974) to show how to calculate it below.

Start by considering the more general setting where $\hat{\theta}(\vec{w})$ is the solution to the equation $\gamma(\hat{\theta}(\vec{w}), \vec{w}) = 0_P$. We assume $\gamma(\cdot, \vec{w})$ is continuously differentiable with full-rank Jacobian matrix; then the derivative $\frac{\partial\hat{\theta}(\vec{w})}{\partial w_n}\Big|_{\vec{w}}$ exists by the implicit function theorem (Krantz and Parks, 2012, Theorem 3.3.1). We can thus use the chain rule

and solve for $\left.\frac{\partial\hat{\theta}(\vec{w})}{\partial w_n}\right|_{\vec{w}}$; here $0_{P\times N}$ is the $P \times N$ matrix of zeros.

$$0_{P\times N} = \left.\frac{\mathrm{d}\gamma(\hat{\theta}(\vec{w}),\vec{w})}{\mathrm{d}\vec{w}^T}\right|_{\hat{\theta}(\vec{w}),\vec{w}} = \left.\frac{\partial\gamma(\theta,\vec{w})}{\partial\theta^T}\right|_{\hat{\theta}(\vec{w}),\vec{w}} \left.\frac{\mathrm{d}\hat{\theta}(\vec{w})}{\mathrm{d}\vec{w}^T}\right|_{\vec{w}} + \left.\frac{\partial\gamma(\theta,\vec{w})}{\partial\vec{w}^T}\right|_{\hat{\theta}(\vec{w}),\vec{w}} \tag{2.8}$$

$$\Rightarrow \left.\frac{\mathrm{d}\hat{\theta}(\vec{w})}{\mathrm{d}\vec{w}^T}\right|_{\vec{w}} = -\left(\left.\frac{\partial\gamma(\theta,\vec{w})}{\partial\theta^T}\right|_{\hat{\theta}(\vec{w}),\vec{w}}\right)^{-1} \left.\frac{\partial\gamma(\theta,\vec{w})}{\partial\vec{w}^T}\right|_{\hat{\theta}(\vec{w}),\vec{w}}, \tag{2.9}$$

where we can take the inverse by our full-rank assumption.

When we apply the general setting above to our special case $\gamma(\theta,\vec{w}) = \sum_{n=1}^{N} w_n G(\theta, d_n)$, we find

$$\left.\frac{\mathrm{d}\hat{\theta}(\vec{w})}{\mathrm{d}\vec{w}^T}\right|_{\vec{w}} = -\left(\sum_{n=1}^{N} w_n \left.\frac{\partial G(\theta,d_n)}{\partial\theta^T}\right|_{\hat{\theta}(\vec{w}),\vec{w}}\right)^{-1} \left(G(\hat{\theta}(\vec{w}),d_1),\ldots,G(\hat{\theta}(\vec{w}),d_N)\right), \tag{2.10}$$

which can again be computed with automatic differentiation software.

2.3 An OLS regression

Before continuing, we illustrate our method with an example. Econometricians often analyze causal relationships — the focus of applied microeconomics — using linear regressions estimated via ordinary least squares (OLS). When using OLS, a researcher rarely believes the conditional mean dependence is truly linear. Rather, researchers use linear regression since it allows transparent and straightforward estimation of an average treatment effect or local average treatment effect. Researchers often invoke the law of large numbers to justify the focus on the sample mean. They invoke the central limit theorem to justify the use of Gaussian confidence intervals — even in the absence of a finite-sample Gaussianity assumption on the regression errors. In the remainder of this section we show that just a single data point can have outsize influence on regression parameters in the finite sample even when the full sample is large.

In view of this central methodology, consider linear mean regression of some outcome y_n on some explanatory variable x_n estimated via OLS. Suppose for simplicity that these variables have been demeaned and that relevant other factors have been partialed out of y_n. We use the model $y = \theta^T x + \epsilon$ and take $d_n = (x_n, y_n)$. Then the (weighted) OLS estimate $\hat{\theta}(w)$ solves the weighted estimating equation (Eq. 2.2) with

$$G(\theta, d_n) = x_n(y_n - \theta^T x_n). \tag{2.11}$$

It follows that the (weighted) OLS estimate is

$$\hat{\theta}(w) = \left(\sum_{n=1}^{N} w_n x_n x_n^T \right)^{-1} \sum_{n=1}^{N} w_n y_n x_n. \tag{2.12}$$

Applying Eq. 2.10 yields

$$\begin{aligned} \left. \frac{\mathrm{d}\hat{\theta}(\vec{w})}{\mathrm{d}w_n} \right|_{\vec{w}=\vec{1}} &= - \left(\sum_{m=1}^{N} w_m x_m x_m^T \right)^{-1} w_n x_n (y_n - \hat{\theta}(\vec{w})^T x_n) \Bigg|_{\vec{w}=\vec{1}} \\ &= - \left(\sum_{m=1}^{N} x_m x_m^T \right)^{-1} x_n (y_n - \hat{\theta}(\vec{1})^T x_n) \end{aligned} \tag{2.13}$$

One might expect that in large samples there ought not to be a small number of data points that wholly determine the results of an OLS regression. We now show that this intuition is misplaced: ψ_n can be very large in practice. Consider as an example the set of seven randomized controlled trials of expanding access to microcredit discussed by Meager (2019). For illustrative purposes we single out the study with the largest sample size: Angelucci et al. (2015). This study has approximately 16,500 households. A full treatment of all seven studies is in Sections 4.3 and 4.4 along with tables and figures of the results discussed below.

We consider the headline results on household business profit regressed on an intercept and a binary variable indicating whether a household was allocated to the treatment group or to the control group. Let Y_{ik} denote the profit measured for household i in site k, and let T_{ik} denote their treatment status. We first estimate the following model via ordinary least squares:

$$Y_{ik} = \beta_0 + \beta_1 T_{ik} + \epsilon_{ik}. \tag{2.14}$$

We confirm the main findings of the study in estimating an average treatment effect (ATE) of -4.55 USD PPP per 2 weeks, with a standard error of 5.88. Here, our parameter is $\theta = (\beta_0, \beta_1)$. We are interested in whether we can change the sign of β_1 from negative to positive, so we take $\phi(\theta) = \beta_1$. We then compute ψ_n for each data point in the sample, which takes less than 2 seconds in R via Python using our implementation below.

Examining $\vec{\psi}$, we find that one household has $\psi_n = 4.95$; removing that single household should flip the sign if the approximation is accurate. In this case we can manually remove the data point and re-run the regression. We indeed find that the ATE is now 0.4 with a standard error of 3.19. Moreover, by removing 15 households we can generate an ATE of 7.03 with a standard error of 2.55: a significant result of the opposite sign. These results and comparable analyses for other microcredit RCTs are presented in Section 4.3.

How is it possible for the absence of a single household to flip the sign of an estimate that was ostensibly based on all the information from a sample of 16,500?

The fact that the original estimate was statistically insignificant plays a role here, but it is not decisive, and we find examples of statistically significant results that can be overturned by removing less than 1% of the sample in Section 4.1 and Section 4.2. One might also suspect that this sensitivity arises because sample means are non-robust in the Huber sense, but using sample means is not decisive either: we find applications in which it is necessary to remove more than 10% of the sample to change the sign, and we can simulate cases in which no amount of removal will change the sign (Section 3.1.2). We also investigate the results of fitting a Bayesian hierarchical model with a more realistic data-generating process to the set of seven experiments and find that taking a Bayesian approach does not resolve the sensitivity either (Section 4.4). Instead, we will see that both our theory and simulations suggest that the major determinant of this sensitivity is the *signal-to-noise ratio* in the data (Section 3.1.1).

2.4 Automated implementation

We provide an open-source software implementation in R using Python's automatic differentiation capacity under the hood. Our implementation automatically computes the Approximate Most Influential Set, Approximate Maximum Influence Perturbation, and the Approximate Perturbation Inducing Proportion for a range of problems. The most computationally expensive part of computing $\vec{\psi}$ is usually computing the partial derivative $\sum_{n=1}^{N} \partial G(\theta, d_n)/\partial\theta$ and its inverse, but this computation is common to all functions ϕ. So these values need only be computed once to investigate a wide range of quantities of interest.

Our package is available on Github in the repository `rgiordan/zaminfluence`. Currently, we handle OLS and IV regression fully automatically, including weighted versions and robust or clustered standard errors. The package can handle general Z-estimators if the user provides a Python implementation of the estimating equation G and functions of interest ϕ. We also facilitate automatically re-running a regression after removing the Approximate Most Influential Set. Thus, the user can obtain a lower bound on the exact Maximum Influence Perturbation or check the quality of the approximation in their application.

To illustrate the ease of use of the package, consider the microcredit example from the previous section. Suppose we have a linear regression for which we care about the estimated coefficient on a variable called "treatment." Once the researcher has run the regression using R's `lm()` function and defined the resulting object as e.g. `reg_fit`, the user need only run the following to compute and rank the $\vec{\psi}$, where

the function of interest is the "treatment" regressor.

```
reg_influence <- ComputeModelInfluence(reg_fit)
     grad_df <- GetTargetRegressorGrads(reg_influence, "treatment")
influence_dfs <- SortAndAccumulate(grad_df)
```

The object `influence_dfs` produces all of the following: a dataframe of the influence scores, the associated change in the coefficient from both individual removal and cumulative removal in rank order, the row at which one can locate the data point in the original dataset, and other metrics. If the researcher is specifically interested in how many data points she needs to remove to change the sign of the result, the significance of the result, or to generate a result of the opposite sign, she would then run the following to produce a table of results similar to those we show in our applications.

```
GetRegressionTargetChange(influence_dfs, "num_removed")
```

Further details and options for these functions can be found at the online repository.

2.5 Example functions of interest

We end this section with some concrete examples of functions of interest. Recall from the start of Section 2 that we are often interested in whether we can change the sign, significance, or both sign and significance of an estimator. Figure 3 illustrates how we might choose ϕ to answer these questions given an estimator $\hat{\theta}$ and an estimate of the standard error, $\hat{\sigma}/\sqrt{N}$.

Suppose we are interested in the p-th component of $\hat{\theta}$ and that, as in the Figure 3, $\hat{\theta}_p$ is positive and statistically significant.

To make $\hat{\theta}_p$ change sign, we can take

$$\phi(\theta, \vec{w}) = -\,\theta_p. \qquad \text{(Change sign)} \tag{2.15}$$

We use $-\theta_p$ instead of θ_p since we have defined ϕ as a function that we are trying to increase (cf. Eq. 2.3 and the discussion after). Increasing $\phi(\hat{\theta}_p, \vec{1})$, for ϕ in Eq. 2.15, by an amount $\Delta = \hat{\theta}_p$ is equivalent to $\hat{\theta}_p$ changing sign from positive to negative.

To make $\hat{\theta}_p$ statistically insignificant, we wish to take the lower bound of the confidence interval to 0. To that, we can take

$$\phi(\theta, \vec{w}) = -\left(\theta_p - \frac{1.96}{\sqrt{N}}\hat{\sigma}_p(\theta, \vec{w})\right). \qquad \text{(Change significance)} \tag{2.16}$$

As in the previous case, we choose Eq. 2.16 with a leading negative sign because we have defined ϕ as a function that we are trying to increase (cf. Eq. 2.3). Increasing

$\phi(\hat{\theta}_p, \vec{1})$, for ϕ in Eq. 2.16, by an amount $\Delta = \hat{\theta}_p - \frac{1.96}{\sqrt{N}}\hat{\sigma}_p$ is equivalent to $\hat{\theta}_p$ becoming statistically insignificant.

Similarly, we can effect a change in both sign and significance by taking

$$\phi(\theta, \vec{w}) = -\left(\theta_p + \frac{1.96}{\sqrt{N}}\hat{\sigma}_p(\theta, \vec{w})\right) \qquad \text{(Change sign and significance)}$$

and $\Delta = \hat{\theta}_p + \frac{1.96}{\sqrt{N}}\hat{\sigma}_p$.

We have written $\hat{\sigma}_p(\theta, \vec{w})$ to emphasize that standard errors are typically given as functions of θ and the weights $\vec{w}$. The sandwich covariance matrix described in Section 3.1.1 is one such example. There are generally several reasonable ways one might define the dependence of standard errors on the weights. For example, one might reasonably define the standard error to be $\hat{\sigma}_p(\theta, \vec{w})/\sqrt{\sum_{n=1}^{N} w_n}$ rather than $\hat{\sigma}_p(\theta, \vec{w})/\sqrt{N}$. See Section 3.1.1 for further discussion of this subtle point.

3 Underlying theory and interpretation

We next provide the theory to support and understand the behavior of the Approximate Maximum Influence Perturbation. First, in Section 3.1, we provide intuition for which aspects of the data and inference problem determine when and how a small fraction of the sample can have a large influence on empirical conclusions detectable by the Approximate Maximum Influence Perturbation. In Section 3.2, we check the quality of our approximation and discuss cases where the approximation may struggle (Section 3.2.3). Finally, we connect the Approximate Maximum Influence Perturbation to existing work on influence functions in Section 3.3.

3.1 What determines robustness?

We next explore a number of natural hypotheses about what might or might not determine robustness according to the Approximate Maximum Influence Perturbation. Throughout this section, we will often just write "robustness" as shorthand for robustness under our metric. For simplicity, in this section we will consider only functions of interest ϕ that do not depend explicitly on the weights, i.e., for which all of the weight dependence is through the optimal parameter $\hat{\theta}(\vec{w})$. From simulation experiments, we will see that the signal-to-noise ratio drives robustness (Section 3.1.1). The same experiments will show that analyses can be robust by the Approximate Maximum Influence Perturbation even if they are not Huber robust (Section 3.1.2). Heavy tails can cause non-robustness insofar as they cause high noise in the signal-to-noise, but (once this noise is controlled for) having a few outliers does not drive non-robustness (Section 3.1.3). Taking $N \to \infty$ does not guarantee robustness, though it does take the standard error to zero (Section 3.1.4). It will follow that statistical non-significance is inherently non-robust (Section 3.1.5).

We conclude by focusing on the special case of linear regression; we will see that particular points are influential if they have simultaneously both high leverage and a large-magnitude residual (Section 3.1.6).

3.1.1 Signal-to-noise drives robustness of the Approximate Maximum Influence Perturbation

We first define the signal, noise, and signal-to-noise ratio (Section 3.1.1.1). We then show via simulation experiments that the signal-to-noise ratio drives robustness in Section 3.1.1.2. The interested reader can see a more formal definition of noise in Section 3.1.1.3 and a proof that the noise can also be expressed as the scale of the influence scores in Section 3.1.1.4.

3.1.1.1 Defining the signal-to-noise ratio

There are two components in the signal-to-noise ratio: the signal and the noise.

Definition 3. The *signal* is Δ, the size of change of interest in our quantity of interest; see Definitions 1 and 2.

The *noise* $\hat{\sigma}_\psi$ is the estimator of the scaled asymptotic standard deviation of our quantity of interest found by combining the sandwich covariance estimator with the delta method; we define it formally in Section 3.1.1.3.

The *signal-to-noise ratio* is their ratio: $\Delta/\hat{\sigma}_\psi$.

Finally, we note that the squared noise $\hat{\sigma}_\psi^2$ can alternatively be expressed as

$$\hat{\sigma}_\psi^2 = N \left\| \vec{\psi} \right\|_2^2. \tag{3.1}$$

In other words, the limiting variance is the same as the scale of the influence scores, and, all else equal, larger scale on the influence scores means larger changes can be effected through dropping fewer points.

We will find Eq. 3.1 useful to calculate the noise in practice, as we see in our simulation experiments below (Section 3.1.1.2). We prove the equivalence from Eq. 3.1 in Section 3.1.1.4.

3.1.1.2 An ordinary least squares simulation to see that signal-to-noise drives robustness

Here we demonstrate via a simulation experiment how signal-to-noise drives robustness according to the Approximate Maximum Influence Perturbation. We will see that even an OLS estimator applied to data simulated from a classic Gaussian linear model can be either robust or non-robust, depending on the signal-to-noise ratio.

We set up our simulation experiment as follows. For $n \in \{1, \ldots, N\}$, we simulate regressors $x_n \in \mathbb{R}$ independently from a Gaussian with mean 0 and some variance

σ_X^2. Let X be the N-long vector (or $N \times 1$ matrix) with x_n as the nth entry. We simulate noise ϵ_n independently from a Gaussian with mean 0 and some variance σ_ϵ^2. Given some scalar β, we set $y_n \in \mathbb{R}$ as $y_n = \beta x_n + \epsilon_n$. Finally we let $\phi = \hat{\theta}$, where $\hat{\theta}$ is the OLS estimator under the model

$$y_n = \theta x_n + \epsilon_n. \tag{3.2}$$

Consider attempting to change the sign of the OLS estimate of θ by removing 1% of the sample. The signal in this case is the magnitude of $\hat{\theta}$, since one must effect a change of that magnitude in $\hat{\theta}$ to put it on the opposite side of zero (cf. Definition 3 and Section 2.5). The noise in this case is given by the sandwich covariance estimate (cf. Eq. 3.7)

$$\hat{\sigma}_\psi^2 = \frac{\frac{1}{N}\sum_{n=1}^N x_n^2 (y_n - x_n\hat{\theta})^2}{\left(\frac{1}{N}\sum_{n=1}^N x_n^2\right)^2}, \tag{3.3}$$

which tends to the ratio $\sigma_\epsilon^2/\sigma_X^2$ as $N \to \infty$. Note that, in our usage, "noise" does not refer only to the variability of the residuals, σ_ϵ^2, but to the variability of the estimator $\hat{\theta}$, which depends on the *relative* variability of the residuals and regressors. Our intuition supports that the ratio $\sigma_\epsilon^2/\sigma_X^2$ forms the relevant noise for our purposes here: reducing σ_X reduces the richness of the variation in our covariate of interest, and increasing σ_ϵ increases the unexplained variation in the inference problem.

First, we investigate a robust case. We choose $\sigma_X = 12.3$, $\sigma_\epsilon = 1.2$, $\beta = -1$, and $N = 10{,}000$. The OLS estimate is close to the true parameter value: $\hat{\theta} = -1.00071$ with a standard error of 0.00195. We now calculate the Approximate Maximum Influence Perturbation with $\alpha = 0.01$. We find that even adversarially removing 1% of the sample produces a change in $\hat{\theta}$ of only about 0.0038 (1/3 of 1% of the original estimate of $\hat{\theta}$). This analysis is robust. In fact, there are data subsets that could be removed to change the sign or significance, as there will necessarily be unless there is full separation in the support of the outcome data associated with the covariate of interest. But here they are sufficiently large that they cannot be detected by our linear approximation. In particular, first notice that the map $\alpha \mapsto \hat{\Psi}_\alpha$ is concave since $\hat{\Psi}_\alpha$ is the cumulative sum of sorted values. Then $\hat{\Psi}_1 \leq 100\hat{\Psi}_{0.01} = 0.38$. That is, even if we remove 100% of the data ($\alpha = 1$), we make an approximate predicted change of at most 0.38 units. Given $\hat{\theta} = -1.00071$, this change is not enough to affect the sign of the estimate. Figure 1 depicts the exact value of $\hat{\theta}$ as we vary the proportion α of points removed. Adversarial removal in one direction gives the upper red line; in the other gives the lower red line. The light blue shaded area gives the 95% confidence interval. We can see that up to 20% removal, sign and significance of the result are unchanged.

Even keeping the same general setup, and fixing $\theta = -1$ (so that the signal $\hat{\theta}$, will remain approximately the same), the empirical results can become non-robust

if we increase the noise. According to Eq. 3.3, decreasing σ_X and increasing σ_ϵ generates more noise. In what follows, we keep $N = 10{,}000$ and vary $\sigma_X \in (0, 4]$ and $\sigma_\epsilon \in (0, 12.5]$. In Figure 2, we make a heatmap with σ_X on the horizontal axis and σ_ϵ on the vertical axis. We plot (as a heatmap) the resulting proportion of the data α necessary to effect a change in (left) sign or (right) significance; the middle plot shows the proportion necessary to effect a significant conclusion of the opposite sign. That is, we plot the Approximate Perturbation Inducing Proportion $\hat{\alpha}^*_\Delta$ for the appropriate Δ in each case. We plot gray for `NA`, when the linear approximation is not able to detect a proportion suitable to make the change. Across all three plots, we see that the cases with a large σ_X relative to σ_ϵ are robust, while those with a large σ_ϵ relative to σ_X are non-robust. Note that, even in this simulated example, there are many cases where removing less than 1% of the sample can generate a significant conclusion of the opposite sign, despite the fact that the regression model is exactly correct and the data is exactly Gaussian. While the inference here is still valid for the population from which the perfectly random sample is drawn, our metric shows that this analysis is exposed to substantial risk should the sampling procedure be compromised, or should the population change in even minor ways.

A key corollary is that, in general, it is neither misspecification nor gross outliers that causes non-robustness according to the Approximate Maximum Influence Perturbation, except insofar as these features decrease the signal-to-noise ratio. Perfect observations from correctly specified models can give rise to non-robust estimators if they attempt to estimate a small effect from noisy data.

3.1.1.3 A formal definition of the noise

We now formally define the noise. To that end, we first consider the scaled asymptotic distribution of our estimator $\hat{\theta}$ and then use that to establish the scaled asymptotic standard deviation of our quantity of interest ϕ. Let $\theta_0 := \lim_{N\to\infty} \hat{\theta}$ be the limit of our sequence of Z-estimators. Then, under standard regularity conditions, $\sqrt{N}(\hat{\theta} - \theta_0)$ converges in distribution to a mean-zero multivariate normal distribution. And the sandwich covariance matrix, described next, is a consistent estimator of the covariance of this distribution (Van der Vaart, 2000, Theorem 5.23, Example 5.25); for more discussion, see Appendix B, particularly Eq. B.7 in Section B.3.

$$\hat{\Sigma}_\theta(\vec{w}) := NH(\vec{w})^{-1}S(\vec{w})H(\vec{w})^{-1}, \text{where} \tag{3.4}$$

$$H(\vec{w}) := \sum_{n=1}^{N} w_n \left.\frac{\partial G(\theta, d_n)}{\partial \theta}\right|_{\theta=\hat{\theta}(\vec{w})} \tag{3.5}$$

$$S(\vec{w}) := \sum_{n=1}^{N} w_n G(\hat{\theta}(\vec{w}), d_n) G(\hat{\theta}(\vec{w}), d_n)^T \tag{3.6}$$

More precisely, $\hat{\Sigma}_\theta(\vec{1})$ gives a consistent estimator of the limiting variance of $\sqrt{N}(\hat{\theta} - \theta_0)$ under standard regularity conditions. In the case of linear models, $\hat{\Sigma}_\theta(\vec{1})$ is also

known as the "robust" standard error covariance.

Though our definition of the noise here requires only $\hat{\Sigma}_\theta(\vec{1})$ and not the fully general $\hat{\Sigma}_\theta(\vec{w})$, we have written Eq. 3.4 with general weights $\vec{w}$ so that we can include it as part of a quantity of interest ϕ to evaluate the sensitivity of statistical significance (cf. Section 2.5). For general weight vectors, whether $\hat{\Sigma}_\theta(\vec{w})$ should be normalized by N or $\sum_{n=1}^N w_n$, and whether $S(\vec{w})$ should be weighted by w_n^2 or w_n, are subtle questions that depend on whether you consider reweighting to be a perturbation of the objective function or empirical distribution, respectively. The question is ultimately one of semantics, and throughout the present paper we prefer the definition given in Eq. 3.4, though we use definitions that match previous authors when analyzing previously published work.[3]

Next we consider the scaled limiting distribution of our quantity of interest, $\sqrt{N}(\phi(\hat{\theta}, \vec{w}) - \phi(\theta_0, \vec{w}))$ as $N \to \infty$. By an application of the delta method, this limiting distribution will be mean-zero univariate normal, and the following quantity consistently estimates its variance:

$$\hat{\sigma}_\psi^2(\vec{w}) := \left.\frac{\partial \phi(\theta, \vec{w})}{\partial \theta^T}\right|_{\hat{\theta}(\vec{w}), \vec{w}} \hat{\Sigma}_\theta(\vec{w}) \left.\frac{\partial \phi(\theta, \vec{w})}{\partial \theta}\right|_{\hat{\theta}(\vec{w}), \vec{w}}. \tag{3.7}$$

As with $\hat{\theta} = \hat{\theta}(\vec{1})$, we will take $\hat{\sigma}_\psi^2$ with no argument to mean $\hat{\sigma}_\psi^2 = \hat{\sigma}_\psi^2(\vec{1})$. Finally, we define $\hat{\sigma}_\psi$ to be the noise.

3.1.1.4 The noise is also the scaled influence-score norm

We now establish the equivalence in Eq. 3.1. To that end, observe from Eq. 2.10, together with the definitions in Eq. 3.5 and Eq. 3.6, that

$$\sum_{n=1}^N \left.\frac{d\hat{\theta}(\vec{w})}{dw_n}\right|_{\vec{w}} \left.\frac{d\hat{\theta}(\vec{w})^T}{dw_n}\right|_{\vec{w}} = H(\vec{w})^{-1} S(\vec{w}) H(\vec{w})^{-1} = \frac{1}{N}\hat{\Sigma}_\theta(\vec{w}). \tag{3.8}$$

We next apply Eqs. 2.7 and 2.4 with the observation that the last term in Eq. 2.7 will be zero, since we assume for this section that ϕ has no dependence on $\vec{w}$ beyond via $\hat{\theta}(\vec{w})$. In this case,

$$\psi_n := \left.\frac{\partial \phi(\vec{w})}{\partial w_n}\right|_{\vec{w}=\vec{1}} = \left.\frac{\partial \phi(\theta, \vec{1})}{\partial \theta^T}\right|_{\theta=\hat{\theta}(\vec{1})} \left.\frac{\mathrm{d}\hat{\theta}(\vec{w})}{\mathrm{d}w_n}\right|_{\vec{w}=\vec{1}}$$

[3]For example, the `lm` function in `R` normalizes by N rather than $\sum_{n=1}^N w_n$, but weights with w_n rather than w_n^2. Putting these choices together arguably forms an incoherent mix of perturbing the objective function and empirical distribution. When comparing with previous authors who used standard errors computed from `lm`, we compute standard errors to match `lm` despite this incoherency. In contrast, the `R` function `sandwich::vcovCL` (Zeileis et al., 2020) matches our Eq. 3.4.

So we then find

$$
\begin{aligned}
N\sum_{n=1}^{N}\psi_n^2 &= N\sum_{n=1}^{N}\left.\frac{\partial\phi(\theta,\vec{1})}{\partial\theta^T}\right|_{\theta=\hat{\theta}(\vec{1})}\left.\frac{\mathrm{d}\hat{\theta}(\vec{w})}{\mathrm{d}w_n}\right|_{\vec{w}=\vec{1}}\left.\frac{\mathrm{d}\hat{\theta}^T(\vec{w})}{\mathrm{d}w_n}\right|_{\vec{w}=\vec{1}}\left.\frac{\partial\phi(\theta,\vec{1})}{\partial\theta^T}\right|_{\theta=\hat{\theta}(\vec{1})} \\
&= \left.\frac{\partial\phi(\theta,\vec{1})}{\partial\theta^T}\right|_{\theta=\hat{\theta}(\vec{1})}\hat{\Sigma}_\theta(\vec{1})\left.\frac{\partial\phi(\theta,\vec{1})}{\partial\theta}\right|_{\theta=\hat{\theta}(\vec{1})} \\
&= \hat{\sigma}_\psi^2,
\end{aligned}
$$

where the final line follows from Section 3.1.1.3 and is the desired result.

3.1.2 An estimator can be non-Huber-robust but robust under the Approximate Maximum Influence Perturbation

Averages are well-known to be non-robust in the Huber sense of "gross error sensitivity" (Huber, 1983; Kim and White, 2004), so one may be tempted to assume this issue explains the reason for non-robustness in our OLS example (Section 2.3). But we note that, in fact, the robustness we examine here is distinct from Huber robustness. Recall that gross error sensitivity concerns the influence that an extremely large change in the value of a small fraction of the data set would have on the inference. In the Approximate Maximum Influence Perturbation, we are concerned only with dropping a small fraction of the data. Indeed, as our simulation results in Figure 2 and Section 3.1.1 show, there are many cases when a mean is highly robust to the small perturbations considered by the Approximate Maximum Influence Perturbation. Indeed, our results show cases where one must remove over 30% or more of the sample to change the significance of the result. And there are many cases where there is no local perturbation that could effect changes to sign or significance.

3.1.3 How heavy tails and outliers affect robustness

A common intuition is that heavy tails and outliers may drive non-robustness. There are at least two hypotheses for why a heavy-tailed distribution might be non-robust: (1) it produces large noise in the data or (2) it generates a few data points relatively far from the majority of data points. We here break down the Approximate Maximum Influence Perturbation $\hat{\Psi}_\alpha$ into terms relating to both hypotheses. We find that while large noise in the data drives non-robustness, the shape of a distribution is less of a determinant — and operates in perhaps unexpected directions.

First we decompose the Approximate Maximum Influence Perturbation into

noise and shape factors. From the definition of $\hat{\Psi}_\alpha$,

$$\hat{\Psi}_\alpha = \sum_{n \in \hat{S}_\alpha} -\psi_n = \hat{\sigma}_\psi \cdot \left\{ -\frac{1}{N} \sum_{n \in \hat{S}_\alpha} \frac{N\psi_n}{\hat{\sigma}_\psi} \right\} = \hat{\sigma}_\psi \Gamma_\alpha, \tag{3.9}$$

$$\text{where } \Gamma_\alpha := -\frac{1}{N} \sum_{n \in \hat{S}_\alpha} \gamma_n \text{ and } \gamma_n := \frac{N\psi_n}{\hat{\sigma}_\psi} \tag{3.10}$$

Here $\hat{\sigma}_\psi$ is the noise, defining the scale of the data, and Γ_α describes a notion of shape. Our results in Section 3.1.1 are in accordance with hypothesis (1) above: a large noise $\hat{\sigma}_\psi$ leads to a large $\hat{\Psi}_\alpha$.

Counter to hypothesis (2) above, we will see that Γ_α is lower (and hence $\hat{\Psi}_\alpha$ is lower) when there are a few extreme influence scores. To that end, we first establish that the γ_n summands that make up Γ_α have empirical mean zero and empirical variance one. In particular, by Eq. 2.1, we have $\sum_{n=1}^N G(\hat{\theta}(\vec{1}), d_n) = 0$; then, by Eq. 2.10, $\sum_{n=1}^N \psi_n = \sum_{n=1}^N \frac{\partial \hat{\theta}(\vec{w})}{\partial w_n}\Big|_{\vec{w}=\vec{1}} = 0$.[4] It follows that $\sum_{n=1}^N \gamma_n = 0$. And, recalling that $\hat{\sigma}_\psi^2 = N\|\vec{\psi}\|_2^2$ (see Eq. 3.1), we have $\frac{1}{N}\sum_{n=1}^N \gamma_n^2 = 1$.

We compare influence scores ψ_n for a light-tailed distribution and heavy-tailed distribution in Figure 4; each distribution is scaled to have unit variance. We leave out $M = 4$ data points, shown in red. The heavy-tailed distribution has one large negative entry, but the other entries are correspondingly smaller due to the constraint $\frac{1}{N}\sum_{n=1}^N \gamma_n^2 = 1$. By contrast, the light-tailed distribution has a moderate number of fairly large entries. The result is that, summed over the four left-out points, the total Γ_α is larger for the light-tailed distribution than for the heavy-tailed distribution. Hence, for the same noise, we expect a larger Γ_α for the light-tailed distribution than for the heavy-tailed distribution here — and thus a larger Approximate Maximum Influence Perturbation for the light-tailed distribution.

In fact, the Γ_α are bounded — unlike the noise $\hat{\sigma}_\psi$, which can range widely. We show in Appendix C that the following (finite-sample) bound holds:

$$|\Gamma_\alpha| \leq \sqrt{\alpha(1-\alpha)}. \tag{3.11}$$

The worst case $\Gamma_\alpha = \alpha(1-\alpha)$ is obtained when all of the influence scores ψ_n, across $n \in \hat{S}_\alpha$, are equal to each other. That is, when there are a few dominating influence scores at the extreme ranks, Γ_α is smaller than when there are nearly equal influence scores. Taking (e.g.) $\alpha = 0.01$ (i.e., removing 1% of datapoints) Eq. 3.11 gives the bound: $\Gamma_\alpha \leq 0.0995$.

We have seen that, perhaps unexpectedly, having a few extreme influence scores can reflect more robustness than having nearly equal influence scores. But even more to the point, we find that Γ_α does not vary much across common distributions and ultimately may not affect robustness much at all. Indeed, note that in Figure 4,

[4] The influence scores in fact sum to zero for any statistical functional that does not depend explicitly on N.

	Distribution	Γ_α
1	Worst case	0.0995
2	Normal	0.0266
3	Exponential	0.0460
4	Flipped exp	0.0099
5	T(10)	0.0300
6	T(3)	0.0408
7	T(2)	0.0361
8	Cauchy	0.0022
9	Uniform	0.0172
10	Binary(0.01)	0.0299
11	Binary(0.1)	0.0299
12	Binary(0.5)	0.0301

Table 1: Values of Γ_α with $\alpha = 0.01$ calculated from 1,000,000 simulated data points from each respective distribution.

the two Approximate Maximum Influence Perturbation values are similar.

We next simulate data by drawing γ_n directly from several common distributions and calculate Γ_α for each set of simulated data, with $\alpha = 0.01$. In each case, we generate $N = 1{,}000{,}000$ random data points from the distribution. See Table 1. We see that the resulting range of Γ_α values in Table 1 is relatively small. Since $\hat{\Psi}_\alpha = \hat{\sigma}_\psi \Gamma_\alpha$, the robustness is largely determined by the scale $\hat{\sigma}_\psi$ rather than the shape Γ_α.

3.1.4 $N \to \infty$ does not guarantee robustness

Consider a small, fixed α as $N \to \infty$; then we expect both Γ_α and $\hat{\sigma}_\psi$ to converge to non-zero quantities. Hence we expect $\hat{\Psi}_\alpha$ to converge to a non-zero quantity.

First, under standard conditions, $\hat{\sigma}_\psi \to \sigma_\psi > 0$. Next, observe that (up to rounding error in the number of left-out points αN),

$$\Gamma_\alpha = -\frac{1}{N} \sum_{n \in \hat{S}_\alpha} \gamma_n = -\frac{\alpha}{|\hat{S}_\alpha|} \sum_{n \in \hat{S}_\alpha} \gamma_n$$

Conditional on $\hat{\sigma}_\psi$, which converges to a constant, the term $-\frac{1}{|\hat{S}_\alpha|} \sum_{n \in \hat{S}_\alpha} \gamma_n$ is a sample average of $|\hat{S}_\alpha|$ observations, all of which typically have the same sign for $\alpha \ll 1/2$. Consequently, we expect that Γ_α converges to a non-zero constant as well.

This observation also allows us to compare using the Approximate Maximum Influence Perturbation to using the standard error, for the purpose of assessing robustness. To illustrate, choose $\alpha = 0.01$, and consider some scalar estimator θ. If the estimator θ is sufficiently regular, the traditional Gaussian asymptotic approach

to quantifying uncertainty yields the (approximate) two-sided 95% interval

$$\hat{\theta} - \theta_0 \in \left\{ -\frac{1.96}{\sqrt{N}}\hat{\sigma}_\psi, \frac{1.96}{\sqrt{N}}\hat{\sigma}_\psi \right\}. \tag{3.12}$$

Due to the $\sqrt{N}$ in the denominator, this interval shrinks to zero as N gets very large.

Now consider the Approximate Maximum Influence Perturbation approach. Let $\Gamma^+_{0.01}$ denote the shape parameter for a target function $\phi(\theta, \vec{w}) = \theta$, and $\Gamma^-_{0.01}$ the shape parameter for $\phi(\theta, \vec{w}) = -\theta$. Applying the decomposition from Eq. 3.9, we find that the uncertainty implied by the Approximate Maximum Influence Perturbation is

$$\hat{\theta}(\vec{w}) - \hat{\theta} \in \{-\Gamma^-_{0.01}\hat{\sigma}_\psi, \Gamma^+_{0.01}\hat{\sigma}_\psi\}. \tag{3.13}$$

By the arguments above, the width of this interval does not shrink to zero as $N \to \infty$.

We see that there may be genuine sensitivities to small perturbations of the data set even in very large samples. These sensitivities would be masked by examining only standard errors or relying on classical Gaussian asymptotics for estimators.

3.1.5 Statistical non-significance is non-robust

We have seen that the sensitivity captured by the $\hat{\Psi}_\alpha$ does not disappear asymptotically; this observation prompts a corollary that non-significance exhibits an inherent non-robustness. To see this non-robustness, take a particular $\hat{\Psi}_\alpha$. Observe that when a result is statistically non-significant, there will generally be some N for which we are able to move the parameter estimate far enough away from zero that the result becomes significant.

Significant results, by contrast, do not exhibit this inherent non-robustness — except in the case where the point estimates are small relative to the size of $\hat{\Psi}_\alpha$ so that the signal-to-noise ratio is small.

These observations are not particular to the Approximate Maximum Influence Perturbation in the sense that they would apply equally to any robustness measure that did not disappear asymptotically.

3.1.6 Simultaneous high leverage and large residual value yield a large influence score

We next focus on linear regression. In this case, we will see that high leverage or a large-magnitude residual alone would not guarantee a large influence score — but the confluence of these events will.

Let X be the $N \times P$ matrix whose nth row is x_n. Let Y be the $N \times 1$ column vector of y_n values. We take the data to be de-meaned. Leverage is often loosely defined as the extent to which results hinge on a single data point — but is formally

defined as the "influence" (in the colloquial sense) of the observed outcome data Y on the predicted values $\hat{Y}$: $h_{nn} := \frac{\partial \hat{y}_n}{\partial y_n}$. For linear regression, define the "hat matrix" $H = X(X^TX)^{-1}X^T$. In this case, the leverage of y_n on its fitted value $\hat{y}_n$ is the nth diagonal of the hat matrix:

$$h_{nn} = x_n^T(X^TX)^{-1}x_n. \tag{3.14}$$

Next we write the influence score of the nth data point with $\phi(\beta, \vec{w}) = \hat{y}_n$ for comparison.

$$\psi_n = \left.\frac{\partial \hat{y}_n(\vec{w})}{\partial w_n}\right|_{\vec{w}=\vec{1}} = \left(\left.\frac{\partial \hat{\beta}(\vec{w})}{\partial w_n}\right|_{\vec{w}=\vec{1}}\right)^T x_n = (y_n - \hat{\beta}(\vec{1})x_n)x_n^T(X^TX)^{-1}x_n. \tag{3.15}$$

We see, then, that the influence score ψ_n is the residual $\epsilon_n = y_n - \hat{\beta}(\vec{1})x_n$ times the leverage h_{nn}. This expression formalizes the conceptual link made by Chatterjee and Hadi (1986) between influence, leverage, and large values of ϵ_n. Note that large values of ϵ_n indicate that data point n is a kind of outlier. When both leverage and residual are large in magnitude, the influence score will be as well.

The univariate regression case may be particularly familiar, so we detail that special case next. In univariate regression, the leverage is

$$h_{nn} = \frac{x_n^2}{\sum_{n=1}^N x_n^2}. \tag{3.16}$$

The influence score is

$$\psi_n = \left.\frac{\partial \hat{y}_n(\vec{w})}{\partial w_n}\right|_{\vec{w}=\vec{1}} = \left.\frac{\partial \hat{\beta}(\vec{w})}{\partial w_n}\right|_{\vec{w}=\vec{1}} x_n = x_n \frac{x_n\epsilon_n}{\sum_{n=1}^N x_n^2} = h_{nn}\epsilon_n,$$

which again is the leverage times the residual.

3.2 Bounds on approximation error

In what follows, we first reiterate that, for any problem where performing estimation a second time is not prohibitively costly, a user can directly provide a lower bound on the exact Maximum Influence Perturbation (Section 3.2.1). Then we provide theoretical bounds on the quality of our linear approximation for ϕ in IV and OLS (Section 3.2.2), as well as a discussion of cases where the approximation may struggle (Section 3.2.3).

3.2.1 An exact lower bound on the Maximum Influence Perturbation

Here we show how to provide an exact lower bound on the Maximum Influence Perturbation when one is willing to incur the cost of a second estimation. To

establish the lower bound, we first find the Approximate Most Influential Set $\hat{S}_\alpha$ using the techniques described above. Let $\vec{w}^{**}$ be the exact Most Influential Set, and let $\vec{w}^*$ be the weight vector with all ones except with zeros at the indices in $\hat{S}_\alpha$. We run the estimation procedure an extra time to recover $\phi(\vec{w}^*)$. Then, by definition,

$$\Psi_\alpha = \phi(\vec{w}^{**}) - \phi(\vec{1}) = \max_{\vec{w} \in W_\alpha} \left(\phi(\vec{w}) - \phi(\vec{1})\right) \geq \phi(\vec{w}^*) - \phi(\vec{1}).$$

Since $\phi(\vec{w}^*) - \phi(\vec{1})$ is a lower bound for Ψ_α, we can use the Approximate Most Influential Set in this way to conclusively demonstrate non-robustness. However, this lower bound will be most useful if it is close to the exact Maximum Influence Perturbation Ψ_α. And we also want to understand how much we can trust any findings suggesting that a data analysis is robust. So we next focus on bounding error in our approximation.

3.2.2 Finite sample and asymptotic bounds on the approximation error for IV and OLS

Our key result in Section 3.2.2 below is the final application of Lemma 1, which provides a finite-sample upper bound on the error between our approximation $\phi^{\text{lin}}(\vec{w})$ and the exact $\phi(\vec{w})$. We also provide an upper bound on the difference $\phi(\vec{w}) - \phi(\vec{1})$ in Lemma 1, since this difference gives the relevant scale for considering the error of the linear approximation. All of our bounds here focus on the special case of IV and OLS. We conjecture similar results may hold for more general Z-estimators.

We first state our bounds in Section 3.2.2.1. Before our key result in Section 3.2.2.1, we provide intermediate and analogous bounds on a linear approximation of the parameter (Theorem 1) since these will be useful, due to the chain rule, in establishing bounds on the final quantity of interest. And we show how our finite-sample results translate into asymptotic statements in Corollaries 1 and 2. We conclude by providing discussion and interpretation of the bounds in Section 3.2.2.2.

3.2.2.1 Statement of the error bounds

We begin by setting up notation for the IV and OLS cases. For $n \in \{1, \ldots, N\}$, consider regressors $x_n \in \mathbb{R}^P$, instruments $z_n \in \mathbb{R}^P$, and responses $y_n \in \mathbb{R}$. The following estimating equation encompasses both IV regression and OLS.

$$G(\theta, \vec{w}) = \frac{1}{N} \sum_{n=1}^{N} w_n (y_n - \theta^T x_n) z_n. \tag{3.17}$$

In particular, we recover OLS by taking $z_n = x_n$.

Now recall from Eq. 2.7 that, by the chain rule, $\phi^{\text{lin}}(\vec{w})$ depends on the linear

approximation to the optimal parameters,

$$\hat{\theta}^{\text{lin}}(\vec{w}) = \left.\frac{\mathrm{d}\hat{\theta}(\vec{w})}{\mathrm{d}\vec{w}^T}\right|_{\vec{1}} (\vec{w} - \vec{1}) + \hat{\theta}.$$

The principal theoretical task is showing the accuracy of $\hat{\theta}^{\text{lin}}(\vec{w})$ as an approximation to $\hat{\theta}(\vec{w})$, which we show in Theorem 1. The accuracy of $\phi^{\text{lin}}(\vec{w})$ as an approximation to $\phi(\vec{w})$ will then follow from smoothness assumptions on ϕ in Lemma 1.

In what follows, we denote the operator norm of a matrix (i.e., its largest eigenvalue) by $\|\cdot\|_{op}$.

Theorem 1. *Let $\vec{w}$ be any vector whose entries are either zero or one. Define the following quantities, which can all be computed from $\hat{\theta}(\vec{1})$ without running any additional regressions:*[5]

$$\mathcal{S} := \{n : w_n = 0\}, \alpha := \frac{|\mathcal{S}|}{N}, \qquad \xi_1 := \left\| \frac{1}{|\mathcal{S}|} \sum_{n \in \mathcal{S}} z_n x_n^T \right\|_2,$$

$$C_{op} := \left\| \left(\frac{1}{N} \sum_{n=1}^{N} z_n x_n^T \right)^{-1} \right\|_{op}, \qquad \xi_2 := \left\| \frac{1}{|\mathcal{S}|} \sum_{n \in \mathcal{S}} z_n (y_n - \hat{\theta}(\vec{1})^T x_n) \right\|_2,$$

$$\tilde{C}_{op} := \frac{3}{2} C_{op}, \qquad \mathscr{B} := \frac{\left\| \phi^{\text{lin}}(\vec{w}) - \phi(\vec{1}) \right\|_2 + 2\alpha^2 \tilde{C}_{op}^2 \xi_1 \xi_2}{1 - 2\alpha^2 \tilde{C}_{op}^2 \xi_1^2}.$$

If $\alpha C_{op} \xi_1 \leq \frac{1}{3}$, then we conclude

$$\left\| \hat{\theta}(\vec{w}) - \hat{\theta}^{\text{lin}}(\vec{w}) \right\|_2 \leq \alpha^2 2 \tilde{C}_{op}^2 \xi_1 (\xi_2 + \mathscr{B} \xi_1) \qquad \text{and}$$
$$\left\| \hat{\theta}(\vec{w}) - \hat{\theta}(\vec{1}) \right\|_2 \leq \alpha C_{op} (\xi_2 + \mathscr{B} \xi_1).$$

Both bounds in Theorem 1 depend on Lemma 2 of Appendix D, which asserts that all of the assumptions for applying the bounds of Giordano et al. (2019) are satisfied. The first bound then follows from Giordano et al. (2019, Corollary 3) and Giordano et al. (2019, Lemma 10). The second result follows from Giordano et al. (2019, Corollary 2). The precise statement of Lemma 2 is slightly notationally burdensome, so we leave its statement and proof to Section D.1.

Next, we state a result that can be combined with Theorem 1 to control $\phi(\vec{w})$.

Lemma 1. *Let $\vec{w}$ be given such that $\left\| \vec{w} - \vec{1} \right\|_2 \leq \alpha N$, and let B_θ be a given convex, compact set containing both $\hat{\theta}$ and $\hat{\theta}(\vec{w})$. Define the normalized weights $\vec{\omega} := N^{-1}\vec{w}$, and let B_ω denote a convex, compact set containing both $N^{-1}\vec{w}$ and $N^{-1}\vec{1}$.*

Assume that ϕ is continuously differentiable in θ and $\vec{w}$ on $B_\theta \times B_\omega$. Assume that the partial derivatives $\left.\frac{\partial \phi(\theta, \vec{w})}{\partial \theta}\right|_{\theta, \vec{w}}$ and $\left.\frac{\partial \phi(\theta, \vec{w})}{\partial \vec{\omega}}\right|_{\theta, \vec{w}}$ are Lipschitz on $B_\theta \times B_\omega$ with

[5] The constant C_{op} requires the smallest eigenvalue of the matrix $\frac{1}{N}\sum_{n=1}^{N} z_n x_n^T$. This matrix is typically factorized in order to evaluate $\hat{\theta}(\vec{1})$, e.g. using a QR decomposition. This factorization can then be re-used to efficiently estimate C_{op} via the power method (Trefethen and Bau, 1997, Sections II, V).

constants L_θ and L_ω, respectively, in the distance induced by the norm

$$\|(\theta, \vec{\omega})\| = \|\theta\|_2 + \|\vec{\omega}\|_2 .$$

Choose any finite constants C_{diff}, C_{lin} C_θ, and C_ω such that

$$\left\|\hat{\theta}(\vec{w}) - \hat{\theta}\right\|_2 \leq C_{\text{diff}}\alpha \qquad \text{and} \qquad \left\|\hat{\theta}(\vec{w}) - \hat{\theta}^{\text{lin}}(\vec{w})\right\|_2 \leq C_{\text{lin}}\alpha^2$$
$$\sup_{\theta\in B_\theta, \vec{\omega}\in B_\omega} \left\| \frac{\partial\phi(\theta, \vec{w})}{\partial\theta}\Big|_{\theta, N\vec{\omega}} \right\|_2 \leq C_\theta \qquad \text{and} \qquad \sup_{\theta\in B_\theta, \vec{\omega}\in B_\omega} \left\| \frac{\partial\phi(\theta, N\vec{\omega})}{\partial\vec{\omega}}\Big|_{\theta, N\vec{\omega}} \right\|_2 \leq C_\omega.$$

Then

$$\left|\phi(\vec{w}) - \phi(\vec{1})\right| \leq (C_\theta C_{\text{diff}} + C_\omega)\,\alpha.$$

and

$$\left|\phi(\vec{w}) - \phi^{\text{lin}}(\vec{w})\right| \leq (L_\theta\,(C_{\text{diff}} + 1)\,C_{\text{diff}} + C_\theta C_{\text{lin}} + L_\omega\,(C_{\text{diff}} + 1))\,\alpha^2.$$

For a proof of Lemma 1, see Section D.2.

From the finite-sample bounds of Theorem 1 and Lemma 1, we can also derive the following asymptotic results by applying the finite-sample results to each N as $N \to \infty$.

Corollary 1. *For each N, choose any sequence of weight vectors $\vec{w}(N)$ with corresponding $\alpha(N)$ such that $\lim_{N\to\infty} \alpha(N) = 0$. Each of the constants from Theorem 1 now depends on N. Assume that each of the $C_{op}(N)$, $\xi_1(N)$, and $\xi_2(N)$ are eventually uniformly bounded as $N \to \infty$. Assume that there exists an N_0 such that, for all $N > N_0$, $\alpha(N)C_{op}(N)\xi_1(N) \leq \frac{1}{3}$ holds. Then, as $N \to \infty$,*

$$\left\|\hat{\theta}(\vec{w}) - \hat{\theta}^{\text{lin}}(\vec{w})\right\|_2 = O(\alpha^2) \to 0$$
$$\left\|\hat{\theta}^{\text{lin}}(\vec{w}) - \hat{\theta}(\vec{1})\right\|_2 = O(\alpha) \to 0.$$

Corollary 2. *Let the conditions of Corollary 1 hold, and suppose that there exists a compact set B_θ containing $\hat{\theta}(\vec{w})$, $\hat{\theta}^{\text{lin}}(\vec{w})$, and $\hat{\theta}(\vec{1})$ for all N. Let the assumptions of Lemma 1 hold for B_θ, specifically that ϕ is continuously differential with Lipschitz partial derivatives. Then, as $N \to \infty$,*

$$\left|\phi(\vec{w}) - \phi^{\text{lin}}(\vec{w})\right| = O(\alpha^2) \to 0$$
$$\left|\phi(\vec{w}) - \phi(\vec{1})\right| = O(\alpha) \to 0.$$

3.2.2.2 Interpretation of the bounds

The bounds of Theorem 1 and Lemma 1 hold for the observed data at hand. Importantly, Theorem 1 applies to any $\vec{w}$, including $\vec{w}$ that are chosen adversarially,

as long as $\vec{w}$ satisfies the (inexpensive-to-check) regularity condition $\alpha C_{op}\xi_1 \leq \frac{1}{3}$. Indeed we choose $\vec{w}$ adversarially in the present paper. For well-behaved functions ϕ, one can immediately apply Theorem 1 with Lemma 1 to derive corresponding bounds for the function of interest.

Our results show that the upper bound on the error in our linear approximations is $O(\alpha^2)$, but the difference that we are trying to approximate is upper bounded only by $O(\alpha)$. Together, these rates form the essence of why the linear approximation provides reliable robustness estimates; in particular, the rates together suggest that the error in the linear approximation is $O(\alpha)$ smaller than the difference it is trying to approximate, and so gives a good approximation for small α. Technically we need a lower bound on $\left\|\phi(\vec{w}) - \phi(\vec{1})\right\|_2$ to ensure it does not decrease even more quickly than a positive constant times α. But the discussion in Sections 3.1.3 and 3.1.4 suggests that, for the $\vec{w}$ that give rise to our robustness metrics, we expect $\left\|\phi(\vec{w}) - \phi(\vec{1})\right\|_2$ will typically be lower bounded by some constant times α. Recall the decomposition of Eq. 3.9, which gives $\phi^{\text{lin}}(\vec{w}) - \phi(\vec{1}) = \hat{\Psi}_\alpha = \hat{\sigma}_\psi \Gamma_\alpha$. We expect that the noise estimate $\hat{\sigma}_\psi$ converges to a non-zero constant, and, by the reasoning in Section 3.1.4, we expect Γ_α converges to α times a non-zero constant as well. Since the error of the linear approximation is upper bounded by $O(\alpha^2)$, it follows that actual difference is also lower bounded by a term of order α.

Note that the assumption in Theorem 1 that $\alpha C_{op}\xi_1 \leq \frac{1}{3}$ essentially requires that the regressors in the left-out set are not too large relative to the average regressors, particularly as $\alpha \to 0$. For example, when the regressors and instruments are bounded (i.e., $\max_{n\in[N]} \|x_n\|_2$ and $\max_{n\in[N]} \|z_n\|_2$ are bounded for all N), then $\alpha C_{op}\xi_1 \leq \frac{1}{3}$ will always hold for sufficiently small α.

Note that, for sequences of α that do not go to zero, the error $\left|\phi^{\text{lin}}(\vec{w}) - \phi(\vec{w})\right|_2$ does not go to zero in general. However, the bounds of Theorem 1 still apply, and we expect the error to be small for small α.

Though it may not be easy to produce explicit error bounds for general Z-estimators (beyond IV and OLS), we expect that the scaling we find in Corollary 1 as $\alpha \to 0$ to be similar given the results in Giordano et al. (2019).

3.2.3 Cases to Approach with Caution

In virtually all cases we examine in our applications in Section 4, we manually re-run the analysis without the data points in the removal set $\hat{S}_\alpha$; in doing so, we find that the change suggested by the approximation is indeed achieved in practice. We find this agreement even when up to 10% of the data is removed. However, there are some notable cases in which it is advisable to approach the Approximate Maximum Influence Perturbation with caution, and we discuss these next.

First and foremost, large changes to the data set are unlikely to yield good approximations due to the Taylor expansion; here "large" might be on the order of removing 30% of the data. The quality of our approximation rests on the similarity

between the exact function and its linearization. We do not view poor approximation at very large α as a major drawback for our robustness metric — since analysts are typically concerned with robustness to small changes. We are unlikely to be concerned about the Maximum Influence Perturbation or its approximation when removing a third of the sample.

We may also detect that the Approximate Maximum Influence Perturbation is incorrect when it reports that there is no feasible way to effect a particular change; i.e., when $\hat{\alpha}^*_\Delta = \texttt{NA}$. We might investigate, for example, a sign change on the treatment effect estimated in a randomized controlled trial. It might be true that $\alpha^*_\Delta = \texttt{NA}$ if there is complete separation in the sample across the treatment and control groups; i.e., every outcome for every individual in the treatment group lies above (respectively, below) those of every individual in the control group. Otherwise, there must be some proportion of the data that can be removed to effect the sign change. When the linear approximation cannot find any set of points whose removal could reverse the desired result, $\hat{\Psi}_\alpha$ is likely not a good approximation for Ψ_α for the large α that would be required to produce such a substantial change in Ψ_α. However, note that for small α, we expect that $\hat{\Psi}_\alpha$ will provide a good approximation of Ψ_α. So, when $\hat{\alpha}^*_\Delta = \texttt{NA}$, we may still confidently assert that there is no *small* α that could reverse our result.

Another case in which the $\hat{\Psi}_\alpha$ can fail is that of bounded parameters whose true value lies near the boundary. Because the linearization of $\hat{\theta}_d(\vec{w})$ is not bounded, the approximation can diverge from the truth when the true θ and original $\hat{\theta}$ are bounded — and will tend to do so near the boundary itself. For linear regression on a continuous outcome, this issue is not a concern for the regression coefficients, but it is a concern for the estimation of associated variance parameters. This concern can also arise in many Bayesian models that use bounds on parameters to improve estimation. And in hierarchical models, variances at different levels may be used for shrinkage as well as inference, and the hypervariances could be quite small in practice. It can help to linearize the problem using unconstrained reparameterizations (e.g., linearly approximating the log variance rather than variance). However, as we discuss in Section 4.4, simply transforming to an unconstrained space is still not guaranteed to produce accurate approximations near the boundary in the original, constrained space.

3.3 The influence function

In this section, we review the influence function Hampel (1974, 1986) and use it to connect our metric to, and contrast our metric with, existing concepts. First, we show that the equivalence between the noise $\hat{\sigma}_\psi$ in our signal-to-noise ratio and the 2-norm on the influence function, proven in Section 3.1.1.4, is a natural consequence of the asymptotic theory of statistical functionals. Next, we will show that the Approximate Maximum Influence Perturbation is itself a seminorm on the

influence function: namely, a supremum over tail means. This interpretation will allow us to contrast our metric with the gross error sensitivity, which is the ∞-norm of the influence function.

3.3.1 Influence function setup

Before moving to our main results, we review the definition of the influence function and its particular form for Z-estimators. The influence function $\mathrm{IF}(d; T, F)$ measures the effect on a statistic T of adding an infinitesimal amount of mass at point d to some base or reference data distribution F (Reeds, 1976; Hampel, 1986). Let δ_d be the probability measure with an atom of size 1 at d. Then

$$\mathrm{IF}(d; T, F) := \lim_{\epsilon \searrow 0} \frac{T(\epsilon\delta_d + (1-\epsilon)F) - T(F)}{\epsilon}.$$

Now we look at the specific case of Z-estimators. First, we consider a generalization of our estimating equation in Eq. 2.2. In this section of the paper, we define $\hat{\theta}(F)$ as the solution to

$$\int G(\hat{\theta}(F), d)\mathrm{d}F(d) = 0, \tag{3.18}$$

as in (Hampel, 1986, Section 4.2c, Def. 5). Eq. 2.1 is the special case of Eq. 3.18 where we set $F = \hat{F}_N$, and $\hat{F}_N$ is the empirical distribution function; that is, $\hat{F}_N$ puts weight N^{-1} on each data point $d_1, \ldots, d_N$. So $\hat{\theta}(\hat{F}_N)$ in this notation would be the solution of Eq. 2.1. Similarly, if we set F to be the distribution function putting weight $N^{-1}w_n$ at data point d_n, we recover Eq. 2.2 from Eq. 3.18.

Before we can express the influence function for Z-estimators, we establish our statistic. As before, we will choose some quantity of interest ϕ. In Section 2.1, we allowed ϕ to have both θ and $\vec{w}$ dependence: $\phi(\theta, \vec{w})$. In this section, for simplicity of exposition, we restrict our attention to $\phi(F) := \phi(\hat{\theta}(F), \vec{1})$. In particular, ϕ may depend on $\vec{w}$ (here, more generally, the dependence is on F) via $\hat{\theta}$ but not in other, additional ways.[6] Then, the influence function in the particular case of the Z-estimator can be written:

$$\mathrm{IF}(d; \phi, F) = -\left.\frac{\partial\phi(\theta, \vec{1})}{\partial\theta^T}\right|_{\hat{\theta}(F)} \left(\int \left.\frac{\partial G(\theta, \tilde{d})}{\partial\theta}\right|_{\hat{\theta}(F)} \mathrm{d}F(\tilde{d})\right)^{-1} G(\hat{\theta}(F), d). \tag{3.19}$$

3.3.2 The noise is the norm of the empirical influence function

In Section 3.1.1.4, we showed that the sandwich covariance estimator of the limiting variance of $\phi(\vec{1})$, after scaling by $\sqrt{N}$, was equal to N times the 2-norm of our

[6] As in ordinary calculus in Euclidean space, allowing for explicit F dependence in ϕ requires adding only an additional influence function describing the dependence of $\phi(\theta, F)$ on F with θ held fixed. This is slightly notationally burdensome and not typical in the analysis of the influence functions for Z-estimators, so we omit this dependence for simplicity.

influence vector, $N\left\|\vec{\psi}\right\|_2^2$. We now show that this equality is a natural consequence of the asymptotic theory of statistical functionals (Mises, 1947; Reeds, 1976; Hampel, 1986).

In the notation before this section, our quantity of interest on the full data set was called $\phi(\vec{1})$; in the notation of this section, we have established above that we will write the same quantity as $\phi(\hat{F}_N)$. The latter notation, which emphasizes the empirical distribution function, makes it especially clear that $\phi(\hat{F}_N)$ will vary as N varies. Let $\phi_0 := \text{plim}_{N\to\infty}\phi(\hat{F}_N)$. Then we are interested in the limiting distribution of $\sqrt{N}(\phi(\hat{F}_N) - \phi_0)$ as $N \to \infty$.

As discussed in Section 3.1.1, we expect this limiting distribution to be normal, since $\phi(\hat{F}_N)$ it is a Z estimator. However, we also may also expect asymptotic normality based on the limiting distribution of smooth functionals of the empirical distribution. In particular, the limiting standard deviation of a statistic T is given by the norm $\|\text{IF}(d;\phi,F)\|_2$ (Hampel, 1986, Eq. 2.1.8) in the following sense:

$$\sqrt{N}(T(\hat{F}_N) - T(F)) \rightsquigarrow \mathcal{N}(0, \|\text{IF}(\,\cdot\,;T,F)\|_2^2),$$

where $\|\text{IF}(\,\cdot\,;T,F)\|_2^2 = \int \text{IF}(d;T,F)^2 \text{d}F(d)$. So the limiting standard deviation of $\sqrt{N}(\phi(\hat{F}_N) - \phi_0)$ is $\|\text{IF}(\,\cdot\,;\phi,F)\|_2$.

A natural estimator of $\|\text{IF}(\,\cdot\,;\phi,F)\|_2$ is the same quantity, but with the empirical distribution substituted for the population distribution: $\left\|\text{IF}(\,\cdot\,;\phi,\hat{F}_N)\right\|_2$. We will now show that this estimator is precisely our delta method standard deviation $\hat{\sigma}_\psi = \left\|\text{IF}(\,\cdot\,;\phi,\hat{F}_N)\right\|_2 = \sqrt{N}\left\|\vec{\psi}\right\|_2$.

First, combine the definition of the influence score $\psi_n := \frac{\partial\phi(\vec{w})}{\partial w_n}\Big|_{\vec{w}=\vec{1}}$ (Eq. 2.4) with Eq. 2.7 and Eq. 2.10 and the notation of the current section to write

$$\psi_n = \left\{\frac{\partial\phi(\theta,\vec{1})}{\partial\theta^T}\Big|_{\hat{\theta}(\hat{F}_N)}\right\} \cdot \left\{-\left(N\int \frac{\partial G(\theta,\tilde{d})}{\partial\theta}\Big|_{\hat{\theta}(\hat{F}_N)} \text{d}\hat{F}_N(\tilde{d})\right)^{-1} G(\hat{\theta}(\hat{F}_N), d_n)\right\}. \tag{3.20}$$

Note that final term in Eq. 2.7 is zero here since we assume no dependence of ϕ on $\vec{w}$ beyond the dependence via $\hat{\theta}$. By comparing Eq. 3.20 to Eq. 3.19, we conclude that

$$\psi_n = \frac{1}{N}\text{IF}(d_n;\phi,\hat{F}_N). \tag{3.21}$$

Using Eq. 3.7, it follows that

$$\begin{aligned}\hat{\sigma}_\psi^2 = N\|\vec{\psi}\|_2^2 = N\sum_{n=1}^N \psi_n^2 = \frac{1}{N}\sum_{n=1}^N \text{IF}(d_n;\phi,\hat{F}_N)^2 \\ = \int \text{IF}(d;\phi,\hat{F}_N)^2 \text{d}\hat{F}_N(d) = \left\|\text{IF}(\,\cdot\,;\phi,\hat{F}_N)\right\|_2^2,\end{aligned}$$

as was to be shown.

3.3.3 The Approximate Maximum Influence Perturbation is a seminorm on the empirical influence function

We have previously seen that our Approximate Maximum Influence Perturbation metric is driven by both the signal and noise, $\hat{\sigma}_\psi$ (Section 3.1.1). We have just shown that the $\hat{\sigma}_\psi$ corresponds to a 2-norm of the influence function, applied at the empirical distribution $\hat{F}_N$. We will now see that in fact the Approximate Maximum Influence Perturbation corresponds to a (different) seminorm on the influence function, also applied at the empirical distribution $\hat{F}_N$.

First, for a particular choice of measure F and strictly positive scalar α, we define the seminorm we will use as follows:

$$\|f\|_{F,\alpha} := \sup_{S:F(S)\leq\alpha} -\int_S f(d)\mathrm{d}F(d).$$

Next, we establish that $\|\cdot\|_{F,\alpha}$ is, in fact, a seminorm. To do so, we check the two standard conditions of a seminorm. (1) Homogeneity follows from linearity of the integral. Namely, for any scalar c, $\|cf\|_{F,\alpha} = |c| \cdot \|f\|_{F,\alpha}$. (2) Subadditivity of $\|\cdot\|_{F,\alpha}$ follows from subadditivity of the supremum:

$$\begin{aligned}
\|f+g\|_{F,\alpha} &= \sup_{S:F(S)\leq\alpha} -\int_S (f(d)+g(d))\,\mathrm{d}F(d) \\
&\leq \left(\sup_{S:F(S)\leq\alpha} -\int_S f(d)\mathrm{d}F(d)\right) + \left(\sup_{S:F(S)\leq\alpha} -\int_S g(d)\mathrm{d}F(d)\right) \\
&= \|f\|_{F,\alpha} + \|g\|_{F,\alpha}\,,
\end{aligned}$$

We conclude that $\|\cdot\|_{F,\alpha}$ is a seminorm.

While non-negativity follows from the two conditions of a seminorm, we can also establish it directly. Since F is a measure and α is strictly positive, $S = \emptyset$ satisfies $F(S) = 0 \leq \alpha$. Then $-\int_{S=\emptyset} f(d)\mathrm{d}F(d) = 0$, so $\|f\|_{F,\alpha} \geq 0$. For $\|\cdot\|_{F,\alpha}$ to be a norm (not just a seminorm), we would also need that $\|f\|_{F,\alpha} = 0$ implies $f = 0$. But for a discrete F with atoms of size strictly greater than α, this implication will not necessarily hold. As a concrete example, consider a measure F with M atoms of size $1/M$ with $1/M > \alpha$; then $\|F\|_{F,\alpha} = 0$ since every S satisfying the condition $F(S) \leq \alpha$ has $F(S) = 0$, but $F \neq 0$. We will see the relevance of this example as we focus on the discrete choice $F = \hat{F}_N$ below.

Finally, we show that our Approximate Maximum Influence Perturbation metric $\hat{\Psi}_\alpha$ (Definition 2) is equal to $\left\|\mathrm{IF}(\,\cdot\,;\phi,\hat{F}_N)\right\|_{\hat{F}_N,\alpha}$. Since $\hat{F}_N$ puts weight N^{-1} at each data point d_n, the constraint $\hat{F}_N(S) \leq \alpha$ effectively restricts to all subsets of at most $\lfloor \alpha N \rfloor$ data points. That is, we obtain the integral in the seminorm by summing $-\frac{1}{N}\mathrm{IF}(\,\cdot\,;\phi,\hat{F}_N)$ over any such subset. Using Eq. 3.21, we conclude that

the supremum collects the data points with the most negative influence scores $\hat{\sigma}_\psi$. But this construction is exactly the Approximate Maximum Influence Perturbation, as seen in Definition 2 and the final line of Eq. 2.5.

As above, we see that if $\alpha < 1/N$ (equivalently, $N > 1/\alpha$), then $\left\|\mathrm{IF}(\,\cdot\,; \phi, \hat{F}_N)\right\|_{\hat{F}_N,\alpha} = 0$. Reasonably, with a data set of size N, we cannot expect to detect sensitivity to leaving out less than $1/N$ proportion of the data. In general, we recommend using an α that allows removal of at least one data point from the data set.

3.3.4 Comparison of the Approximate Maximum Influence Perturbation to the gross error sensitivity

The gross error sensitivity of a statistical function is defined as $\sup_d |\mathrm{IF}(d; \phi, F)|$, or, equivalently, $\|\mathrm{IF}(d; \phi, F)\|_\infty$ (Hampel, 1986, Eq. 2.1.13). Now that we have expressed our Approximate Maximum Influence Perturbation metric as a seminorm, we can see two key differences between our metric and the gross error sensitivity: (1) our metric uses the empirical influence function rather than the population influence function as the argument to its seminorm, and (2) our metric integrates over the empirical distribution function whereas the gross error sensitivity takes the infinity norm of the influence function.

Regarding (1), note that a central goal of classical robust statistics is to design estimators for which the gross error sensitivity is bounded. When the goal is designing estimators, rather than applying the gross sensitivity directly as an evaluation metric, using the population distribution may be seen as a strength. By contrast, our goal in the present paper is to evaluate sensitivity for the data we saw and the estimator we chose, not an idealized population or asymptotic limit. For this goal, we see our use of the empirical influence function as a strength.

Regarding (2), note that — by using the infinity norm — the gross error sensitivity measures the worst possible change in our statistic that could result from a small change to the population distribution function at a single point. By contrast, we are not concerned with corrupted data, much less adversarially corrupted data. In our setting, we allow that it is possible for all of our data to come from the same population distribution and even for the presumed population distribution to be an adequate description of the data we have collected. And yet it could still be the case that our conclusions are sensitive to dropping just a few data points — i.e., sensitive according to our metric. In this case, we would still be concerned about the broader implications of our conclusions, as we outlined at the start of this manuscript.

4 Applied experiments

4.1 The Oregon Medicaid experiment

In this subsection we show that even the conclusions of empirical analyses that display very little classical uncertainty can be sensitive to the removal of less than 1% of the sample. We consider the Oregon Medicaid study (Finkelstein et al., 2012). We focus on the impact of Medicaid on health outcomes. These empirical analyses exhibit standard errors that are small relative to effect size; against a null hypothesis of no effect, the p values are small: 0.019, 0.009, and the remaining five values rounded to 0.001 or smaller. We find that some results are robust up to the removal of 5% of the sample. But in others, removing less than 1% and even less than 0.05% of the sample will produce a significant result of the opposite sign to the full-sample analysis.

4.1.1 Background and replication

First we provide some context for the analysis and results of Finkelstein et al. (2012). In early 2008, the state of Oregon opened a waiting list for new enrollments in its Medicaid program for low-income adults. Oregon officials then drew names by lottery from the 90,000 people who signed up, and those who won the lottery could sign up for Medicaid along with any of their household members. This setup created a randomization into treatment and control groups at the household level. The Finkelstein et al. (2012) study measures outcomes one year after the treatment group received Medicaid. About 25% of the treatment group did indeed have Medicaid coverage by the end of the trial (a somewhat low compliance rate). The main analysis both investigates treatment assignment as treatment itself ("intent to treat" or ITT analysis) and uses treatment assignment as an instrumental variable for take-up of insurance coverage ("local average treatment effect" or LATE analysis).

The outcomes of interest are grouped into health care use indicators, compliance with recommended preventative care, financial strain related to medical expenditures, and health outcomes themselves (both physical and mental). We here focus on the final group: health outcomes, which appear in Panel B from Table 9 of Finkelstein et al. (2012). Each of these J outcomes is denoted by y_{ihj} for individual i in household h for outcome type j. The data sample to which we have access consists of survey responders ($N = 23{,}741$); some responders are from the same household. All regressions include a set of covariates X_{ih} including household size fixed effects, survey wave fixed effects, and the interaction between the two. They also include a set of optional demographic and economic covariates V_{ih}. To infer the Intention-to-treat (ITT) effects of winning the Medicaid lottery, the authors

estimate the following model via OLS:

$$y_{ihj} = \beta_0 + \beta_1 \text{LOTTERY}_h + \beta_2 X_{ih} + \beta_3 V_{ih} + \epsilon_{ihj}.$$

To infer the Local Average Treatment Effects (LATE) of taking up Medicaid on compliers, the authors employ a Two Stage Least Squares strategy. The first stage is:

$$\text{INSURANCE}_{ih} = \delta_0 + \delta_1 \text{LOTTERY}_h + \delta_2 X_{ih} + \delta_3 V_{ih} + \nu_{ihj}.$$

The second stage is:

$$y_{ihj} = \pi_0 + \pi_1 \text{INSURANCE}_{ih} + \pi_2 X_{ih} + \pi_3 V_{ih} + \mu_{ihj}.$$

All standard errors are clustered on the household, and all regressions are weighted using survey weights defined by the variable "weight_12m". We have access to the following seven variables, presented in Panel B of Table 9 in the following order: a binary indicator of a self-reported measure of health being good/very good/excellent (not fair or poor), a binary indicator on self-reported health not being poor, a binary indicator on health being about the same or improving over the last six months, the number of days of good physical health in the past 30 days, the number of days on which poor physical or mental health did not impair usual activities, the number of days mental health was good in the past 30 days, and an indicator on not being depressed in last two weeks. We replicate Panel B of Table 9 of Finkelstein et al. (2012) exactly, both for the ITT effect ($\hat{\beta}_1$) for the entire population and for the LATE on compliers ($\hat{\pi}_1$). In both cases, the results show very strong evidence for positive effects on all health measures, with most p values well below 0.01.

4.1.2 Applying our metric

Consider first the ITT analysis, which Finkelstein et al. (2012) conducted with a variety of control variables. For each health outcome in Panel B from Table 9 of Finkelstein et al. (2012), we apply our metric to assess how many data points one need remove to change the sign, the significance, and produce a significant result of the opposite sign. Table 4 summarizes our results, with all fixed effects and controls included and clustering at the household level. The table demonstrates that there are variables for which the sign can be changed by removing 0.05% of the data or less, or around 100 data points in a sample of approximately 22,000. It typically requires the removal of around 1% of the data to produce a significant result of the opposite sign — although some of the results are more robust and require almost 5% removal to be reversed. In Figure 6, we show how results vary with the proportion of points removed.

Consider now the LATE analysis, which Finkelstein et al. (2012) performed

using the two-stage-least-squares estimator. Table 5 shows the results of applying our metric to this analysis, with all fixed effects and controls included and with clustering at the household level. We find that the robustness of the IV results is very similar to the ITT case. Recent authors including Young (2019) have suggested that the uncertainty intervals for IV may be more poorly calibrated than the intervals for OLS. The fact that we find IV to be similarly robust, under our metric, to OLS does not contradict this conclusion. Young (2019) examines whether test size and power are close to nominal when data are sampled from a static population, and find that finite-sample inference for IV based on limiting normal approximations may not perform in line with asymptotic theory. We assess whether a statistic or function of interest to an analyst (typically interesting because the analyst accepts its validity, often based on asymptotic arguments) can be meaningfully altered by removing a few data points. Recall from Section 3.1 that even correctly specified models can be non-robust in terms of our metric when there is a low signal-to-noise ratio in the analysis problem. Similarly, confidence intervals that are robust to dropping influential points may yet have poor asymptotic coverage because they were formed under unrealistic assumptions. The two modes of failure for classical inference are orthogonal, though both may be a concern.

4.1.3 Checking approximation quality

Our results above were generated using our linear approximation to the combinatorial problem. To check the accuracy of the linear approximation, we re-run with the implicated data points removed. In this application, we find that our approximation always delivers the reversal of the results that it aims to deliver.

In the ITT case, we consider the impact of Medicaid on each of the seven outcomes. We re-run each regression after manually removing the data points in the Approximate Most Influential Set, $\hat{S}_{\hat{\alpha}^*_\Delta}$. Table 6 shows the results. We see that, even when 5% of the sample is removed, the linear approximation here still reliably uncovers combinations of data points that can deliver the claimed changes. As discussed above, the observed difference with the Approximate Most Influential Set removed forms a lower bound on the sensitivity. In particular, it may have been possible to discover a smaller set of points achieving the same change in results. But to check whether there is such a smaller set of points, one would need to solve the full (and prohibitively expensive) combinatorial optimization problem across all sufficiently small data subsets.

We similarly check the linear approximation for the LATE analyses. In this case, Table 7 shows the results of re-running the analyses with the points in $\hat{S}_{\hat{\alpha}^*_\Delta}$ removed. The linear approximations in these cases also perform well.

4.2 Cash transfers

In this subsection we show that removing outliers may not protect an empirical analysis from displaying sensitivity to removal of less than 1% of the sample. To that end, we now apply our techniques to examine the robustness of the main analysis from Angelucci and De Giorgi (2009), one of the flagship studies showing the impact of cash transfers on ineligible households, also known as "spillover effects." The authors employ a randomized controlled trial to study the impact of Progresa, a social program giving cash gifts to eligible poor households in Mexico. The randomization occurs at the village level. Therefore, the authors can study not only the main effect on the poor households selected to receive Progresa but also the impact on the non-eligible "non-poor" households located in the same villages as Progresa-receiving poor households. The analysis on the poor households is very robust, but the analysis on the non-poor households – whom the trimming protocol actually affects – is less robust.

4.2.1 Background and replication

The main results of the paper show that there are strong positive impacts of Progresa on total household consumption measured as an index both for eligible poor households and for the non-eligible households; see Table 1 of Angelucci and De Giorgi (2009). This variable is denoted C_ind_{it} for household i in time period t. The authors study three different time periods separately to detect any change in the impact between the short and long term. They further condition on a large set of variables (household poverty index, land size, head of household gender, age, whether speak indigenous language, literacy; at the locality level, poverty index and number of households) to ensure a fair comparison between households in the treatment and control villages. In this case these controls are important; the effects on the "nonpoor" households are significant at the 5% level when the controls are included, but they are only significant at the 10% level in a simple regression on a dummy for treatment status.

The full data for the paper is available on the website of the *American Economic Review* due to the open-data policies of the journal and the authors. We are able to successfully replicate the results of their analysis with the controls and without, and we proceed with the controls in this exercise as their preferred specification. We consider the time periods indexed as $t = 8, 9, 10$ in the dataset provided, though we note that the authors do not rely on the results at $t = 8$ as these are very early-stage. We employ K control variables, where X_{itk} is the kth variable for household i in period t. Then we run the following regression:

$$C\_ind_{it} = \beta_0 + \beta_1 \text{treat}_{poor,i} + \beta_2 \text{treat}_{nonpoor,i} + \sum_{k=1}^{K} \beta_{2+k} X_{itk} + \epsilon_{it}.$$

We are able to exactly replicate the results of Table 1 of Angelucci and De Giorgi (2009), which all show positive effects. We focus on the latter two time periods, as households had received only partial transfers in the first time period — but we show all three for completeness.

4.2.2 Applying our metric

We apply our approximate metric to perform a sensitivity analysis to assess how many data points one need remove to change the sign, the significance, or to generate a significant result of the opposite sign to that found in the full sample. Table 3 shows our results. We find that the inferences on the direct effects are quite robust, but the inferences on the indirect effects are more sensitive. For the analysis of the poor ("treatp"), one typically needs to remove around 5% and even up to 10% of the sample to effect these changes. For the analysis of the nonpoor ("treatnp"), one need remove less than 0.5% of the data to make these large changes, and removing only 3 data points in a sample of approximately 10,000 households can change the significance status for both $t = 9$ and $t = 10$. These differential robustness results likely reflect the differential signal to noise ratio that one might expect comparing direct and indirect effects. Our results also suggest the merits of a cautious approach to the spillovers literature more broadly.

In truncating the consumption variable, the authors of this study made what is typically considered a conservative choice in view of classical robustness concerns. Knowing that conditional mean estimates are sensitive to large data values, they deleted households with consumption indices greater than 10,000 units from the analysis (and both of these households were in the treatment group). The robustness of the direct effects on the poor households is not generated by the truncation because they are not in the truncated set. As our results show, even the truncated analysis remains non-robust for the nonpoor households. These observations further support our discussion in Section 3.1.6; namely, in the case of OLS linear regression, it is not simply large values of the outcome that produce large influence scores. Rather large regression errors combined with high leverage produce large influence scores. Hence, truncating based on large values of the outcome does not necessarily remove the highest influence data points for a given analysis.

4.2.3 Checking approximation quality

We again check the quality of our approximation. Table 3 shows the results of manually re-running each analysis after removing the implicated data points. In most cases the linear approximation correctly identifies a combination of data points that can make the claimed changes to the conclusions of the study. When the approximation falls short and does not actually result in as much change as claimed, we can see that it is nevertheless very close to achieving the claimed reversal.

4.3 Seven RCTs of microcredit: Linear regression analysis

We now show that even a simple 2-parameter linear model that performs a comparison of means between the treatment and control group of a randomized trial can be highly sensitive. To that end, we consider the analysis from seven randomized controlled trials of expanding access to microcredit, first aggregated in Meager (2019). In the next subsection, we will consider a Bayesian hierarchical model to see whether a Bayesian approach alleviates the sensitivity detected here.

4.3.1 Background

Each of the seven microcredit studies was conducted in a different country, and each study selected certain communities to randomly receive greater access to microcredit. Researchers either built a branch, or combined building a branch with some active outreach, or randomly selected borrowers among those who applied. The selected studies are: Angelucci et al. (2015), Attanasio et al. (2015), Augsburg et al. (2015), Banerjee et al. (2015), Crépon et al. (2015), Karlan and Zinman (2011), and Tarozzi et al. (2015). Six of these studies were published in a special issue of the *American Economics Journal: Applied Economics* on microcredit. All seven studies together are commonly considered to represent the most solid evidence base for understanding the impact of microcredit.

We first follow the original studies and Meager (2019) in analyzing the impact of this access itself as the treatment of interest. The studies range in their sample sizes from around 1,000 households in Mongolia (Attanasio et al., 2015) to around 16,500 households in Mexico (Angelucci et al., 2015). We consider the headline results on household business profit regressed on an intercept and a binary variable indicating whether a household was allocated to the treatment group or to the control group. For household i in site k, let Y_{ik} denote the profit measured, and let T_{ik} denote the treatment status. We estimate the following model via ordinary least squares:

$$Y_{ik} = \beta_0 + \beta T_{ik} + \epsilon_{ik}. \tag{4.1}$$

It is hard to imagine a more straightforward analysis. This regression model compares the means in the treatment and control groups and estimates the difference as $\hat{\beta}$. With a sample size of 1,000 or 16,500 and a typical econometrics education, one might believe that the resulting estimate of these two means — and thus their difference — would be highly accurate. We follow Meager (2019) in omitting the control variables or fixed effects from the regressions in order to examine the robustness of this fundamental procedure. But in principle this omission should

make no difference to the estimate $\hat{\beta}$, and indeed it does not (Meager, 2019).[7]

4.3.2 Applying our metric

We now analyze the robustness of the microcredit results for household profit to the removal of a small fraction of the data. Our results appear in Table 8. In this set of studies, we see that removing a small number of data points can change the sign, the significance, and generate a result of the opposite sign that would be deemed significant at the 5% level. As discussed in Section 2.3, the largest study is among the most sensitive; a single data point among the 16,561 households in Mexico determines the sign. To change both the sign and significance — that is, to turn Mexico's noisy negative result into a "strong" positive result — one need remove only 15 data points, i.e. less than 0.001% of the sample. Mongolia, the smallest study in terms of sample size, is among the most robust in terms of sign changes; it takes 2% of the sample to change the sign. The Philippines has the largest standard error yet is the most robust in terms of our ability to generate a significant result of the opposite sign, which would require the removal of more than 5% of the sample. Figure 7 illustrates how removing different proportions of data change the results in each case.

It may seem unsurprising that some of these results are highly non-robust, as they are all non-significant. Yet some of these non-significant results are more robust than some of the significant results in the Cash Transfers and Oregon Medicaid examples; consider the Philippines study, for example. It might also seem natural to implicate the fat tails of the household profit variable, a phenomenon well-documented by Meager (2020) and known to reduce the efficiency of the mean as an estimator of location. But we have shown in Section 3.1.3 that fat tails generally matter most in their role in generating outcome data with very large scale, so in what follows we focus on scale issues. Moreover, in Section 3.1 we showed that what matters is not the absolute scale but the relative scale of the estimated effect size and the tails of the influence function.

To show that it is relative scale of the data and $\hat{\beta}$ that matters, not the absolute scale, we now provide an examination of the much less variable outcome, and see that it reveals a similar sensitivity to the profit outcome. We consider household consumption spending on temptation goods such as alcohol, chocolate, and cigarettes. This variable had a much smaller scale than household profits, and Meager (2019) estimated it with the greatest precision of all six considered variables. Table 10 shows the results of applying our approximate metric to these analyses.

[7] The omission may in principle make a difference to the inference on β by affecting the standard errors. However, it turns out that in these studies the additional covariates make very little difference to the standard errors. We also do not cluster the standard errors at the community level for the same reason; the results are not substantially changed. Running the regression above in each of the seven studies delivers almost identical results to the preferred specification, as it should if intracluster correlations are weak and covariates are not strongly predictive of household profit.

While these analyses are somewhat more robust than the profit analyses, the difference is not large. Mongolia and India are now faring quite well, but it is still possible to change the sign of many of the estimates by removing around 1% of each sample or less. And it is possible to generate a "strong" result of the opposite sign by removing 2% of the data or less. Understanding that the relative scale is what matters rather than the absolute scale, it is not surprising that the temptation analysis is only marginally more robust than the profit analysis; the temptation analysis also had by far the smallest estimated $\hat{\beta}$ of all variables.

4.3.3 Checking approximation quality

We again test how well our linear approximation performs. Table 9 shows the results of manually re-running the profit analysis with the largest influence points removed. Table 11 shows the re-run results for the temptation analysis. The desired reversals are always attained for both variables. Again, we emphasize that it is possible that there are sets of fewer points that lead to even greater changes. But, as we have seen, this single re-run of the analysis offers a lower bound on the sensitivity. In this case, we see substantial changes when the percentage of removed points is very small, so the lower bound property guarantees that the true sensitivity is high.

4.4 Seven RCTs of microcredit: Bayesian hierarchical tailored mixture model

In this subsection, we show that the results of Bayesian analyses can also display major sensitivity to the removal of a small fraction of the sample. We specifically focus on the tailored mixture model from Meager (2020). One might hope that any of the following might alleviate sensitivity: the use of hierarchical Bayesian evidence aggregation, the regularization from incorporation of priors, or the somewhat more realistic data-generating process captured in this specific tailored likelihood. Indeed, the approach of Meager (2020) was specifically motivated by the desire to capture important features of the data-generating process such as fatter tails. We find that instead, the average effects remain sensitive according to our metric. But we also find that the estimated variance in treatment effects across studies is somewhat more robust than these averages.

4.4.1 Background

Following Meager (2020), we fit the model to all the data from the seven RCTs. We model each of the seven distributions using a spike at zero and two lognormal tail distributions, one for the positive realizations of profit and one for the negative realizations. Within the model, microcredit can affect the proportion of data assigned to each of these three components as well as affecting the location and

scale of the lognormal tails. There is a hierarchical shrinkage element to the model for each parameter. The hypervariances of the treatment effects are of particular interest because these capture the variation in effects across studies; this variation provides information about the transportability of results across settings. Here we will focus on the treatment effect of microcredit on the location parameters of the tail distributions.

The models in the original paper were fit via Hamiltonian Monte Carlo (HMC) in the software package Stan (Carpenter et al., 2017). It is possible to compute the Approximate Maximum Influence Perturbation for HMC, or for any Markov Chain Monte Carlo method, using the tools of Bayesian local robustness (Gustafson, 2000), but the sensitivity of simulation-based estimators is beyond the scope of this paper. However, there are ways to estimate Bayesian posteriors via optimization; perhaps the most notable among these are Variational Bayes (VB) techniques (Blei et al., 2016). Therefore, to proceed in this section, we fit the model using a variant of Automatic Differentiation Variational Inference (ADVI) (Kucukelbir et al., 2017) and apply our sensitivity analysis to the ADVI estimates. Specifically, we apply the version of ADVI described in Giordano et al. (2018, Section 5.2). Since the posterior uncertainty estimates of vanilla ADVI are notoriously inaccurate, we estimated posterior uncertainty using linear response covariances, again from Giordano et al. (2018, Section 5.2). We verified that the posterior means and covariance matrices of the ADVI and MCMC procedures agreed closely.

4.4.2 Applying our metric

Within our VB implementation, we compute the influence score ψ_n for each data point. Consider the effect of microcredit on the location parameter of the positive tail of profit, where the majority of the data in the samples is located. For each of the seven countries, Figure 8 shows the changes that one can make by removing data points successively from the most influential point in order of decreasing influence. We consider the removal of up to 1% of the total sample for each of the individual treatment effects on the means of the positive tails; the results are similar for the negative tails. The light blue bands in this case show the central 95% posterior interval. While a Bayesian analyst is not generally concerned with statistical significance, our analyst might be concerned that these marginal posteriors show a large sensitivity to the removal of a small fraction of the sample.[8] Parameter value regions that had a posterior mass of less than 2% in the original analysis can end up with a posterior mass of 50% when we remove only 1% of the sample.

The hypermeans and hypervariances for these location parameters are somewhat more robust than the country-specific parameters, and the hypervariances even more than the hypermeans. Figure 9 shows the change in the posterior marginals for these parameters for both the negative and positive tails. The hypervariances are some-

[8] And indeed, she is.

what more robust, as the removal of 1% of the sample would not materially alter the original paper's conclusions about effect heterogeneity across settings. This observation is quite interesting since, from a conventional robust statistics viewpoint, variances tend to be highly non-robust statistics. Our finding of a somewhat more robust hypervariance suggests that while it is possible to change the individual treatment effects by changing certain points, there must be different points removed for the different effects. It is easy to move each of the treatment effects around individually, but it is harder to move all of them in different directions away from each other by removing only a small part of the sample. Thus, the aggregate conclusions are somewhat more robust than the individual conclusions of the papers on which the analysis is based, although they are still somewhat sensitive.

4.4.3 Checking approximation quality

We again check the accuracy of our approximation. In this case, we uncover an instance where the approximation fails. We took the function of interest to be the log of the hypervariance and attempted make it as negative as possible. That is, we tried to find a set of points which, once removed, would make the hypervariance close to zero. Though the Approximate Maximum Influence Perturbation suggests that such a change is possible, Figure 10 shows that the actual estimator diverges as the hypervariance approaches the boundary at zero. In future work, it might be interesting to explore higher-order approximations to see if they might resolve this issue.

5 Conclusion

There are different ways of quantifying the dependence between the sample and the conclusions of statistical inference. While the idea of dependence on the vagaries of finite sample realisation has become synonymous with standard errors in frequentist statistics, the notions are equivalent only under a certain paradigm that considers a hypothetical random resampling exercise for the purpose of evaluating a specific parameter within a given model. Yet this hypothetical exercise may not always adequately capture all the data sensitivity relevant to applied social science. Indeed, much of 20th century statistics, with its focus on standard errors and sampling uncertainty, was motivated by randomized agricultural trials. In these cases, the difference in yield across multiple fields is well-modeled by independent sampling variation. Contrast with microcredit, where we do not believe the statistical model is an exact description of the true data-generating process. And in microcredit, the average effect is but a convenient summary; if the average profit were to increase slightly through one individual becoming wealthy while leaving all others destitute, one could consider the intervention a failure. By contrast, if a single plant produced an entire harvest's worth of corn, the outcome would still be desirable, if strange.

We here posit that there are other ways of conceiving of and quantifying the dependence of empirical results on the sample data, beyond standard errors. Sensitivity under our metric does not necessarily imply a problem with the sample. But the goal of inference is not to learn about the sample, but rather to learn about the population. Moreover in practice researchers often wish to conceive of that population in a relatively broad manner. If minor alterations to the sample can generate major changes in the inference, and we know the environment in which we do economics is changing all the time, we ought to be less confident that we have learned something fundamental about this broader population we seek to understand, for whom we ultimately seek to make policy. This conclusion does not necessarily mean that the original analysis is invalid according to classical sampling theory — and we do not recommend that researchers abandon the original full-sample results. However, reporting our metric alongside standard errors would improve our ability to understand and interpret the findings of a given analysis.

Working within an established general framework for local sensitivity, we have provided a computable metric, the Approximate Maximum Influence Perturbation, to quantify robustness to small perturbations of data in a finite sample. The Approximate Maximum Influence Perturbation can be computed automatically for many common analysis methods, and the quality of the approximation can be checked in practice at the cost of one additional data analysis. We find that common methods for data analysis in economics can, and often do, display important sensitivities to small fractions of the data. These sensitivities are present in all of the empirical applications we studied here, even when the inferential procedures are straightforward, although to differing extents. In many cases, removing less than 1% of the sample data can generate a strong, statistically significant result of the opposite sign to that claimed in the study.

However, we do find more-robust cases in certain applications, suggesting that applied economic analyses do differ in their sensitivities according to the Approximate Maximum Influence Perturbation. Analyses that are more robust are not obviously identifiable in terms of having smaller standard errors or fatter tails or inferior significance status. This observation confirms our theoretical findings that the sensitivity measured by Approximate Maximum Influence Perturbation is not captured by conventional metrics of sampling uncertainty. Instead, the sensitivity we analyze appears to be largely driven by the relative scale of the outcome and regressors as well as the size of their covariation, a set of factors succinctly captured in the signal to noise ratio for the given statistical problem. This intuition is also distinct from that underlying general Huber or robust statistics; indeed, we have shown that conditional means can perform well if the signal to noise ratio is high. We thus recommend that our metric be routinely computed and reported.

It now seems desirable to develop new statistical methods to address the presence of this kind of sensitivity, and moreover, to develop these methods in view of

the actual goals and uses of economics research rather than relying on a classical resampling paradigm that bears little resemblance to the practice of applied social science.

6 Figures

Table 2: Number of Data Points Affecting Conclusions of Cash Transfers Analysis

Case	N	Change Sign	Change Significance	Change Both
treatnp, t=10	4,266	30	3	101
treatnp, t=9	3,838	21	3	53
treatnp, t=8	4,624	5	17	28
treatp, t=10	10,518	697	435	1,049
treatp, t=9	9,630	345	146	653
treatp, t=8	10,936	252	83	520

Table 3: Manual Re-Runs Of The Cash Transfers Analysis

Case	Beta (SE)	Re-run for sign	Re-run for significance	Re-run for both
treatnp, t=10	21.49 (9.41)	-0.57 (6.75)	16.26 (8.93)	-11.96 (6.84)
treatnp, t=9	22.85 (10)	-0.37 (7.54)	16.51 (9.11)	-11.83 (7.11)
treatnp, t=8	-5.44 (7.13)	0.26 (6.41)	-12.97 (6.78)	10.44 (5.86)
treatp, t=10	33.86 (4.47)	-2.56 (3.54)	4.81 (3.68)	-11.34 (3.99)
treatp, t=9	27.92 (5.77)	-1.38 (4.41)	7.08 (4.55)	-11.26 (4.78)
treatp, t=8	17.31 (4.59)	-0.66 (3.75)	7.28 (4.1)	-9.11 (3.67)

Table 4: Number of Data Points Affecting Conclusions of Oregon Medicaid Table 9 ITT Results

Outcome	N	Change Sign	Change Significance	Change Both
health_genflip_bin_12m	23,361	286	163	422
health_notpoor_12m	23,361	156	101	224
health_chgflip_bin_12m	23,407	198	106	292
notbaddays_tot_12m	21,881	74	11	147
notbaddays_phys_12m	21,384	88	20	166
notbaddays_ment_12m	21,601	124	42	216
nodep_screen_12m	23,147	123	43	225

Table 5: Number of Data Points Affecting Conclusions of Oregon Medicaid Table 9 IV Results

Outcome	N	Change Sign	Change Significance	Change Both
health_genflip_bin_12m	23,361	275	162	383
health_notpoor_12m	23,361	155	100	219
health_chgflip_bin_12m	23,407	197	106	292
notbaddays_tot_12m	21,881	73	10	145
notbaddays_phys_12m	21,384	87	20	165
notbaddays_ment_12m	21,601	123	42	215
nodep_screen_12m	23,147	123	42	220

Table 6: Manual Re-Runs Of The Oregon Medicaid ITT Analysis

Outcome	Beta (SE)	Re-run for sign	Re-run for significance	Re-run for both
health_genflip_bin_12m	0.039 (0.008)	-0.004 (0.008)	0.013 (0.008)	-0.021 (0.008)
health_notpoor_12m	0.029 (0.005)	-0.001 (0.005)	0.008 (0.005)	-0.009 (0.004)
health_chgflip_bin_12m	0.033 (0.007)	-0.002 (0.006)	0.011 (0.007)	-0.015 (0.006)
notbaddays_tot_12m	0.381 (0.162)	-0.013 (0.157)	0.306 (0.161)	-0.309 (0.153)
notbaddays_phys_12m	0.459 (0.175)	-0.017 (0.169)	0.328 (0.172)	-0.344 (0.165)
notbaddays_ment_12m	0.603 (0.184)	-0.027 (0.178)	0.34 (0.181)	-0.381 (0.175)
nodep_screen_12m	0.023 (0.007)	-0.001 (0.007)	0.013 (0.007)	-0.015 (0.007)

Table 7: Manual Re-Runs Of The Oregon Medicaid IV Analysis

Outcome	Beta (SE)	Re-run for sign	Re-run for significance	Re-run for both
health_genflip_bin_12m	0.133 (0.026)	-0.006 (0.025)	0.044 (0.026)	-0.047 (0.024)
health_notpoor_12m	0.099 (0.018)	-0.003 (0.015)	0.027 (0.016)	-0.03 (0.015)
health_chgflip_bin_12m	0.113 (0.023)	-0.006 (0.022)	0.039 (0.022)	-0.051 (0.022)
notbaddays_tot_12m	1.317 (0.563)	-0.023 (0.535)	1.078 (0.558)	-1.035 (0.524)
notbaddays_phys_12m	1.585 (0.606)	-0.04 (0.577)	1.131 (0.597)	-1.169 (0.568)
notbaddays_ment_12m	2.082 (0.64)	-0.062 (0.607)	1.171 (0.625)	-1.293 (0.603)
nodep_screen_12m	0.078 (0.025)	-0.005 (0.024)	0.046 (0.024)	-0.05 (0.023)

Table 8: Number of Data Points Affecting Conclusions of Microcredit Profit Analysis

Country	N	Change Sign	Change Significance	Change Both
Mexico	16,560	1	14	15
Mongolia	961	16	2	38
Bosnia	1,195	14	1	40
India	6,863	6	1	32
Morocco	5,498	11	2	30
Philippines	1,113	9	10	63
Ethiopia	3,113	1	45	66

Table 9: Manual Re-Runs Of The Microcredit Profit Analysis

Country	Beta (SE)	Re-run for sign	Re-run for significance	Re-run for both
Mexico	-4.55 (5.88)	0.4 (3.19)	-10.96 (5.57)	7.03 (2.55)
Mongolia	-0.34 (0.22)	0.02 (0.18)	-0.44 (0.22)	0.36 (0.15)
Bosnia	37.53 (19.78)	-2.23 (15.63)	43.73 (18.89)	-34.93 (14.32)
India	16.72 (11.83)	-0.5 (8.22)	22.89 (10.27)	-16.64 (7.54)
Morocco	17.54 (11.4)	-0.57 (9.92)	21.72 (11)	-18.85 (9.01)
Philippines	66.56 (78.13)	-4.01 (57.2)	155.89 (77.37)	-135.41 (53.51)
Ethiopia	7.29 (7.89)	-0.05 (2.51)	15.36 (7.76)	-8.75 (1.85)

Table 10: Number of Data Points Affecting Conclusions of Microcredit Temptation Analysis

Country	N	Change Sign	Change Significance	Change Both
Mexico	16,435	12	14	55
Mongolia	961	3	12	162
Bosnia	996	10	1	33
India	6,827	41	8	85
Morocco	5,487	3	14	23

Table 11: Manual Re-Runs Of The Microcredit Temptation Analysis

Country	Beta (SE)	Re-run for sign	Re-run for significance	Re-run for both
Mexico	-0.08 (0.09)	0 (0.09)	-0.18 (0.09)	0.18 (0.09)
Mongolia	1.52 (2.1)	-0.03 (0.97)	4.21 (2.08)	-7.37 (2.41)
Bosnia	-5.8 (2.82)	0.39 (2.13)	-4.87 (2.69)	5.13 (1.98)
India	-1.64 (0.58)	0.04 (0.51)	-1.05 (0.54)	1.06 (0.49)
Morocco	-0.42 (0.72)	0.05 (0.67)	-1.35 (0.67)	1.25 (0.6)

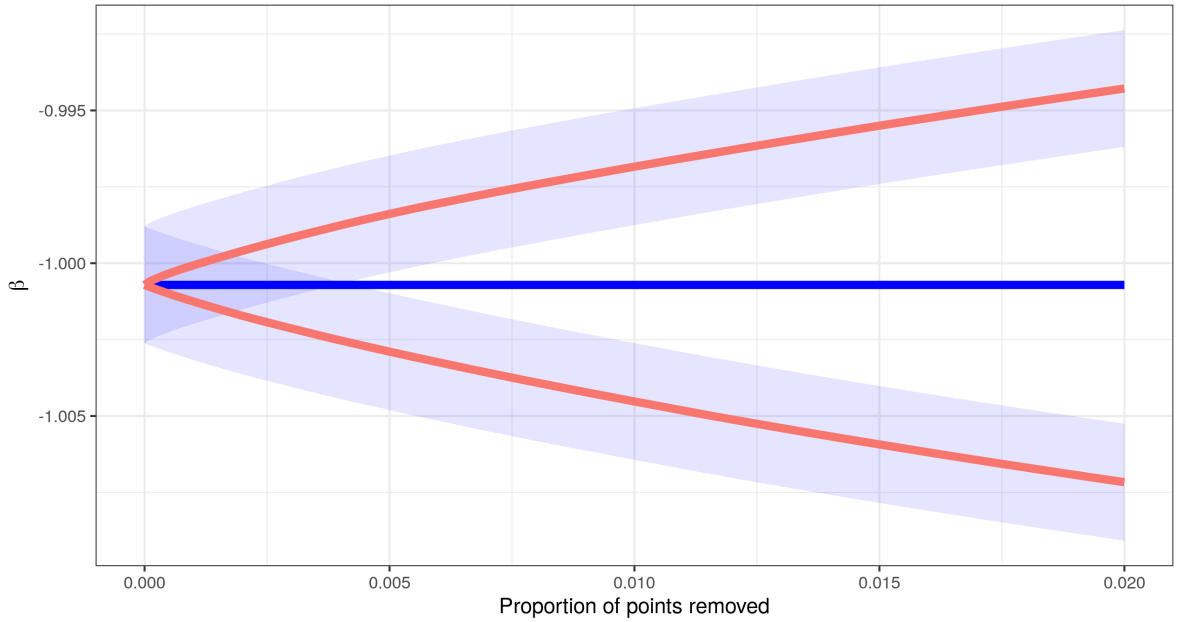

Figure 1: Simulation results for linear regression. Values of $\hat{\beta}$ are on the vertical axis; values of α (proportion of the data removed) are on the horizontal axis. The dark blue line shows the original $\hat{\beta}$ value. The red lines show how $\hat{\beta}$ can be altered by adversarial removal in both directions; the light blue shaded area is the 95% confidence interval.

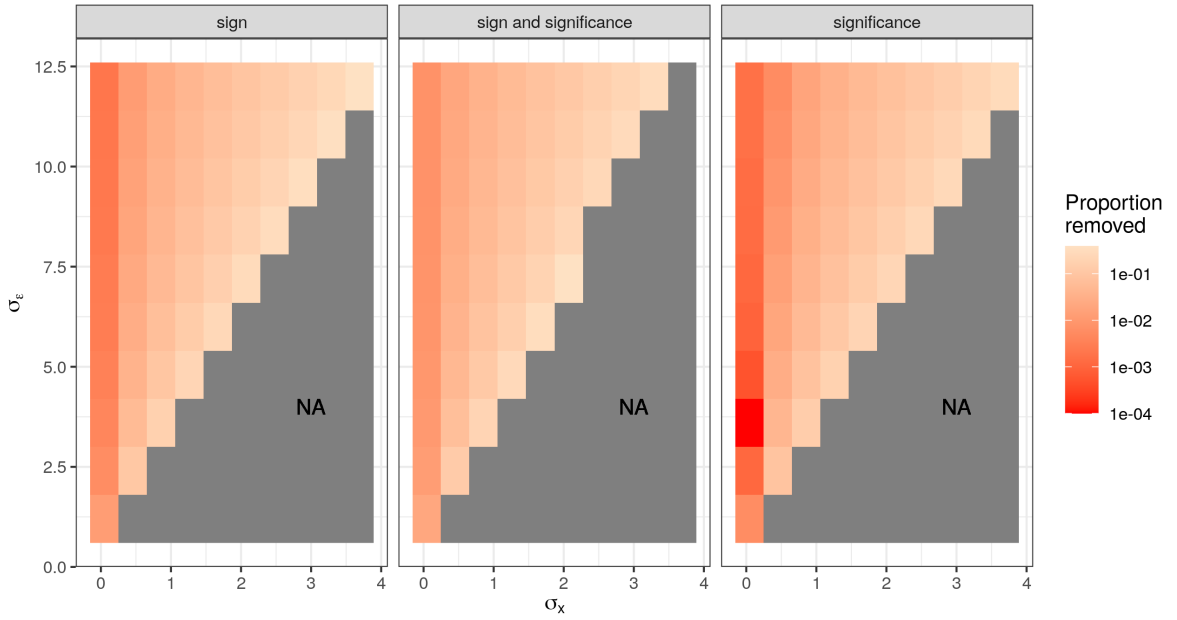

Figure 2: Simulation results for linear regression at differing scales of σ_X and σ_ϵ. A darker red colour indicates a highly sensitive analysis, in which only a small proportion of the sample needs to be removed to effect these three major changes: generating the opposite sign (left), changing the significance (right), and generating a significant result of the opposite sign (middle). A lighter red colour indicates greater robustness. The grey areas indicate $\hat{\Psi}_\alpha = \texttt{NA}$, a failure of the linear approximation to locate any way to effect these changes.

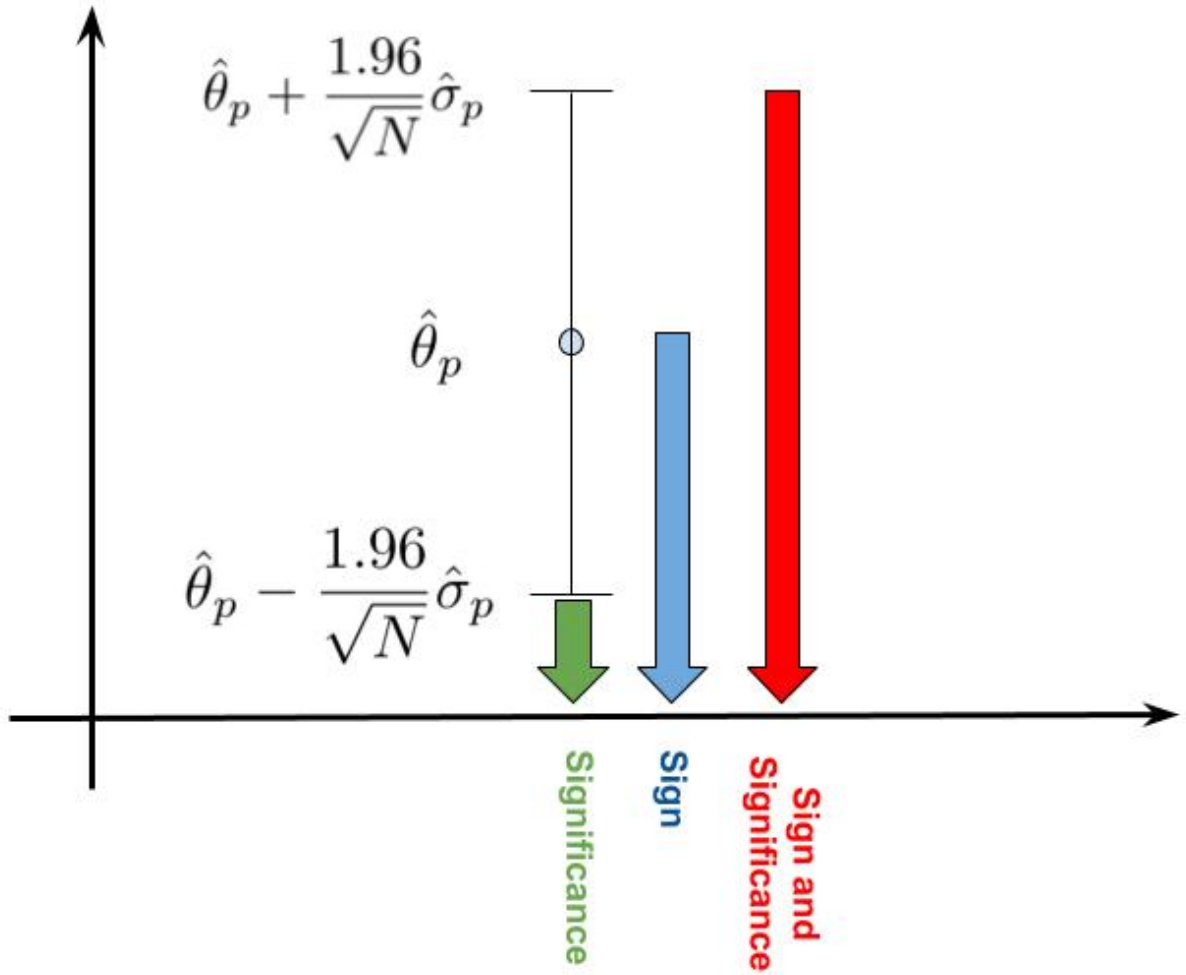

Figure 3: Illustration of the Δ required to effect a change of significance, sign, or both sign and significance of a positive, statistically significant effect. By definition, ϕ is a function of interest that we are trying to increase, so ϕ would be taken to be the negative of the illustrated quantities.

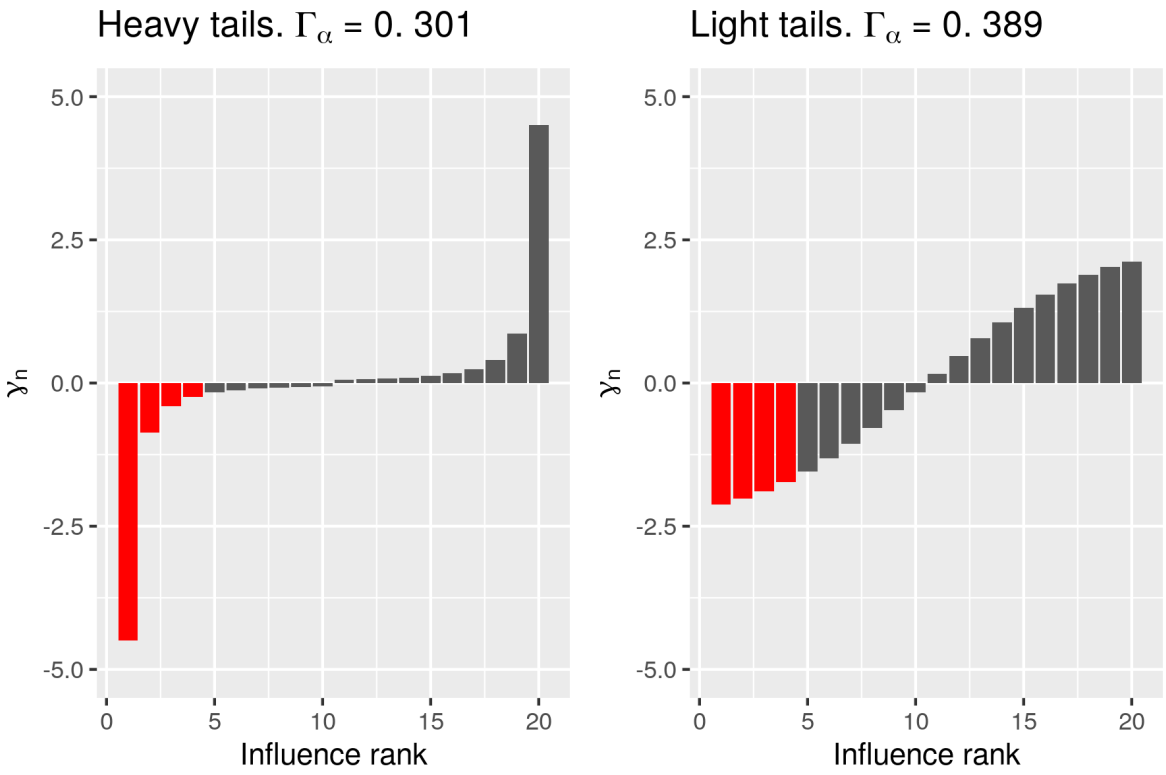

Figure 4: Simulations. Illustrative distributions of $\vec{\gamma}$ with traditionally-defined heavy and light tails. Here, $N = 20$, $\alpha = 0.2$, and $M = 4$. For both plots, $\sum_{n=1}^{N} \gamma_n^2 = 1$.

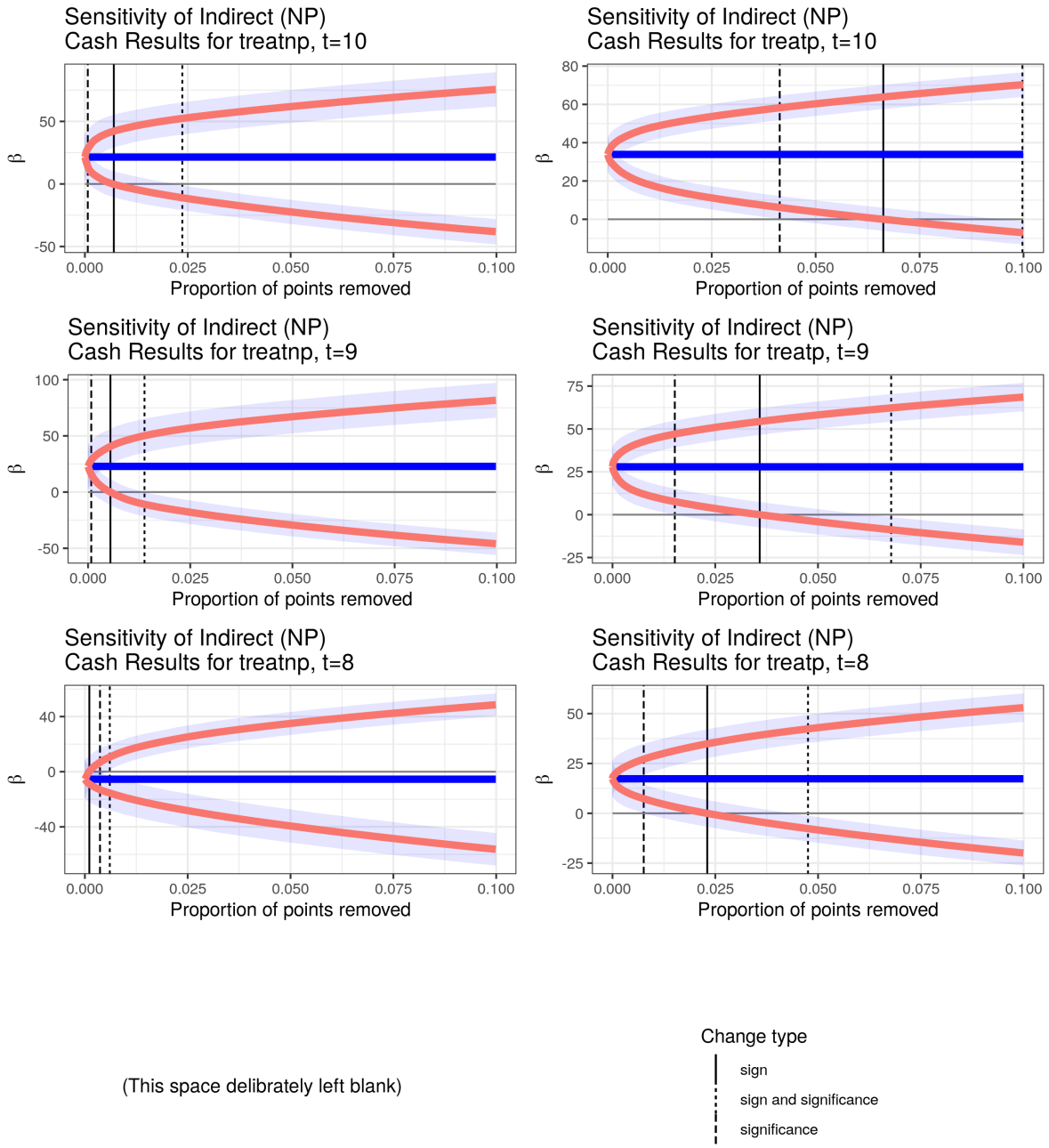

Figure 5: Cash Transfers Analysis. Values of $\hat{\beta}$ are on the vertical axis; values of α (proportion of the data removed) are on the horizontal axis. The dark blue line shows the original $\hat{\beta}$ value. The red lines show how $\hat{\beta}$ can be altered by adversarial removal in both directions; the light blue shaded area is the 95% confidence interval.

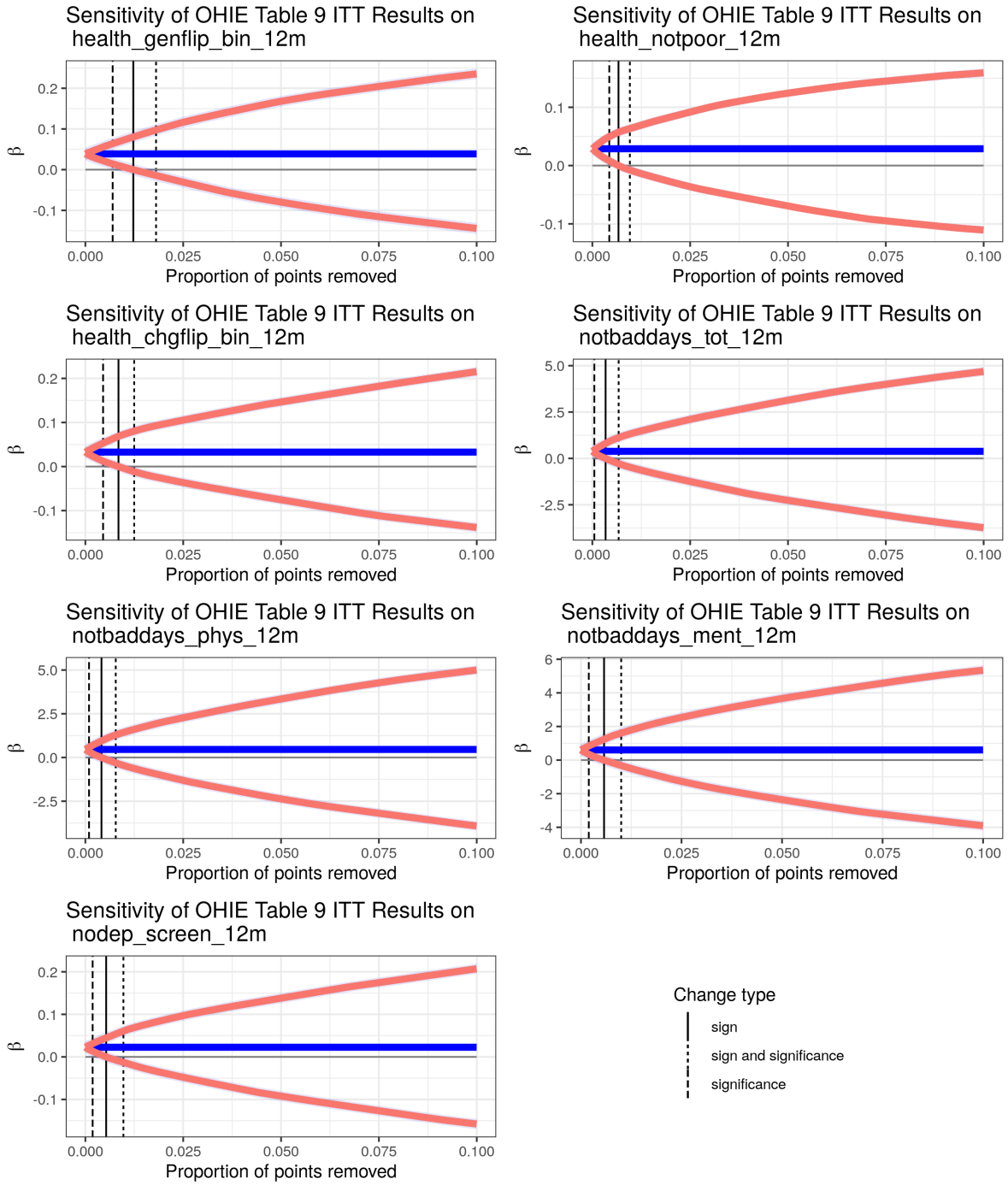

Figure 6: Oregon Medicaid Analysis. Values of $\hat{\beta}$ are on the vertical axis; values of α (proportion of the data removed) are on the horizontal axis. The dark blue line shows the original $\hat{\beta}$ value. The red lines show how $\hat{\beta}$ can be altered by adversarial removal in both directions; the light blue shaded area is the 95% confidence interval. In this case, the light blue shaded area is almost indistinguishable from the red lines.

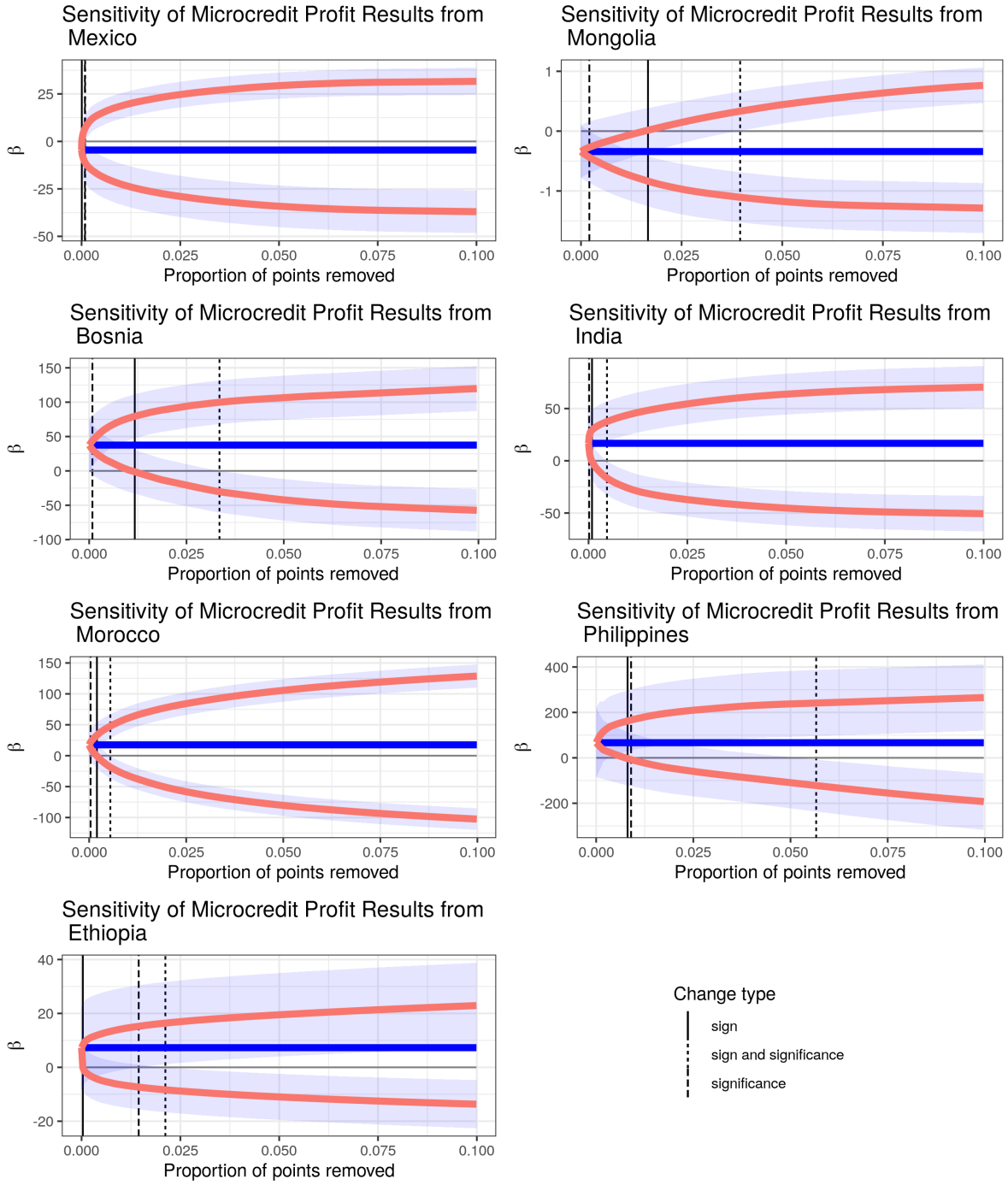

Figure 7: Microcredit Analysis: Linear. Values of $\hat{\beta}$ are on the vertical axis; values of α (proportion of the data removed) are on the horizontal axis. The dark blue line shows the original $\hat{\beta}$ value. The red lines show how $\hat{\beta}$ can be altered by adversarial removal in both directions; the light blue shaded area is the 95% confidence interval.

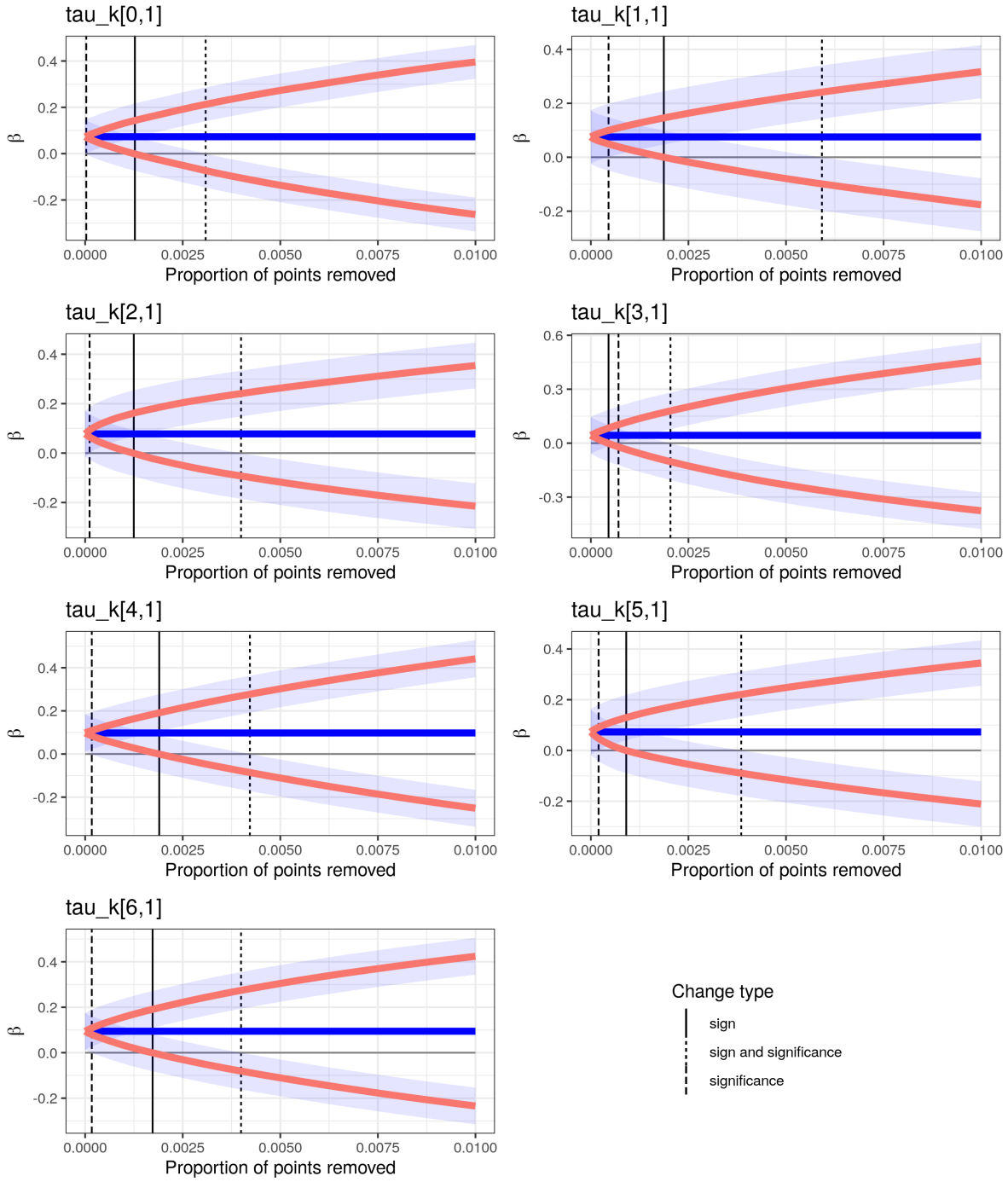

Figure 8: Microcredit Analysis: Bayesian Hierarchical. We first consider the 7 study-specific treatment effects on the location parameter of the positive tail of profit (within the lognormal specification). Parameter values are on the vertical axis; values of α (proportion of the data removed) are on the horizontal axis. The dark blue line shows the original marginal posterior mean value. The red lines show how the marginal posterior mean can be altered by adversarial removal in both directions; the light blue shaded area is the central 95% posterior interval.

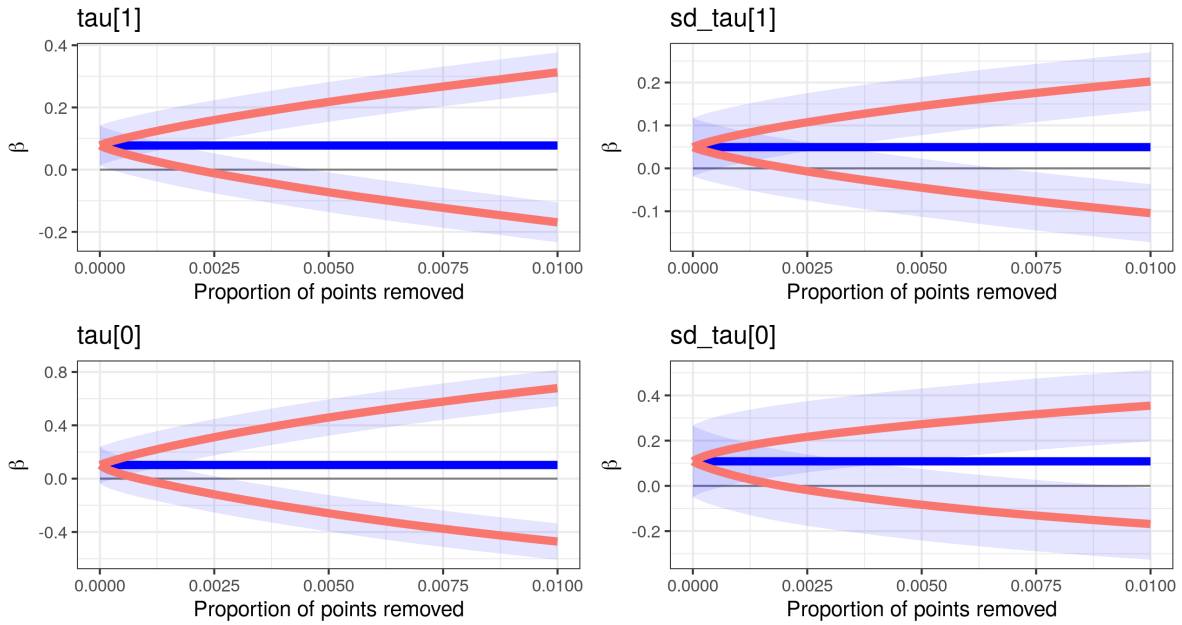

Figure 9: Microcredit Analysis: Bayesian Hierarchical. We now consider the hypermean and hypervariance of the treatment effect on the location parameter of the positive tail of profit (first row) and the negative tail of profit (second row). Parameter values are on the vertical axis; values of α (proportion of the data removed) are on the horizontal axis. The dark blue line shows the original marginal posterior mean value. The red lines show how the marginal posterior mean can be altered by adversarial removal in both directions; the light blue shaded area is the central 95% posterior interval.

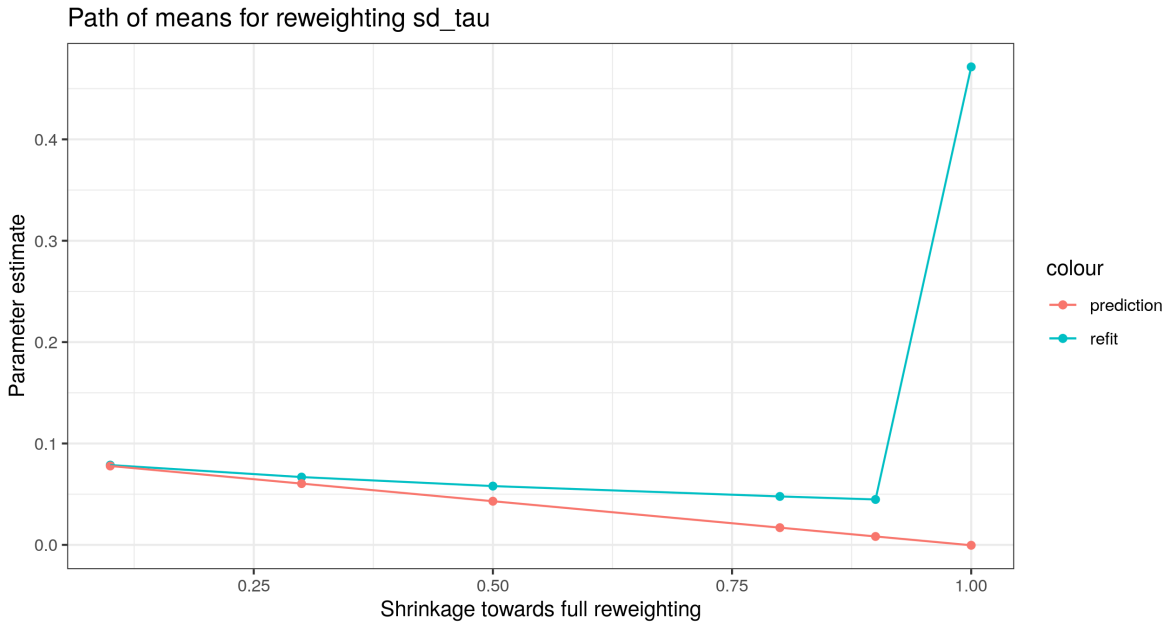

Figure 10: Comparison of the change in the posterior indicated by the approximation with the actual change in the hypervariance (here of the negative tail) achieved by re-running the analysis with the data points removed.

References

Angelucci, M. and De Giorgi, G. (2009). Indirect effects of an aid program: How do cash transfers affect ineligibles' consumption? *American Economic Review*, 99(1):486–508.

Angelucci, M., Karlan, D., and Zinman, J. (2015). Microcredit impacts: Evidence from a randomized microcredit program placement experiment by Compartamos Banco. *American Economic Journal: Applied Economics*, 7(1):151–82.

Attanasio, O., Augsburg, B., De Haas, R., Fitzsimons, E., and Harmgart, H. (2015). The impacts of microfinance: Evidence from joint-liability lending in Mongolia. *American Economic Journal: Applied Economics*, 7(1):90–122.

Augsburg, B., De Haas, R., Harmgart, H., and Meghir, C. (2015). The impacts of microcredit: Evidence from Bosnia and Herzegovina. *American Economic Journal: Applied Economics*, 7(1):183–203.

Banerjee, A., Duflo, E., Glennerster, R., and Kinnan, C. (2015). The miracle of microfinance? Evidence from a randomized evaluation. *American Economic Journal: Applied Economics*, 7(1):22–53.

Baydin, A., Pearlmutter, B., Radul, A., and Siskind, J. (2017). Automatic differentiation in machine learning: A survey. *The Journal of Machine Learning Research*, 18(1):5595–5637.

Blei, D., Kucukelbir, A., and McAuliffe, J. D. (2016). Variational inference: A review for statisticians. *arXiv preprint arXiv:1601.00670*.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1).

Chatterjee, S. and Hadi, A. (1986). Influential observations, high leverage points, and outliers in linear regression. *Statistical Science*, 1(3):379–393.

Chen, X., Tamer, E., and Torgovitsky, A. (2011). Sensitivity analysis in semiparametric likelihood models. *Cowles Foundation discussion paper*.

Crépon, B., Devoto, F., Duflo, E., and Parienté, W. (2015). Estimating the impact of microcredit on those who take it up: Evidence from a randomized experiment in Morocco. *American Economic Journal: Applied Economics*, 7(1):123–50.

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

Giordano, R., Broderick, T., and Jordan, M. I. (2018). Covariances, robustness and variational Bayes. *The Journal of Machine Learning Research*, 19(1):1981–2029.

Giordano, R., Jordan, M. I., and Broderick, T. (2019). A higher-order Swiss army infinitesimal jackknife. *arXiv preprint arXiv:1907.12116*.

Gustafson, P. (2000). Local robustness in Bayesian analysis. In *Robust Bayesian Analysis*, pages 71–88. Springer.

Hampel, F. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393.

Hampel, F. (1986). *Robust statistics: the approach based on influence functions*, volume 196. Wiley-Interscience.

Hansen, L. and Sargent, T. (2008). *Robustness*. Princeton University Press.

He, X., Jurečková, J., Koenker, R., and Portnoy, S. (1990). Tail behavior of regression estimators and their breakdown points. *Econometrica: Journal of the Econometric Society*, pages 1195–1214.

Huber, P. (1983). Minimax aspects of bounded-influence regression. *Journal of the American Statistical Association*, 78(381):66–72.

Karlan, D. and Zinman, J. (2011). Microcredit in theory and practice: Using randomized credit scoring for impact evaluation. *Science*, 332(6035):1278–1284.

Kim, T. and White, H. (2004). On more robust estimation of skewness and kurtosis. *Finance Research Letters*, 1(1):56–73.

Krantz, S. and Parks, H. (2012). *The implicit function theorem: History, theory, and applications*. Springer Science & Business Media.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. (2017). Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474.

Leamer, E. (1984). Global sensitivity results for generalized least squares estimates. *Journal of the American Statistical Association*, 79(388):867–870.

Leamer, E. (1985). Sensitivity analyses would help. *The American Economic Review*, 75(3):308–313.

Maclaurin, D., Duvenaud, D., and Adams, R. (2015). Autograd: Effortless gradients in numpy. In *ICML 2015 AutoML Workshop*, volume 238.

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

Meager, R. (2019). Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics*, 11(1):57–91.

Meager, R. (2020). Aggregating distributional treatment effects: A Bayesian hierarchical analysis of the microcredit literature. *LSE working paper*.

Mises, R. (1947). On the asymptotic distribution of differentiable statistical functions. *The Annals of Mathematical Statistics*, 18(3):309–348.

Mosteller, F. and Tukey, J. (1977). *Data Analysis and Regression: A Second Course In Statistics*. Pearson, USA.

Reeds, J. (1976). *On the definition of von Mises functionals*. PhD thesis, Statistics, Harvard University.

Saltelli, A. (2004). Global sensitivity analysis: An introduction. In *Proc. 4th International Conference on Sensitivity Analysis of Model Output (SAMO '04)*, pages 27–43.

Sobol, I. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation*, 55(1-3):271–280.

Tarozzi, A., Desai, J., and Johnson, K. (2015). The impacts of microcredit: Evidence from Ethiopia. *American Economic Journal: Applied Economics*, 7(1):54–89.

Trefethen, L. and Bau, D. (1997). *Numerical linear algebra*, volume 50. Siam.

Van der Vaart, A. (2000). *Asymptotic statistics*, volume 3. Cambridge University Press.

Young, A. (2019). Consistency without inference: Instrumental variables in practical application. http://personal.lse.ac.uk/YoungA/CWOI.pdf. Accessed: 2020-11-27.

Zeileis, A., Köll, S., and Graham, N. (2020). Various versatile variances: An object-oriented implementation of clustered covariances in R. *Journal of Statistical Software*, 95(1):1–36.

Appendix A Motivating Examples

Example 1 (Linear regression)**.** Let y_n denote the response and x_n denote the regressors for linear regression. In this case, $d_n = (y_n, x_n)$ and the regression coefficient solves Eq. 2.2 with $G(\theta, d_n) = x_n(y_n - x_n^T\theta)$. This can be thought of as a method of moments estimator imposing the condition that the residuals $y_n - x_n^T\theta$ be orthogonal to the regressors.

Example 2 (Smooth optimization problems)**.** Suppose that $f(\theta, d_n) \in \mathbb{R}$ is a three-times continuously differentiable objective function, and we wish to find θ solving $\hat{\theta} := \text{argmin}_\theta \frac{1}{N}\sum_{n=1}^N f(\theta, d_n)$. Under appropriate regularity conditions, the optimization problem is equivalent to satisfying the first order condition $\frac{1}{N}\sum_{n=1}^N \frac{\partial f(\theta,d_n)}{\partial\theta}\Big|_{\hat{\theta}} = 0$. So the optimization problem is equivalent to Eq. 2.2 with $G(\theta, d_n) = \frac{\partial f(\theta,d_n)}{\partial\theta}\Big|_\theta$.

Regression is, of course, a special case of a smooth optimization problem, as are generalized method of moments estimators. Many common econometrics methods can be cast as optimization problems, but not all, as the following example shows.

Example 3 (Instrumental variables)**.** Suppose we have responses y_n, regressors x_n, and instruments z_n. Then $d_n = (y_n, x_n, z_n)$ and the instrumental variables (IV) estimator θ solves Eq. 2.2 with $G(\theta, d_n) = z_n(y_n - \theta^T x_n)$. Note that the IV estimator does not, in general solve an optimization problem. It if did, then $\frac{1}{N}\sum_{n=1}^N \partial G(\theta, d_n)/\partial\theta = \frac{1}{N}\sum_{n=1}^N z_n x_n^T$ would be the Hessian matrix of the objective function, but this is impossible, because $\frac{1}{N}\sum_{n=1}^N z_n x_n^T$ is not symmetric in general.

Example 4 (Sequences of optimization problems)**.** Suppose that $\theta = (\theta_1, \theta_2)$, and we estimate θ_1 by first solving an optimization problem

$$\hat{\theta}_1 := \text{argmin}_{\theta_1} \frac{1}{N}\sum_{n=1}^N f_1(\theta_1, d_n),$$

and then use $\hat{\theta}_1$ as a hyperparameter for a subsequent optimization problem:

$$\hat{\theta}_2 := \text{argmin}_{\theta_2} \frac{1}{N}\sum_{n=1}^N f_2(\hat{\theta}_1, \theta_2, d_n).$$

Jointly, this is equivalent to solving Eq. 2.2 with

$$G(\theta, d_n) = \begin{pmatrix} \frac{1}{N}\sum_{n=1}^N \frac{\partial f_1(\theta_1,d_n)}{\partial\theta_1}\Big|_{\theta_1} \\ \frac{1}{N}\sum_{n=1}^N \frac{\partial f_2(\theta_1,\theta_2,d_n)}{\partial\theta_2}\Big|_{\theta_1,\theta_2} \end{pmatrix}.$$

Example 5 (Sample mean)**.** A simple sample mean is a case where S_α and Ψ_α are analytically tractable. Let $\vec{x}$ be a vector of N scalar observations, and let $G(\theta, x_n) = \theta - x_n$, so that $\hat{\theta} = \frac{1}{N}\sum_{n=1}^N x_n$. The re-weighted estimate is given by $\hat{\theta}(\vec{w}) = \sum_{n=1}^N w_n x_n / \sum_{n=1}^N w_n$. We will take $\phi(\theta, \vec{w}) = \theta$. Without loss of generality,

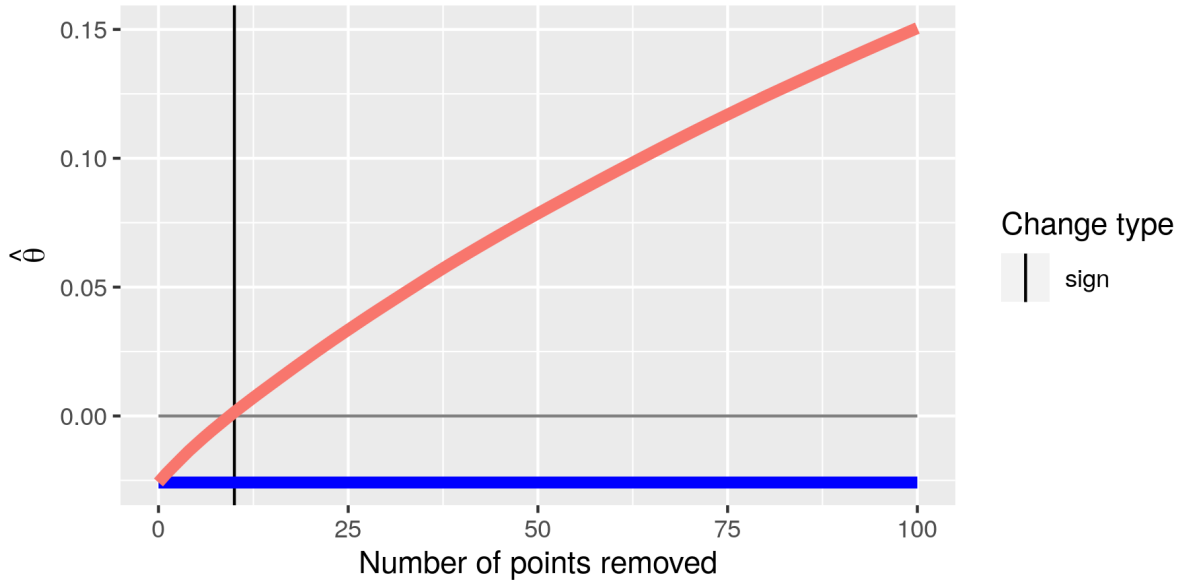

Figure 11: A graph of Example 5 for $1 \leq M \leq 100$ using 1000 standard normal datapoints. The horizontal blue line shows the original sample mean, and the red curve shows $\hat{\theta}(\vec{w})$ as more and more points are left out. The vertical black line shows that the sign of $\hat{\theta}(\vec{w})$ changes when 10 points are left out.

let $\vec{x}$ be sorted so that $x_1 \leq x_2 \leq \ldots \leq x_N$. When there are N' points remaining in the sample, the effect on $\hat{\theta}$ of removing datapoint n is $-x_n/N'$. Consequently, the most influential datapoint to remove is always the most negative observation remaining, so $S_\alpha = \{1 \ldots M\}$, $\hat{\theta}(\vec{w}^*) = \frac{1}{N-M} \sum_{n=M+1}^{N} x_n$, and $\Psi_\alpha = \hat{\theta}(\vec{w}^*) - \hat{\theta}$.

Figure 11 shows this analysis for 1000 standard normal datapoints which happened to have a sample mean of -0.026. We took $M \leq 100$ ($\alpha < 10\%$). The difference between the red curve, which shows $\hat{\theta}(\vec{w})$, and the horizontal blue line, which shows $\hat{\theta}$, is Ψ_α for increasing M. The re-weighted $\hat{\theta}(\vec{w})$ crosses zero at $M = 10$ ($\alpha = 1\%$). If the conclusions of an analysis rested on the fact that the sign of $\hat{\theta}$ were negative, one would take $\Delta = 0.026$ and find that $M = 10$ sufficed to produce a change of size Δ and overturn the analysis. If 1% of the sample were not considered too large, then the analysis would not be considered robust.

Example 6 (Linear regression)**.** Consider linear regression as defined in Example 1. In this case,

$$\hat{\theta}(\vec{w}) = \left(\sum_{n=1}^{N} w_n x_n x_n^T \right)^{-1} \sum_{n=1}^{N} w_n y_n x_n.$$

The fact that the weights occur in the $\left(\sum_{n=1}^{N} w_n x_n x_n^T\right)^{-1}$ term mean that the ordering of the datapoints' influence depends on which datapoints have already been removed. For example, suppose that the regressors would be nearly colinear were it not for two similar datapoints, x_1 and x_2. The effect of removing only one of x_1 or x_2 alone might be small, but the effect of removing both of them could be

very large. For this reason, exactly computing S_α and Ψ_α for linear regression is hard in general.

Note that if one were to "fix" the regressors in the re-weighting, defining the weighted estimating equation

$$\frac{1}{N}\sum_{n=1}^{N} x_n \left(w_n y_n - x_n^T \hat{\theta}(\vec{w})\right) = 0,$$

then S_α and Ψ_α would have closed forms as in Example 5. For the remainder of the paper we take the view that removing a datapoint involves removing the regressor as well, though which is most appropriate depends on the context.

Example 7 (Optimization.)**.** In the setting of Example 2, the Jacobian of $G(\theta, d_n)$ is the Hessian of the objective function $f(\theta, d_n)$, and Eq. 2.10 takes the form

$$\left.\frac{d\hat{\theta}(\vec{w})}{d\vec{w}^T}\right|_{\vec{w}} = -\left(\sum_{n=1}^{N} w_n \left.\frac{\partial^2 f(\theta,\vec{w})}{\partial\theta\partial\theta^T}\right|_{\hat{\theta}(\vec{w}),\vec{w}}\right)^{-1} \left(\left.\frac{\partial f(\theta,d_1)}{\partial\theta}\right|_{\hat{\theta}(\vec{w})}, \ldots, \left.\frac{\partial f(\theta,d_N)}{\partial\theta}\right|_{\hat{\theta}(\vec{w})}\right).$$

Note that, in order for Eq. 2.10 to apply, the Hessian matrix must be non-degenerate at $\hat{\theta}(\vec{w}), \vec{w}$.

Example 8 (A component of $\hat{\theta}$.)**.** When ϕ simply picks out one entry of the vector θ, i.e. $\phi(\theta) = \theta_p$, then ψ_n is simply the (p, n)-th entry of the matrix $d\hat{\theta}(\vec{w})/d\vec{w}^T$.

Appendix B Asymptotic Properties of the Influence Function

B.1 Covariance of M-estimators: standard and robust versions.

Suppose we have an objective function, g, that decomposes as a sum over datapoints, d_n, $n = 1, ..., N$. Let $x = (d_1, ..., d_N)$. Let an estimate of the parameter $\theta \in \mathbb{R}^P$ be defined as a root of the summed objective function, i.e.,

$$\hat{\theta} := \theta \text{ such that } \sum_{n=1}^{N} g\left(d_n, \theta\right) =: G\left(x, \theta\right) = 0. \tag{B.1}$$

By definition, $G\left(\hat{\theta}, x\right) = 0$. The MLE of smooth likelihoods is such an estimator, where g is the gradient of the log likelihood. Let θ_0 denote the "true" value (that is, the root of Eq. B.1). Assume all the smoothness and regularity you need, Taylor

expand a single term around θ_0, and evaluate at $\hat{\theta}$ to get

$$0 = G\left(x, \hat{\theta}\right) = G\left(x, \theta_0\right) + \left.\frac{dG}{d\theta^T}\right|_{\theta_0}\left(\hat{\theta} - \theta_0\right) + O\left(\left\|\hat{\theta} - \theta_0\right\|^2\right) \Rightarrow$$

$$\hat{\theta} - \theta_0 = -\left(\left.\frac{dG}{d\theta^T}\right|_{\theta_0}\right)^{-1} G\left(x, \theta_0\right) + O\left(\left\|\hat{\theta} - \theta_0\right\|^2\right). \tag{B.2}$$

This is sort of a "master formula", of which different regression standard errors arise as special cases.

B.2 Correctly specified likelihoods

First, suppose that g is the gradient of a correctly specified log likelihood, $\ell\left(d_n, \theta\right)$. Then

$$\begin{aligned} g\left(d_n, \theta\right) &= \nabla \ell\left(d_n, \theta\right) \\ G\left(x, \theta_0\right) &= \sum_{n=1}^{N} \nabla \ell\left(d_n, \theta_0\right) \\ \left.\frac{dG}{d\theta^T}\right|_{\theta_0} &= \sum_{n=1}^{N} \nabla^2 \ell\left(d_n, \theta_0\right) \end{aligned}$$

By standard properties of correctly-specified likelihoods,

$$\mathbb{E}\left[\nabla \ell\left(d_n, \theta\right)\right] = 0, \tag{B.3}$$

and

$$\mathrm{Cov}\left(\nabla \ell\left(d_n, \theta_0\right)\right) = \mathcal{I}, \tag{B.4}$$

where $\mathcal{I}$ is the Fisher information. By the law of large numbers, and again a property of correctly-specified likelihoods,

$$\frac{1}{N}\left.\frac{dG}{d\theta^T}\right|_{\theta_0} = \frac{1}{N}\sum_{n=1}^{N} \nabla^2 \ell\left(d_n, \theta_0\right) = \mathbb{E}\left[\nabla^2 \ell\left(d_n, \theta_0\right)\right] \to -\mathcal{I}. \tag{B.5}$$

By the Central limit theorem,

$$\frac{1}{\sqrt{N}} G\left(x, \theta_0\right) = \frac{1}{\sqrt{N}}\sum_{n=1}^{N} \nabla \ell\left(d_n, \theta_0\right) \overset{dist}{\to} \mathcal{N}\left(0, \mathrm{Cov}\left(\nabla \ell\left(d_n, \theta_0\right)\right)\right) = \mathcal{N}\left(0, \mathcal{I}\right). \tag{B.6}$$

By smoothness assumptions, $O\left(\sqrt{N}\left\|\hat{\theta}-\theta_0\right\|^2\right) \to 0$. Putting this all together in Eq. B.2,

$$\begin{aligned}\sqrt{N}\left(\hat{\theta}-\theta_0\right) &= -\left(\frac{1}{N}\left.\frac{dG}{d\theta^T}\right|_{\theta_0}\right)^{-1}\frac{1}{\sqrt{N}}G\left(x,\theta_0\right)+O\left(\sqrt{N}\left\|\hat{\theta}-\theta_0\right\|^2\right)\\ &\overset{dist}{\to}\mathcal{I}^{-1}\mathcal{N}\left(0,\mathcal{I}\right)\\ &=\mathcal{N}\left(0,\mathcal{I}^{-1}\right).\end{aligned}$$

Typically, $\mathcal{I}^{-1}$ is estimated with the negative inverse Hessian of the log likelihood, i.e., the "observed Fisher information":

$$\hat{\mathcal{I}} := -\frac{1}{N}\sum_{n=1}^{N}\nabla^2\ell\left(d_n,\hat{\theta}\right) = -\frac{1}{N}\left.\frac{dG}{d\theta^T}\right|_{\hat{\theta}}.$$

I will briefly observe that this could be motivated by considering the non-asymptotic sensitivity estimator at $\hat{\theta}$ rather than θ_0. The formal difference is merely forming the Taylor expansion around $\hat{\theta}$ and not θ_0. The conceptual difference is interesting, but beyond the scope.

B.3 Robustness to misspecification.

Let us consider how results change if the likelihood is not correct. In this case, we are simply calculating the asymptotic behavior of a smooth optimization problem. The results of Section B.2 used the fact that the model was correctly specified in two places. First, it was used in Eq. B.3. This is not a particularly material assumption; we simply define the "true" θ_0 as the one for which Eq. B.3 is true. Less benign is the assumption that the quantity $\mathcal{I}$ in Eq. B.4 and Eq. B.5 is the same. This is not, in general, true. Let us define

$$\mathbb{E}\left[\nabla^2\ell\left(d_n,\theta_0\right)\right] := -H.$$

Then, everything else follows as before, except we get

$$V = \mathbb{E}\left[\mathrm{Cov}\left(\nabla\ell\left(d_n,\theta_0\right)\right)\right],$$

where the d_n is drawn according to whatever distribution generates your data – they must be independent, but might not be identically distributed. And then,

$$\begin{aligned}\sqrt{N}\left(\hat{\theta}-\theta_0\right) &\overset{dist}{\to} H^{-1}\mathcal{N}\left(0,V\right)\\ &= \mathcal{N}\left(0,H^{-1}VH^{-1}\right).\end{aligned} \tag{B.7}$$

Now we need to estimate two different quantities.

$$\begin{aligned}\hat{H} &:= -\frac{1}{N}\sum_{n=1}^{N}\nabla^2\ell\left(d_n,\hat{\theta}\right)\\ \hat{V} &:= \frac{1}{N}\sum_{n=1}^{N}\nabla\ell\left(d_n,\hat{\theta}\right)\nabla\ell\left(d_n,\hat{\theta}\right)^T.\end{aligned} \tag{B.8}$$

Now, $\hat{V}$ is the sample variance of the observed scores (using the fact that $\frac{1}{N}\sum_{n=1}^{N}\nabla\ell\left(d_n,\hat{\theta}\right) = 0$ by definition). In Section B.2 we effectively assumed that $\hat{H} = \hat{V}$. In finite sample, even this is not necessarily true; it was motivated in Section B.2 by purely asymptotic assumptions.

Note that, in regression problems, $\nabla\ell\left(d_n,\hat{\theta}\right)$ is given by

$$\begin{aligned}\ell\left(d_n,\theta\right) &= \frac{1}{2}\left(y_n - d_n^T\theta\right)^2\\ \nabla\ell\left(d_n,\hat{\theta}\right) &= d_n\left(y_n - d_n^T\hat{\theta}\right).\end{aligned}$$

This shows that Eq. B.7 is in fact the standard heteroskedasticity-consistent robust standard error for regressions.

B.4 Robustness to within-group covariances.

Finally, let us consider grouping data together. Both Section B.2 and Section B.3 required that $d_1, ..., d_N$ be independent of one another. Otherwise, the expected score covariance as measured by Eq. B.5 or by Eq. B.8 is not the same variance to be used in the Central limit theorem, Eq. B.6.

As a first step, note that, even when the $d_1, ..., d_N$ are believed to be independent, there is in fact a choice to be made. For if $d_1, ..., d_N$ are independent, then the pairs $(d_1, d_2), (d_3, d_4), ..., (d_{N-1}, d_N)$ are also independent. So it is perfectly well-motivated to write

$$\begin{aligned}z_m &= \left(d_{2m+1}, d_{2m+2}\right), \text{ for } m = 1, ..., \frac{N}{2}\\ \ell\left(z,\theta\right) &= \sum_{m=1}^{M}\ell\left(z_m,\theta\right)\\ &= \sum_{m=1}^{M}\left(\ell\left(d_{2m+1},\theta\right) + \ell\left(d_{2m+2},\theta\right)\right)\\ &= \sum_{n=1}^{N}\ell\left(d_n,\theta\right)\\ &= \ell\left(x,\theta\right)\end{aligned} \tag{B.9}$$

and so replace Eq. B.8 with

$$\hat{V}_{\text{paired}} := \frac{1}{M}\sum_{m=1}^{M} \nabla\ell\left(z_m, \hat{\theta}\right) \nabla\ell\left(z_m, \hat{\theta}\right)^T \tag{B.10}$$
$$= \frac{1}{N/2}\sum_{m=0}^{N/2-1} \left(\nabla\ell\left(d_{2m+1}, \hat{\theta}\right) + \nabla\ell\left(d_{2m+2}, \hat{\theta}\right)\right)\left(\nabla\ell\left(d_{2m+1}, \hat{\theta}\right) + \nabla\ell\left(d_{2m+2}, \hat{\theta}\right)\right)^T.$$

The existence of cross-terms in Eq. B.10 shows that $\hat{V}_{\text{paired}}$ is not equal to $\hat{V}$ in Eq. B.8, though if the d_n are truly independent then the difference disappears asymptotically. The same argument could be made for any grouping, or, indeed, groupings of different sizes where you imagine that the group size is independent and random.

We see no reason to use $\hat{V}_{\text{paired}}$ if you believe that $d_1, ..., d_N$ are independent. However, if you believe that some of the d_n are dependent within a certain grouping, but that the groups are independent of one another, then you can simply re-write the problem using these groups:

$$z_m = \left(d_{g_m(1)}, ..., d_{g_m(N_m)}\right), \text{ for } m = 1, ..., M,$$

and use the reasoning of Section B.3 applied to the grouped random variables as in Eq. B.10. A simple example is the problem we're considering, where the groups are villages, and the d_n are observations for people. The objective function is unchanged, as seen in Eq. B.9. The score covariance estimator will be noisier in general, since it contains the cross-terms that would otherwise be absent in Eq. B.8. But this may be a small price to pay for a correct model specification.

Appendix C Bounds on the shape parameter

As this is α times the truncated mean of a non-generate distribution, Γ_α^+ will not tend to 0 as N increases. Indeed, it will remain large to the extent that this truncated mean is large, which occurs when the scale of the γ_n variable is large. In fact, it is possible to derive an upper bound on this object, to understand the worst-case scenario of maximal sensitivity for a fixed N and α given some observed standard errors σ_ϵ. Consider the problem of choosing $\{\gamma_n\}_{n=1}^N$ to maximize Γ_α^+ subject to the two constraints that $N^{-1}\sum_{n=1}^N \gamma_N^2 = 1$ and $N^{-1}\sum_{n=1}^N \gamma_N = 0$. For a given α and N, what matters is the value of the influence scores of the data points we are going to discard: this is the set for which $n \in \hat{S}_\alpha$, and we denote members of this set by γ_m for $m = 1, 2, ...M$ where applicable. Thus the Lagrangian form of the problem that defines the worst case scenario is

$$\mathcal{L} = \inf_{\lambda_\mu, \lambda_\sigma} \max_{\{\gamma_n\}_{n=1}^N} \left(\sum_{m=1}^{M} \gamma_m + \lambda_\mu \sum_{n=1}^{N} \gamma_n + \frac{1}{2}(\sum_{n=1}^{N} \gamma_n^2 - N)\right) \tag{C.1}$$

Taking first order conditions with respect to both some candidate γ_m and some retained γ_n (with slight abuse of notation),

$$\begin{aligned} 1 + \lambda_\mu + \lambda_\sigma \gamma_m &= 0 \\ \lambda_\mu + \lambda_\sigma \gamma_n &= 0 \end{aligned}$$

Summing over the indices in the first equation and plugging in the constraint $\sum_{n=1}^{N} \gamma_n = 0$ gives $\lambda_\mu = -M/N = -\alpha$. Squaring the second equation and summing over the indices and using the constraint $\sum_{n=1}^{N} \gamma_n^2 = N$ delivers $\lambda_\sigma^2 = \alpha(1-\alpha)$. Putting these together, and employing the negative root of λ_σ, we have

$$\begin{aligned} \gamma_m =& \frac{-(1+\lambda_\mu)}{\lambda_\sigma} \\ =& \frac{1-\alpha}{\sqrt{\alpha(1-\alpha)}} \\ =& \sqrt{\frac{1-\alpha}{\alpha}} \end{aligned}$$

This is the worst possible value of some candidate γ_m. Thus, plugging that into the definition of Γ_α^+, in general it turns out that

$$\Gamma_\alpha^+ \leq \sqrt{\alpha(1-\alpha)} \tag{C.2}$$

and that this upper bound is attained when exactly M values are omitted and when γ_m from these omitted points all take the same value.

Appendix D IV and OLS Proofs

D.1 Proof of Theorem 1

In this section we prove the key step in Theorem 1 using the results from Giordano et al. (2019), which we will abbreviate as HOIJ. For example, we will refer to Giordano et al. (2019, Assumption 1) as HOIJ Assumption 1. Additionally, for this section only, we will exclusively use the notation of Giordano et al. (2019). Our Section 3.2 draws the connection between this section's notation and the notation of the rest of the present paper.

Suppose we observe regressors, $x_n \in \mathbb{R}^P$, instruments $z_n \in \mathbb{R}^P$, and responses, $y_n \in \mathbb{R}$, for $n = 1, \ldots, N$. We assume that the observations are exchangeable. It will be convenient to define the residual $\varepsilon_n(\theta) := \theta^T x_n - y_n$. Our estimating equation

partial derivatives are then

$$\begin{aligned} G(\theta, \vec{w}) :=& = \frac{1}{N} \sum_{n=1}^{N} w_n \varepsilon_n(\theta) z_n \\ d_\theta^1 G(\theta, \vec{w}) =& H(\theta, w) := \frac{1}{N} \sum_{n=1}^{N} w_n z_n x_n^T \\ d_\theta^k G(\theta, \vec{w}) =& 0 \text{ for } k > 1. \end{aligned}$$

We now consider each of the assumptions and conditions of HOIJ Section 4.1 in this context, proving the following lemma:

Lemma 2. *Under the assumption $\alpha C_{op} \xi_1 \leq \frac{1}{3}$ of Theorem 1 of Section 3.2, HOIJ Assumptions 1-5 are satisfied with $\rho = 1/3$ and*

$$\begin{aligned} \Omega_\theta(\mathscr{B}) &:= \left\{ \theta : \left\| \theta - \hat{\theta} \right\|_2 < \mathscr{B} \right\} \quad \textit{with} \\ \mathscr{B} &= \frac{\left\| \hat{\theta}_1^{\mathrm{IJ}}(w) - \hat{\theta} \right\|_2 + 2\alpha^2 \tilde{C}_{op}^2 \xi_1 \xi_2}{1 - 2\alpha^2 \tilde{C}_{op}^2 \xi_1^2} > 0. \end{aligned}$$

Proof. For HOIJ Assumption 1, we will use HOIJ Lemma 8 to choose a suitable value of $\mathscr{B}$ once the other constants are established.

HOIJ Assumption 2 follows immediately by inspection, as the estimating equation is linear in θ.

In this case, $H(\theta, 1_N) = \hat{H} = \frac{1}{N} \sum_{n=1}^{N} z_n x_n^T$ does not depend on θ. So HOIJ Assumption 3 is satisfied with

$$\text{HOIJ Assumption 3:} \quad C_{op} := \left\| \hat{H}^{-1} \right\|_{op}.$$

HOIJ Assumption 4 is satisfied for $k \geq 2$ with $M_k = 0$. For $k = 1$, we observe again that $d_\theta^1 G(\theta, 1_N)$ does not depend on θ, and so we can take

$$\begin{aligned} \text{HOIJ Assumption 4:} \quad M_1 &:= \left\| \hat{H} \right\|_2 \\ M_2 &:= 0, \ldots. \end{aligned}$$

Note that HOIJ Assumption 4 does not require control for $k = 0$.

HOIJ Assumption 5 is more involved and depends on the types of weights we are considering. Let us adopt the notation of Section 3.2 that w is zero in entries indexed by a set $\mathscr{N} \subseteq [N]$ and one otherwise. It will turn out to be convenient to define the following quantities (which match the definitions of Theorem 1 but in the

present section's notation).

$$\begin{aligned}\alpha :=& \frac{|\mathscr{N}|}{N} \\ \xi_1 :=& \left\| \frac{1}{|\mathscr{N}|} \sum_{n \in \mathscr{N}} z_n x_n^T \right\|_2 \\ \xi_2 :=& \left\| \frac{1}{|\mathscr{N}|} \sum_{n \in \mathscr{N}} \varepsilon_n(\hat{\theta}) z_n \right\|_2 .\end{aligned}$$

The stochastic and asymptotic behavior of ξ_1 and ξ_2 will obviously depend on how the weights w are chosen and on the tail behavior of x_n and y_n.

For $k \geq 2$, we can take $\lambda_k = 0$. For $k = 1$, $d_\theta^1 G(\theta, w)$ is again independent of θ, so we have

$$\sup_{\theta \in \Omega_\theta(\mathscr{B})} \left\| d_\theta^1 G(\theta, w) - d_\theta^1 G(\theta, 1_N) \right\|_2 = \alpha \xi_1.$$

For $k = 0$, we must rely on the definition of $\Omega_\theta(\mathscr{B})$.

$$\begin{aligned}&\sup_{\theta \in \Omega_\theta(\mathscr{B})} \|G(\theta, w) - G(\theta, 1_N)\|_2 \\ =&\alpha \sup_{\theta \in \Omega_\theta(\mathscr{B})} \left\| \frac{1}{|\mathscr{N}|} \sum_{n=1}^N \varepsilon_n(\theta) z_n \right\|_2 \\ =&\alpha \sup_{\theta \in \Omega_\theta(\mathscr{B})} \left\| \frac{1}{|\mathscr{N}|} \sum_{n=1}^N \left(\varepsilon_n(\theta) + \varepsilon_n(\hat{\theta}) - \varepsilon_n(\hat{\theta}) \right) z_n \right\|_2 \\ \leq&\alpha \left(\sup_{\theta \in \Omega_\theta(\mathscr{B})} \left\| (\theta - \hat{\theta})^T \frac{1}{|\mathscr{N}|} \sum_{n \in \mathscr{N}} z_n x_n^T \right\|_2 + \xi_2 \right) \\ \leq&\alpha \left(\mathscr{B} \xi_1 + \xi_2 \right).\end{aligned}$$

We can then make use of HOIJ Lemma 1 to set

$$\begin{aligned}\text{HOIJ Assumption 5}: \quad \lambda_0 :=& \alpha \left(\mathscr{B} \xi_1 + \xi_2 \right) \\ \lambda_1 :=& \alpha \xi_1 \\ \lambda_2 :=& 0, \ldots\end{aligned}$$

Next, we turn to HOIJ Condition 1, for which we will require that

$$\text{HOIJ Condition 1:} \quad \rho := C_{op} \lambda_1 + C_{op}^2 M_2 \lambda_0 = \alpha C_{op} \xi_1 < 1.$$

Should HOIJ Condition 1 fail to be satisfied, then our bounds cannot be applied, essentially because $d_\theta^1 G(\theta, w)$ is not smooth enough to guarantee the strong convexity of $H(\tilde{w})$ using HOIJ Lemma 2. Supposing that HOIJ Condition 1 is satisfied with $\rho < 1$, we define $\tilde{C}_{op} := \frac{1}{1-\rho} C_{op}$ as in HOIJ Lemma 2. In Theorem 1, we take

$\rho = 1/3$, for reasons to be described below.

Finally, we can calculate the approximation and error bounds (though we cannot use the bounds to control the error until we have chosen $\mathscr{B}$ to satisfy HOIJ Lemma 8). Let us consider $k_{\mathrm{IJ}} = 1$ for simplicity. We have

$$\begin{aligned}
\hat{\theta}_1^{\mathrm{IJ}}(w) &= \hat{\theta} + \delta_w^1 \hat{\theta}(1_N) \\
&= \hat{\theta} + \alpha \left(\frac{1}{N} \sum_{n=1}^{N} x_n x_n^T \right)^{-1} \frac{1}{|\mathscr{N}|} \sum_{n \in \mathscr{N}} x_n \varepsilon_n(\hat{\theta}) \quad \text{(HOIJ Equation 3)} \\
\left\| \hat{\theta}_1^{\mathrm{IJ}}(w) - \hat{\theta}(w) \right\|_2 &\leq \tilde{C}_{op}^3 M_2 \lambda_0^2 + 2\tilde{C}_{op}^2 \lambda_1 \lambda_0 + \tilde{C}_{op}^3 \lambda_2 \lambda_0^2 \quad \text{(HOIJ Corollary 3)} \\
&= 2\alpha^2 \tilde{C}_{op}^2 \xi_1 (\mathscr{B}\xi_1 + \xi_2).
\end{aligned}$$

In order to apply HOIJ Lemma 8 to show that $\Omega_\theta(\mathscr{B})$ satisfies HOIJ Assumption 1, it will suffices to choose $\mathscr{B}$ so that

$$\begin{aligned}
\mathscr{B} &> \left\| \hat{\theta}_1^{\mathrm{IJ}}(w) - \hat{\theta} \right\|_2 + 2\alpha^2 \tilde{C}_{op}^2 \xi_1 (\mathscr{B}\xi_1 + \xi_2) \Leftrightarrow \\
\mathscr{B} &> \frac{\left\| \hat{\theta}_1^{\mathrm{IJ}}(w) - \hat{\theta} \right\|_2 + 2\alpha^2 \tilde{C}_{op}^2 \xi_1 \xi_2}{1 - 2\alpha^2 \tilde{C}_{op}^2 \xi_1^2}.
\end{aligned} \tag{D.1}$$

As long as $2\alpha^2 \tilde{C}_{op}^2 \xi_1^2 < 1$, a positive solution to Eq. D.1 exists and can be readily computed from the data at hand.

Recall that, to satisfy HOIJ Condition 1 above, we took $\rho = \alpha C_{op} \xi_1 < 1$. Noting this fact allows the interpretation of Eq. D.1 as an additional, stricter condition on ρ, since

$$\begin{aligned}
2\alpha^2 \tilde{C}_{op}^2 \xi_1^2 &\leq 1 \Leftrightarrow \\
\left(\frac{\rho}{1-\rho} \right)^2 &< \frac{1}{2} \Leftrightarrow \\
\rho \left(1 + \frac{1}{\sqrt{2}} \right) &< \frac{1}{\sqrt{2}} \Leftrightarrow \\
\rho &< \frac{1}{1+\sqrt{2}} < 1.
\end{aligned}$$

Consequently, to satisfy both HOIJ Condition 1 and HOIJ Lemma 8, it suffices to have

$$\alpha C_{op} \xi_1 < \frac{1}{1+\sqrt{2}}, \tag{D.2}$$

which can be satisfied by requiring $\alpha C_{op} \xi_1 < \rho := 1/3$, the value assumed in Theorem 1.

□

D.2 Proof of Lemma 1

Proof. For a fixed ϕ, $\hat{\theta}$, and $\vec{w}$, define the linear interpolation functions of a scalar $t \in [0,1]$,

$$\begin{aligned}\theta(t) :=& \hat{\theta} + t(\hat{\theta}(\vec{w}) - \hat{\theta}) \\ \vec{\omega}(t) :=& N^{-1}\left(\vec{1} + t(\vec{w} - \vec{1})\right),\end{aligned}$$

and then define the function $\Phi(t)$ as

$$\Phi(t) := \phi(\theta(t), N\vec{\omega}(t)).$$

Thus, $\Phi(1) = \phi(\vec{w})$ and $\Phi(0) = \phi(\vec{1})$. By the fundamental theorem of calculus,

$$\begin{aligned}\Phi(1) - \Phi(0) = \int_0^1 \left.\frac{\partial \Phi(t)}{\partial t}\right|_{t=t'} dt' \Rightarrow \\ \left|\phi(\vec{w}) - \phi(\vec{1})\right| \leq \sup_{t' \in [0,1]} \left|\left.\frac{\partial \Phi(t)}{\partial t}\right|_{t=t'}\right|.\end{aligned} \tag{D.3}$$

By the chain rule,

$$\left.\frac{\partial \Phi(t)}{\partial t}\right|_t = \left.\frac{\partial \phi(\theta, \vec{w})}{\partial \theta^T}\right|_{\theta(t), N\vec{\omega}(t)} (\hat{\theta}(\vec{w}) - \hat{\theta}) + \left.\frac{\partial \phi(\theta, N\vec{\omega})}{\partial \vec{\omega}^T}\right|_{\theta(t), N\vec{\omega}(t)} N^{-1}(\vec{w} - \vec{1}). \tag{D.4}$$

So, by the triangle inequality and the Cauchy-Schwartz,

$$\begin{aligned}\left|\phi(\vec{w}) - \phi(\vec{1})\right| \leq& C_\theta \left\|\hat{\theta}(\vec{w}) - \hat{\theta}\right\|_2 + C_\omega N^{-1} \left\|\vec{w} - \vec{1}\right\|_2 \\ \leq& \left(C_\theta C_{\text{diff}} + C_\omega\right) \alpha.\end{aligned}$$

proving the first result.

Next, recall from Eqs. 2.4 and 2.7 that

$$\phi^{\text{lin}}(\vec{w}) - \phi(\vec{1}) = \left.\frac{\partial \phi(\theta, \vec{w})}{\partial \theta^T}\right|_{\hat{\theta}, \vec{1}} (\hat{\theta}^{\text{lin}}(\vec{w}) - \hat{\theta}) + \left.\frac{\partial \phi(\theta, N\vec{\omega})}{\partial \vec{\omega}^T}\right|_{\hat{\theta}, \vec{1}} N^{-1}(\vec{w} - \vec{1}).$$

Combining with Eqs. D.3 and D.4,

$$\begin{aligned}
&\phi(\vec{w}) - \phi^{\text{lin}}(\vec{w}) \\
&= \left(\int_0^1 \left. \frac{\partial \phi(\theta, \vec{w})}{\partial \theta^T} \right|_{\theta(t), N\vec{\omega}(t)} dt \right) (\hat{\theta}(\vec{w}) - \hat{\theta}) - \left. \frac{\partial \phi(\theta, \vec{w})}{\partial \theta^T} \right|_{\hat{\theta}, \vec{1}} (\hat{\theta}^{\text{lin}}(\vec{w}) - \hat{\theta}) + \\
&\quad \left(\int_0^1 \left. \frac{\partial \phi(\theta, N\vec{\omega})}{\partial \vec{\omega}^T} \right|_{\theta(t), N\vec{\omega}(t)} dt \right) N^{-1}(\vec{w} - \vec{1}) - \left. \frac{\partial \phi(\theta, N\vec{\omega})}{\partial \vec{\omega}^T} \right|_{\hat{\theta}, \vec{1}} N^{-1}(\vec{w} - \vec{1}) \\
&= \left(\int_0^1 \left. \frac{\partial \phi(\theta, \vec{w})}{\partial \theta^T} \right|_{\theta(t), N\vec{\omega}(t)} dt - \left. \frac{\partial \phi(\theta, \vec{w})}{\partial \theta^T} \right|_{\hat{\theta}, \vec{1}} \right) (\hat{\theta}(\vec{w}) - \hat{\theta}) + \\
&\quad \left. \frac{\partial \phi(\theta, \vec{w})}{\partial \theta^T} \right|_{\hat{\theta}, \vec{1}} (\hat{\theta}(\vec{w}) - \hat{\theta}^{\text{lin}}(\vec{w})) + \\
&\quad \left(\int_0^1 \left. \frac{\partial \phi(\theta, N\vec{\omega})}{\partial \vec{\omega}^T} \right|_{\theta(t), N\vec{\omega}(t)} dt - \left. \frac{\partial \phi(\theta, N\vec{\omega})}{\partial \vec{\omega}^T} \right|_{\hat{\theta}, \vec{1}} \right) N^{-1}(\vec{w} - \vec{1}).
\end{aligned}$$

By the Lipschitz property of the partial derivatives,

$$\begin{aligned}
&\sup_{t \in [0,1]} \left\| \left. \frac{\partial \phi(\theta, \vec{w})}{\partial \theta} \right|_{\theta(t), N\vec{\omega}(t)} dt - \left. \frac{\partial \phi(\theta, \vec{w})}{\partial \theta} \right|_{\hat{\theta}, \vec{1}} \right\|_2 \\
&\quad \leq L_\theta \sup_{t \in [0,1]} \left(\left\| \theta(t) - \hat{\theta} \right\|_2 + \left\| \vec{\omega}(t) - \vec{1} \right\|_2 \right). \\
&\quad \leq L_\theta \left(\left\| \hat{\theta}(\vec{w}) - \hat{\theta} \right\|_2 + N^{-1} \left\| \vec{w} - \vec{1} \right\|_2 \right). \\
&\quad \leq L_\theta \left(C_{\text{diff}} + 1 \right) \alpha,
\end{aligned}$$

with analogous reasoning giving

$$\sup_{t \in [0,1]} \left\| \left. \frac{\partial \phi(\theta, N\vec{\omega})}{\partial \vec{\omega}} \right|_{\theta(t), N\vec{\omega}(t)} dt - \left. \frac{\partial \phi(\theta, N\vec{\omega})}{\partial \vec{\omega}} \right|_{\hat{\theta}, \vec{1}} \right\|_2 \leq L_\omega \left(C_{\text{diff}} + 1 \right) \alpha.$$

Applying these Lipschitz bounds, together with the triangle inequality and Cauchy-Schwartz, then give

$$\begin{aligned}
\left| \phi(\vec{w}) - \phi^{\text{lin}}(\vec{w}) \right| &\leq L_\theta \left(C_{\text{diff}} + 1 \right) \alpha \cdot C_{\text{diff}} \alpha + C_\theta C_{\text{lin}} \alpha^2 + L_\omega \left(C_{\text{diff}} + 1 \right) \alpha \cdot \alpha \\
&\leq \left(L_\theta \left(C_{\text{diff}} + 1 \right) C_{\text{diff}} + C_\theta C_{\text{lin}} + L_\omega \left(C_{\text{diff}} + 1 \right) \right) \alpha^2,
\end{aligned}$$

which proves the second result.

□