Sensitivity to Model Misspecification

Dmitry Arkhangelskiy
Stanford GSB

Evgeni Drynkin
Stanford GSB

January 18, 2016

Abstract

In this paper we construct a measure that captures sensitivity of estimates to model misspecification in a particular GMM-like framework. This framework includes estimation of treatment effect under unconfoundedness, instrumental variables, demand estimation and general semiparametric models. Our sensitivity measure can be easily computed even for large datasets and complex models. It captures the worst case scenario and can be used to construct worst case bounds for the parameter of interest. Somewhat surprisingly the worst case scenario is connected with a norm of an observable function. We demonstrate its performance using familiar (LaLonde 1986) dataset.

1 Introduction

In this paper we construct a measure of sensitivity of key parameters of an econometric model to misspecification. Our measure quantifies the worst case sensitivity and thus can be used to construct conservative bounds on parameter value.

Our theoretical model is a particular instance of GMM framework (see Hansen 1982). It includes as a special case treatment effect models both with instruments and under unconfoundedness, random coefficient demand models, and some general semiparametric models. One common factor that unites all these models is that there is an unknown nuisance function which is important for estimation of the main parameter. In case of treatment effects this is a control function that is crucial for estimating the main parameter – treatment effect. In case of demand models this is unknown function of attributes that is essential for estimating price elasticity.

It is clear that all models used by applied economists are misspecified: they don't include all necessary variables. This fact by itself isn't important, because it is a priori not clear that the misspecification is harmful. For example, there is a general understanding that in a well-designed experiment one can estimate treatment effects using misspecified model.

Main goal of our paper is to quantify the effect of model misspecification on the key parameters of the model. In order to do this we need to formalize what precisely we mean by misspecification. For this formalization we use a particular GMM-like framework. The main conceptual problem with misspecification is that it is often difficult to disentangle two different effects: change in the model and change in question that researcher is asking.

Formally each time we are considering a new model we are implicitly changing the question that we are trying to answer. As a result, one can always argue that it's not the model that is misspecified but rather the question is different.

This logic is unbreakable and is universally applicable. At the same time, for some applications we can talk about misspecification and keep the main research question essentially constant. This frameworks are characterized by two properties: (i) all model parameters can be divided into key parameters and nuisance parameters; (ii) key parameters can be expressed as a function of nuisance parameters. Then we think that the model is misspecified when nuisance parameters aren't modeled correctly.

Our focus on misspecification is additionally motivated by the following consideration. When economists report their estimates in empirical papers it is customary to report not just the estimated value but also its standard error. The common ground is that parameters without their standard errors are meaningless. It is clear that standard errors capture only one particular source of uncertainty – randomness of data. We can model this randomness differently, either as sampling uncertainty or as pure model-based uncertainty in Bayesian fashion. At the end of the day all different concepts of standard errors all boil down to conceptually similar objects – some sort of sensitivity of estimated value to perturbations in data.

There is one problem with this kind of uncertainty: it assumes that we are using the right model. Moreover, purely mathematically this uncertainty decreases as we have more and more data. This is a built-in property of mathematical models that are used to model uncertainty: independent random variables imply concentration of measure as sample size grows. As a result, with modern-sized datasets we observe economically insignificant standard errors. If we take these errors at face value then we are bound to believe that we essentially know the answer without any uncertainty.

In practice researchers don't take these errors completely seriously because everybody understands that there is a different and arguably more important source of uncertainty – the model used might be wrong. In absence of any formal way to quantify this type of uncertainty researchers engage into several ad hoc practices. The most popular one is to report estimates from several models. Typically in published papers these reports show that key estimates don't change from specification to specification.

There are at least three conceptual problems with this otherwise reasonable practice. First, specifications are chosen informally and there is no way to tell whether the reported specifications are in fact the most illustrative ones. Second, this whole process is a multiple comparison procedure and thus should be treated as such, which requires additional sophisticated statistical analysis. Finally, in complex models one simply can't computationally afford to report results from different specifications. Beginning from (Leamer 1983) the first two problems received significant attention in the econometric literature. These concerns are somewhat understood on the theoretical level but are for the most part ignored in the empirical literature.

Our measure solves all these problems: it is formally defined and quantifies the worst case scenario, it can be consistently estimated with reasonable accuracy and even for complex models it takes seconds to compute. Of course, all these nice properties don't come for free. The price that we pay is the following: we use a specific framework, which, being rich enough, of course doesn't include all interesting models; our measure is inherently local and thus quantifies sensitivity to local perturbations; and finally, we measure the worst case sensitivity which can be far from the actual one.

We demonstrate usefulness of our measure using a familiar (LaLonde 1986) dataset. In his original paper LaLonde poses a very important question: can economists using sophisticated econometric models with observational data replicate the results of experiments? His answer was negative and motivated a long lasting line of research (see Dehejia and Wahba 1999).

LaLonde's data consists of two parts: experimental dataset and observational dataset that designed to mimic experimental data. We use his dataset for that we a priori know that experiment should be robust to model misspecification, while lots of papers, starting from LaLonde's original one, show that observational dataset is highly sensitive to model considered. As a result we expect our measure to be small in experimental dataset and large in the observational one, which is indeed the case.

Our empirical example is there to demonstrate that our measure behaves reasonably in a situation in which we already know relative importance of misspecification. We encourage readers to apply our measure to other studies in which we don't a priori know the sensitivity and report their results.

Related literature: Our paper doesn't readily fit into one particular strand of literature. Many papers discuss model selection and costs of misspecification but most don't propose any measure of sensitivity.

The closed paper in motivation and approach is (Athey and Imbens 2015) in which authors propose a procedure that looks at different splits of sample and reports estimates of main parameters resulting from such splits. Splits are selected by a decision-tree-like algorithm. Although formally very different this approach is similar in spirit to our measure. One advantage of our measure is that we have a closed-form solution that can be estimated in seconds. Moreover our measure quantifies a well-defined worst-case scenario, while approach of (Athey and Imbens 2015) is somewhat arbitrary.

Another paper that is motivated by misspecification and also constructs a measure of sensitivity is (Gentzkow and Shapiro 2015). In this paper authors consider dependence of parameter of interest on different moments of data. We think of our work as complementary to theirs: their focus is mainly on what moments that are used in estimation are most important for the estimates of key parameters. We essentially answer the orthogonal question: how important are those moments that were not included in estimation procedure. We believe that there might be some deep connections between our approaches but we think that this is a question of a separate paper.

Two theoretical examples that we consider throughout the text are related to treatment effect literature. Excellent recent source on these models is (Imbens and Rubin 2015). As we show, in the model with unconfoundedness sensitivity can be interpreted as a particular measure of imbalance. It was long recognized in experimental literature that sample balance is crucial for empirical work. Check (Imbens and Rubin 2015) for the discussion of importance of balance.

From statistical point of view one can see our work as a study of bias. The other side of the story is a study of variance in misspecified models which has a long tradition in statistics and econometrics. For the textbook treatment of inference in misspecified models check (Anatolyev and Gospodinov 2011) which discusses it at length. Variance is an important issue, at the end of the day, we would prefer correct standard errors to incorrect ones, but as sample grows variance goes to zero, while bias might stay large independent of sample size.

For our estimation procedure we use the nonparametric noise estimator of (Liitiäinen, Corona, and Lendasse 2008). This kind of estimators are not very popular among econometricians, despite their simplicity and computational attractiveness. One particular case of use of a similar estimator is presented in (Abadie and Imbens 2006).

From mathematical point of view we are using some simple facts from theory of Hilbert spaces. One can use (Bickel et al. 1993) and (Fernholz 2012) for introduction into Hilbert spaces, related infinite-dimensional calculus concepts and their use in statistics.

2 Model

In this section we discuss probabilistic model that we are using throughout the paper. The most important fact about our model is that it allows to talk about misspecification without changing the question that researchers are trying to answer. This is achieved by focusing on a specific form a misspecification – we assume that nuisance parameters of the probability model can be misspecified while the main parameter always stays the same. We formalize this logic expressing the answer to economic question of interest as a functional of nuisance parameters. Definition of nuisance parameters can change from model to model but the functional stays constant.

We explain this structure on two simple examples: estimation of treatment effect under unconfoundedness and with instrumental variables. Main mathematical apparatus that we are using is that of Hilbert spaces. Reader can safely think about all considered Hilbert spaces as ordinary finite-dimensional Euclidean spaces.

2.1 Notation

We use capital Latin letters $(Y, X, Z, H, \dots)$ to denote observed random variables. We use small latin letters $(f, g, h, c, \dots)$ to denote real-valued functions of random variables. Perhaps, $f(X)$ would have been a better notation, but we reduce it to f for better readability. We use script letters $(\mathcal{F}, \mathcal{S}, \mathcal{C}, \dots)$ for spaces of real-valued functions and measures.

We use $\|\cdot\|_{2,\mu}$ to denote an $L^2(\mu)$ norm and sometimes use $\|\cdot\|_2$ when the underlying measure if clear from the discussion. Small greek letters $(\alpha, \beta, \gamma, \dots)$ are used for real-valued fixed parameters. We use a.s. shortcut to describe events that hold almost surely with respect to some underlying measure. We use notation $a := b$ to state that a is defined as b.

2.2 Technical definitions

The most important definition here is the definition of two spaces $\mathcal{F}$ and $\mathcal{S}$ which are used extensively throughout the text. These are just two Hilbert spaces of functions.

Let (X, H) be a pair of random variables which takes values in $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ and is distributed according to some measure μ. Let $(\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}, \sigma(X, H), \mu)$ be a corresponding probability space, where $\sigma(X, H)$ is a (Borel) σ-algebra generated by (X, H).

Assumption 2.1. *Measure μ is absolutely continuous with respect to fixed known σ-finite measure ν on $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$.*

Assumption 2.2. *Let K be a fixed compact subset of $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$. Under ν random pair $(X, H) \in K$ almost surely.*

For any σ-algebra $\mathcal{A}$ let $m\mathcal{A}$ be a set of all real-valued random variables measurable with respect to $\mathcal{A}$. Let $\mathcal{F} = \{f \in m\sigma(X, H) : \int f^2 d\nu < \infty\}$ and let $\mathcal{S} = \{f \in m\sigma(X) : \int f^2 d\nu < \infty\}$. That is $\mathcal{F}$ is a set of square-integrable functions of (X, H) and $\mathcal{S}$ is a set of square-integrable functions of X only. Let $\|f\|_{2,\nu}$ be an $L^2(\nu)$ norm of f, that is $\|f\|_{2,\nu} = \left(\int f^2 d\nu\right)^{\frac{1}{2}}$. By construction both $\mathcal{F}$ and $\mathcal{S}$ are separable Hilbert spaces (under $\|\cdot\|_{2,\nu}$) and $\mathcal{S}$ is a closed linear subspace of $\mathcal{F}$.

Let $C > 0$ be some fixed constant and define the following set of measures: $\mathcal{M} = \{\mu : \mu \geq 0,\, \nu(\mu) = 1, \mu \ll \nu \text{ and } \frac{d\mu}{d\nu} < C\}$. That is $\mathcal{M}$ is a set of probability measures absolutely continuous with respect to ν and with bounded density (Radon-Nikodym derivative).

Observe that for any $\mu \in \mathcal{M}$ we have that $\|\cdot\|_{2,\mu} \leq C\|\cdot\|_{2,\nu}$ and thus both $\mathcal{F}$ and $\mathcal{S}$ are Hilbert spaces for any μ (closed subspaces of $L^2(\mu)$). This is the main reason why we need Assumption 2.1 and consider μ with bounded Radon-Nikodym derivative. Of course there are weaker assumptions that will guarantee the same result, but we choose these two for their simplicity.

2.3 GMM framework

In this paper we are using a familiar GMM framework (see Hansen 1982), although our focus is somewhat different from the usual one. Our main objective is to work with misspecified models and in order to do this we need to introduce two sets of moment conditions: one set describes the "true" model, while the second set describes the model that a researcher is going to use in practice.

One particular conceptual problem with misspecified models is that typically when the researcher changes the model she frequently (and typically implicitly) changes the **question** she is trying to answer. This makes misspecification controversial – one can always argue that the answer is right, but the question is different.

Below we consider a specific type of models and a specific type of misspecification that allows one to talk about changing the model without essentially changing the question. Clearly, the logic "different model"-"different question" still applies, but we believe that in our setup it is mostly sophistical.

To introduce the first model we need some additional definitions. Let l be some fixed number, let $K \leq \infty$ and consider the following three functions: $\psi_1 : \mathbb{R} \times \mathbb{R} \times \mathbb{R}^{d_2} \times \mathbb{R}^l \to \mathbb{R}$, $\psi_2 : \mathbb{R} \times \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \times \mathbb{R}^l \to \mathbb{R}^K$ and $\psi_3 : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \times \mathbb{R}^l \to \mathbb{R}^l$.

Our set of moments is given by the following system:

$$\begin{cases} \mathbb{E}_\mu[\psi_1(\theta, f(X), H, \gamma)] = 0 \\ \mathbb{E}_\mu[\psi_2(f(X), X, H, \gamma)] = 0 \\ \mathbb{E}_\mu[\psi_3(X, H, \gamma)] = 0 \end{cases} \tag{1}$$

where $f \in \mathcal{S}$, $\theta \in \mathbb{R}$, $\gamma \in \mathbb{R}^l$ and $\mu \in \mathcal{M}$. We emphasize that both ψ_1 and ψ_2 have as its argument value $f(X)$, not a function f. At the same time solution of system (1) defines some particular $f^\star$.

In this paper we abstract from all technical issues regarding existence and uniqueness of solution to (1). We make the following assumption:

Assumption 2.3. *All moment condition systems considered in the paper have a unique solution.*

Some sufficient conditions under which this assumption holds can be found in many different papers and we are not going to state them here. Let $(\theta^\star, \gamma^\star, f^\star)$ be the solution of system (1). By construction we have that $f^\star \in \mathcal{S}$.

The main parameter of interest in system (1) is θ, while f and γ are two distinct nuisance parameters.[1] One should think about $\gamma \in \mathbb{R}^l$ as a low-dimensional parameter. If n is a potential sample size, then $l \ll n$. Note that γ and f are defined separately from θ. Second moment condition is K-dimensional, where, potentially, K can be infinite, and thus our framework includes conditional moment restrictions.

Our main interest in system (1) is in its first condition:

$$\mathbb{E}_\mu[\psi_1(\theta, f(X), H, \gamma)] = 0 \tag{2}$$

We view restriction (2) as a **definition** of the main parameter of interest θ given f, γ and μ. In order to make this view operational we need additional assumption.

Let $\Phi : \mathbb{R} \times \mathcal{S} \times \mathbb{R}^l \times \mathcal{M} \to \mathbb{R}$ be defined in the following way: $\Phi(\theta, f, \mu, \gamma) := \mathbb{E}_\mu[\psi_1(\theta, f(X), H, \gamma)]$.

Assumption 2.4. *For any $(f, \mu, \gamma) \in \mathcal{S} \times \mathcal{M} \times \mathbb{R}^l$ equation $\Phi(\theta, f, \mu, \gamma) = 0$ has a unique solution. Denote it by $\tau : \mathcal{S} \times \mathcal{M} \times \mathbb{R}^l \to \mathbb{R}$, that is $\Phi(\tau(f, \mu, \gamma), f, \mu, \gamma) = 0$ for all (f, μ, γ).*

We explicitly assume existence, instead of giving some sufficient conditions on function ψ_1 and measure μ that will guarantee it. The main reason for this is our belief that in applications it will be more or less straightforward to verify this assumption directly, constructing the functional τ, at least locally, instead of verifying some sufficient conditions which necessarily will be to stringent to be universally applicable.

Functional τ will be the main object of interest in this paper. Its meaning is straightforward: given the nuisance parameters f, γ and measure μ it gives us the “answer” to the question of interest.

System (1) describes the “true” model. However, we assume that in reality researcher will use a different system. Let $\tilde{\psi_2} : \mathbb{R} \times \mathbb{R} \times \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \times \mathbb{R}^l \times \mathbb{R}^m \to \mathbb{R}^m$ and $\tilde{\psi_3} : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \times \mathbb{R}^{\tilde{l}} \to \mathbb{R}^{\tilde{l}}$, where $\tilde{l} > l$ and consider the following set of moments:

$$\begin{cases} \mathbb{E}_\mu[\psi_1(\theta, f(\eta, X), H, \gamma_1)] = 0 \\ \mathbb{E}_\mu[\tilde{\psi_2}(f(\eta, X), \theta, X, H, \gamma_1, \gamma_2, \eta] = 0 \\ \mathbb{E}_\mu[\tilde{\psi_3}(X, H, \gamma_1, \gamma_2)] = 0 \end{cases} \tag{3}$$

[1] Technically speaking in this setup we have more nuisance parameters than just (f, γ), since system (1) need not fully specify the probability model. In what follows we will be interested only in f and thus we won't need to define these additional nuisances.

where $\eta \in \mathbb{R}^m$, $\gamma_1 \in \mathbb{R}^l$, $\gamma_2 \in \mathbb{R}^{\tilde{l}-l}$ and $f(\eta) \in \mathcal{S}$ for each η. The main difference between systems (1) and (3) is that nuisance parameters are defined differently. Note that parameter η is m-dimensional, and one should think of m as small relative to reasonable sample size. The primary reason why researches might use system (3) instead of "true" system (1) is the simplicity of the former. We look at two particular examples of such simplification in the next subsection.

The main parameter θ is still defined using the same moment condition and thus we can use the same functional τ to answer the question of interest. Let $(\theta^{\star\star}, \eta^{\star\star}, \gamma_1^{\star\star}, \gamma_2^{\star\star})$ be the solution of system (3), then we have $\theta^{\star\star} = \tau(f(\eta^{\star\star}), \gamma_1^{\star}, \mu)$.

This emphasizes once again the role of τ as a **defintion** of the parameter of interest. Researchers can use different sets of moment conditions to pin down the nuisance parameters but the definition of the main parameter stays the same. Specific structure of systems (1) and (3), mainly the fact that ψ_1 is always used to define θ, allows us to meaningfully talk about changing the "model" f without changing the "question" τ.

2.4 Examples

System (1) is general enough to include lots of applications of interest. We will discuss two simple applications below. In both application functional τ will be linear in f.[2]

2.4.1 Constant treatment effects under unconfoundedness

We work in traditional potential outcome framework (see Imbens and Rubin 2015). Let $H = (Y, T)$ where Y is an outcome variable and T is a treatment assignment; X is a vector of attributes. Let $Y(T)$ be a potential outcome function, then by definition we have the following:

$$Y = TY(1) + (1-T)Y(0) = \mathbb{E}_\mu[Y(0)] + \mathbb{E}_\mu[Y(0) - \mathbb{E}_\mu[Y(0)]|X] + \\ (Y(1) - Y(0))\,T + \varepsilon \quad (4)$$

Assuming that $Y(1) - Y(0) = \theta$, we have the following:

$$Y = \alpha + f(X) + \theta T + \varepsilon \quad (5)$$

where $\alpha = \mathbb{E}_\mu[Y(0)]$, $f(X) = \mathbb{E}_\mu[Y(0) - \mathbb{E}_\mu[Y(0)]|X]$ and $\varepsilon = Y(0) - \mathbb{E}_\mu[Y(0)|X]$. We assume identifying moment condition (unconfoundedness):

$$\mathbb{E}_\mu[\varepsilon|T, X] = 0 \quad (6)$$

which implies the following:

$$\begin{cases} \mathbb{E}_\mu[(Y - f(X) - \theta T)(T - \mathbb{E}_\mu[T])] = 0 \\ \mathbb{E}_\mu[Y - \mathbb{E}_\mu[Y|T=0] - f(X)|X, T=0] = 0 \end{cases} \Rightarrow \\ \begin{cases} \theta = \frac{\mathbb{E}_\mu[(Y - f(X))(T - \mathbb{E}_\mu[T])]}{\mathbb{V}_\mu[T]} \\ f(X) = \mathbb{E}_\mu\left[Y - \mathbb{E}_\mu[Y|T=0]|X, T=0\right] \end{cases} \quad (7)$$

[2] Technically speaking these functionals are affine but we abuse definition slightly for better readability.

In this setup we are interested in the following functional:

$$\tau(f,\mu,\gamma) = \frac{\mathbb{E}_{\mu}\left[(Y-f(X))(T-\mathbb{E}_{\mu}[T])\right]}{\mathbb{V}_{\mu}[T]} \tag{8}$$

Note, that it indeed depends on μ and f; also observe that for a fixed μ this functional is in fact linear in f.

It is logical to view (8) as a **definition** of a parameter of interest in this setup. Observe that $f(X)$ is non-parametrically identified here separately from θ using a conditional moment restriction in (7). At the same time, in empirical work one is likely to use a particular parametric family for f (e.g., first-order polynomials) and might use different moment conditions to identify parameters of this family. This potentially leads to a misspecification of f. At the same time, this misspecification doesn't change definition (8). We present a particular case below.

Clearly this example fits in our general GMM framework. Let $\{g_1,\dots,\}$ be a countable basis of $\mathcal{S}$. In this case we have the following functions:

$$\begin{cases} \psi_1(\theta,f,H,\gamma) = (Y-f(X)-\theta T)(T-\gamma_1) \\ \psi_2(f,X,H,\gamma) = (Y-\gamma_2-f(X))(1-T)g_i, \quad \text{for } i=1,\dots,n,\dots \\ \psi_3(X,H,\gamma) = \begin{pmatrix} T-\gamma_1 \\ Y(1-T)-\gamma_2(1-T)\end{pmatrix} \end{cases} \tag{9}$$

In this case the second moment restriction is in fact infinite-dimensional and in practice researchers will somehow simplify it. One way to accomplish it is to assume a particular parametric structure on f.

Let η be a k-dimensional parameter and consider a potentially nonlinear parameterization $f(x)=f(\eta,x)$. For this specification researchers are likely to derive moment restriction as the first order conditions to the following problem:

$$\mathbb{E}_{\mu}\left[(Y-\alpha-f(\eta,X)-\theta T)^2\right] \to \min_{\alpha,\eta,\theta} \tag{10}$$

In this case we can rewrite solution to this problem in the following way:

$$\begin{cases} \mathbb{E}_{\mu}\left[(Y-f(\eta,X)-\theta T)(T-\gamma_1)\right]=0 \\ \mathbb{E}_{\mu}\left[(Y-\alpha-f(\eta,X)-\theta T)\left(\frac{\partial f(\eta,X)}{\partial\eta}\right)^T\right]=0 \\ \mathbb{E}_{\mu}\left[(Y-\alpha-f(\eta,X)-\theta T)\right]=0 \\ \mathbb{E}_{\mu}\left[(T-\gamma_1)\right]=0 \end{cases} \tag{11}$$

where we assume all necessary differentiability and integrability conditions. In this case we have the following definition of $\tilde{\psi}_2$:

$$\tilde{\psi}_2 = \begin{pmatrix} ((Y-\alpha-f(\eta,X)-\theta T)\left(\frac{\partial f(\eta,X)}{\partial\eta}\right)^T \\ Y-\alpha-f(\eta,X)-\theta T \end{pmatrix} \tag{12}$$

Note that $\tilde{\psi}_2$ depends on θ.

Clearly, there are other potential models that researcher can consider in order to go from the general model (9) to something more tractable. In our next example we will discuss another way to simplify the model.

2.4.2 Constant treatment effects with one sided compliance

Let $H = (Y, T, Z)$, where Z is a binary instrumental variable, and all other variables are the same as in the previous example. One sided compliance implies that $T\{Z = 0\} = 0$. We again assume that $Y(1) - Y(0) = \theta$ and end up with the same model:

$$Y = \alpha + f(X) + \theta T + \varepsilon \tag{13}$$

where $\alpha = \mathbb{E}_\mu[Y(0)]$, $f(X) = \mathbb{E}_\mu[Y(0) - \mathbb{E}_\mu[Y(0)]|X]$ and $\varepsilon = Y(0) - \mathbb{E}_\mu[Y(0)|X]$. This time we assume a different moment restriction:

$$\mathbb{E}_\mu[\varepsilon|X, Z] = 0 \tag{14}$$

which now implies the following:

$$\begin{cases} \mathbb{E}_\mu[(Y - f(X) - \theta T)(Z - \mathbb{E}_\mu[Z])] = 0 \\ \mathbb{E}_\mu[Y - \mathbb{E}_\mu[Y|Z = 0] - f(X)|X, Z = 0] = 0 \end{cases} \Rightarrow$$

$$\begin{cases} \theta = \frac{\mathbb{E}_\mu[(Y - f(X))(Z - \mathbb{E}_\mu[Z])]}{\mathbb{E}_\mu[T(Z - \mathbb{E}_\mu[Z])]} \\ f(X) = \mathbb{E}_\mu\left[Y - \mathbb{E}_\mu[Y|Z = 0]|X, Z = 0\right] \end{cases} \tag{15}$$

In this setup we are interested in the following functional:

$$\tau(f, \mu, \gamma) = \frac{\mathbb{E}_\mu\left[(Y - f(X))(Z - \mathbb{E}_\mu[Z])\right]}{\mathbb{E}_\mu\left[T(Z - \mathbb{E}_\mu[Z])\right]} \tag{16}$$

which is again linear in f for a fixed μ.

As before (16) can be viewed as a definition of parameter of interest which depends on some non-parametrically identified function f. Again, in practice one can use different restrictions to identify f, but still use (16) to compute the parameter of interest. This allows us to consider misspecification of f separately from definition of treatment effect.

This example fits in the general GMM framework using the following functions ($\{g_1, \dots\}$ is a basis of $\mathcal{S}$ as before):

$$\begin{cases} \psi_1(\theta, f, H, \gamma) = (Y - f(X) - \theta T)(Z - \gamma_1) \\ \psi_2(f, X, H, \gamma) = (Y - \gamma_2 - f(X))(1 - Z)g_i, \quad \text{for } i = 1, \dots, n, \dots \\ \psi_3(X, H, \gamma) = \begin{pmatrix} Z - \gamma_1 \\ Y(1 - Z) - \gamma_2(1 - Z) \end{pmatrix} \end{cases} \tag{17}$$

In reality researchers might use the following slightly different and simpler set of moment restrictions:

$$\begin{cases} \mathbb{E}_\mu\left[(Y - \sum_{i=1}^m \beta_i g_i - \theta T)(Z - \gamma)\right] = 0 \\ \mathbb{E}_\mu\left[(Y - \alpha - \sum_{i=1}^m \beta_i g_i - \theta T)g_i\right] = 0, \quad \text{for } i = 1, \dots, m \\ \mathbb{E}_\mu\left[(Y - \alpha - \sum_{i=1}^m \beta_i g_i - \theta T)\right] = 0 \\ \mathbb{E}_\mu\left[(Z - \gamma)\right] = 0 \end{cases} \tag{18}$$

To construct moment conditions in (18) we used restriction (14) directly and first m functions from $\{g_1, \dots, \}$. In this case $\tilde{\psi_2}$ is given by the following function:

$$\tilde{\psi_2} = \begin{pmatrix} (Y - \alpha - \sum_{i=1}^{m} \beta_i g_i - \theta T) g_i \quad \text{for } i = 1, \dots, m \\ Y - \alpha - \sum_{i=1}^{m} \beta_i g_i - \theta T \end{pmatrix} \tag{19}$$

In practice researchers might experiment with different bases and select m based on some sample size related procedure.

2.5 Discussion

The main purpose of this section was to formally define τ in a way that allows us to discuss misspecification. The fact that τ is defined for any (f, γ) allows us to change definition of nuisance without changing the question.

The price that we pay for this convenience is high: there are other important forms of misspecification that are left out of the picture. Of course we would like to consider frameworks in which ψ_1 is misspecified as well but we don't know how to do this in a coherent way. At the same time, lots of applications either fall directly in our framework or can be reframed in a way that would fit in.

Indeed, when writing a model researcher has in mind several different ways to set it up, without changing the question she is trying to answer. This implies that there is separation (at least mental) of the question and the model that is used to answer it. In our framework we formalize this separation in one particular way which was chosen for its simplicity and operational convenience.

Important limitation of our discussion is that both f and θ are assumed to be one-dimensional. There is nothing in the current setup that precludes both of them to be multidimensional. We focus on one-dimensional setup because of its simplicity and because it already spans lots of applications.

Another important limitation of our framework is that we focus on exactly identified systems. Technically, this is not a limitation, because any estimation method that we are familiar with, transforms overidentified system into exactly identified system.[3] On the conceptual level over-identified systems are tricky: if we allow for misspecification then over-identified system might not have a solution (in population) which implies that the definition of parameter of interest depends on the method that we are using to solve the over-identified system. This is an important point by itself and we believe that it deserves a separate treatment.

3 Sensitivity

In this section we define the most important object of the paper – sensitivity of τ to local perturbations in f. Sensitivity is nothing more than just a particular kind of derivative of τ with respect to f. The main gain comes from the fact that this derivative can be easily computed, estimated in sample and interpreted.

[3] For example, this is true for GMM that EL. In the former we transform the overidentified system using random linear transformation that depends on the weight matrix, while in the latter we add parameters in such a way that the resulting system is exactly identified.

3.1 Motivation

In notation of the previous section the main object of interest is $\tau(f^\star, \mu, \gamma^\star)$ while in practice researchers focus on $\tau(f^{\star\star}, \mu, \gamma^{\star\star})$. Starting from this section we will use f for $f^\star$ and $\hat{f}$ for $f^{\star\star}$. We assume that $\gamma^\star = \gamma^{\star\star}$ and thus the only problem that researcher faces is that $f \neq \hat{f}$. We abuse notation and write $\tau(f, \mu)$, suppressing dependence on γ.

In this setup the main object of interest is the size of discrepancy between $\tau(\hat{f}, \mu)$ and $\tau(f, \mu)$: $|\tau(\hat{f}, \mu) - \tau(f, \mu)|$. We separate this bias into two parts:

$$|\tau(\hat{f}, \mu) - \tau(f, \mu)| = \|\hat{f} - f\|_2 \frac{|\tau(\hat{f}, \mu) - \tau(f, \mu)|}{\|\hat{f} - f\|_2} = e \times \delta \tag{20}$$

where $e = \|\hat{f} - f\|_2$ and $\delta = \frac{|\tau(\hat{f},\mu) - \tau(f,\mu)|}{\|\hat{f} - f\|_2}$.

This factorization is useful because it allows us to think about the bias as a combination of the size of misspecification e and sensitivity to misspecification δ. In practice it would be typically impossible to asses neither e nor δ. At the same time, operationally these two objects are quite different: in order to bound e we need to have a good idea what f looks like, while δ is in a sense a derivative and thus might be somehow bounded without any knowledge of f.

In both our examples we have an operational definition of f – it is a particular conditional expectation. This fact will later allow us to compute e. However, in other setups f can be just a nuisance parameter that doesn't have any operational definition and thus it will be hard to estimate e. On more abstract level we might be unsure what we actually mean by "true" f and might have several exclusive definitions of it.

In practice it is often unreasonable to assume that e is huge – typically we choose specifications that we believe are close to the "correct" one. As a result, one can somewhat arbitrarily bound e by $\alpha\|\hat{f}\|_2$ for some small α (e.g., $\alpha = 0.05$).

Small e suggests that in order to bound δ we should focus on behavior of τ in the neighborhood of $\hat{f}$. This makes our approach inherently local, which at the end allows us to bound δ by an easily computable quantity.

3.2 Definition

Throughout this section we assume that μ is known, essentially working in population. Given the functional τ and function $\hat{f}$ we would like to compute sensitivity of τ to small changes in $\hat{f}$. Since we are working in a linear space of functions it is natural to consider perturbations from some closed subspace of $\mathcal{S}$. Let $\mathcal{G}$ be (potentially infinite dimensional) subspace of $\mathcal{S}$.[4] We define sensitivity in the following way:

Definition 3.1. *Given $\mathcal{G}$, τ, $\hat{f}$ and $\varepsilon > 0$ define $\delta(\tau, \hat{f}, \mathcal{G}, \varepsilon)$ as follows:*

$$\delta(\tau, \hat{f}, \mathcal{G}, \varepsilon) \equiv \sup_{g \in \mathcal{G}: \|g\|_2 \leq \varepsilon} \left\{ \frac{\tau(\hat{f} + g, \mu) - \tau(\hat{f}, \mu)}{\varepsilon} \right\} \tag{21}$$

[4] Note that since $\mathcal{G}$ is a linear subspace it contains zero function.

Essentially $\delta(\tau, \hat{f}, \mathcal{G}, \varepsilon)$ quantifies a maximal sensitivity given a ε-small perturbation of $\hat{f}$. This definition has two major problems: it depends on the ε and it is completely non-operational – we can't compute $\delta(\tau, \hat{f}, \mathcal{G}, \varepsilon)$ with it. We solve both of these problems assuming that τ can be well approximated by a linear functional up to second order terms.

Assumption 3.2. *For any $\mu \in \mathcal{M}$, τ is a Fréchet differentiable functional of f.*

In this subsection we explicitly assume differentiability of τ. We will give sufficient conditions for this assumption to hold in the GMM framework in the next subsection.

By definition of differentiability we have that τ can be approximated in L^2 norm by the continuous linear functional. Since we are working in Hilbert space of functions, by the Reisz representation theorem we have the following corollary:

Corollary 3.3. *For each fixed μ and f τ has the following representation:*

$$\tau(f+g,\mu) = \tau(f,\mu) + \mathbb{E}_\mu[\nabla_\tau(f)g] + o(\|h\|) \tag{22}$$

where $\nabla_\tau(f) \in \mathcal{F}$.

By definition $\nabla_\tau(f)$ is a function from $\mathcal{F}$. Since h in the corollary above is the member of $\mathcal{S}$ it is clear that $\nabla_\tau(f)$ can't be unique – only its projection on $\mathcal{S}$ matters for the linear part. As we will show below it doesn't matter for the sensitivity. In our examples we will select a particular version of $\nabla_\tau(f)$.

We assumed differentiability of the functional in order to use its linear part in the definition of sensitivity. This leads as to the following:

Definition 3.4. *Given τ, $\hat{f}$ and $\mathcal{G}$ linearized sensitivity is defined by the following expression:*

$$\Delta(\tau,\hat{f},\mathcal{G}) \equiv \sup_{g\in\mathcal{G}:\|g\|_2\le\varepsilon}\left\{\frac{\mathbb{E}_\mu[\nabla_\tau(f)g]}{\varepsilon}\right\} = \sup_{g\in\mathcal{G}:\|g\|_2\le 1}\left\{\mathbb{E}_\mu[\nabla_\tau(\hat{f})g]\right\} \tag{23}$$

Note that dependence on ε goes away, since in the linear case it is just scaling. As a result we can assume that $\varepsilon = 1$.

Observe that by construction we have the following relationship:

$$|\delta(\tau,\hat{f},\mathcal{G},\varepsilon) - \Delta(\tau,\hat{f},\mathcal{G})| \le \sup_{g\in\mathcal{G}:\|g\|_2\le\varepsilon}\left\{\left|\frac{\tau(\hat{f}+g,\mu)-\tau(\hat{f},\mu)}{\varepsilon} - \frac{1}{\varepsilon}\mathbb{E}_\mu[\nabla_\tau(f)g]\right|\right\} \tag{24}$$

Fréchet differentiability is a very strong notion of differentiability that implies that the right-hand side of the inequality above goes to zero. This allows us to state the following corollary.

Corollary 3.5. $\Delta(\tau,\hat{f},\mathcal{G}) = \lim_{\varepsilon\to 0}\delta(\tau,\hat{f},\mathcal{G},\varepsilon)$

Corollary shows that linearized sensitivity is a limit of the ordinary one as ε goes to zero. Linearized sensitivity has one main advantage over the ordinary sensitivity: it can be easily computed, as the following trivial lemma shows.

Lemma 3.6. *Let $P_{\mathcal{G}}$ be an operator of orthogonal projection on $\mathcal{G}$. Then $\Delta(\tau, \hat{f}, \mathcal{G}) = \|P_{\mathcal{G}}(\nabla_\tau(\hat{f}))\|_2$.*

Proof. Since $\mathcal{G}$ is a closed linear subspace operator $P_{\mathcal{G}}$ is well defined. Since $\nabla_\tau(\hat{f})$ is a vector from $\mathcal{F}$ and $\mathcal{G}$ is a closed subspace of $\mathcal{F}$ we have the following representation: $\nabla_\tau(\hat{f}) = \lambda g^\star + h$, where $g^\star \in \mathcal{G}$, $\|g^\star\|_2 = 1$ and h is orthogonal to $\mathcal{G}$. As a result $\Delta(\tau, \hat{f}, \mathcal{G}) = |\lambda| \sup_{g \in \mathcal{G}: \|g\|_2 \le 1} \{\mathbb{E}_\mu[g^\star g]\} = |\lambda| = \|P_{\mathcal{G}}(\nabla_\tau(\hat{f}))\|_2$. □

This lemma shows that linearized sensitivity is nothing more than a length of a projection of a certain vector on the linear subspace $\mathcal{G}$. This suggests a way of actually computing sensitivity in samples and emphasizes the role of vector $\nabla_\tau(\hat{f})$. If one can easily find $\nabla_\tau(\hat{f})$ then the sensitivity can be computed. As we show below this is true in our examples.

3.3 Examples

3.3.1 Constant treatment effects under unconfoundedness

In this example our functional is given by the following formula:

$$\tau(f, \mu) = \frac{\mathbb{E}_\mu\left[(Y - f(X))(T - \mathbb{E}_\mu[T])\right]}{\mathbb{V}_\mu[T]} \tag{25}$$

As we discussed before this a linear functional and thus one can readily observe that Freshet derivative is equal to the following:

$$\nabla_\tau(f) = \frac{(T - \mathbb{E}_\mu[T])}{\mathbb{V}_\mu[T]} \tag{26}$$

Observe that $\nabla_\tau(f)$ is observable in sample up to recentering and scaling (two unknown constants which are trivial to estimate). As a result, sensitivity will be large if function from $\mathcal{G}$ can predict T well enough. Also observe that sensitivity will be large if $\mathbb{V}_\mu[T]$ is small which will happen if distribution of T is skewed: either control or treatment group is disproportionally large.

In this case we can actually represent sensitivity in slightly different way. Assume that $\mathcal{G}$ is a finite dimensional linear subspace and let $\{g_1, \dots, g_d\}$ be an orthonormal basis of this space. By construction we have the following:

$$\mathbb{E}_\mu[\nabla_\tau(f) g_i] = \frac{\mathbb{E}_\mu[(T - \mathbb{E}_\mu[T]) g_i]}{\mathbb{V}_\mu[T]} = \mathbb{E}_\mu[g_i | T = 1] - \mathbb{E}_\mu[g_i | T = 0] = \beta_i \tag{27}$$

and one can see that β_i is essentially a measure of **imbalance** in distribution of T with respect to direction g_i. By orthogonality we have the following:

$$P_{\mathcal{G}}(\nabla_\tau(f)) = \sum_{i=1}^d \beta_i g_i \tag{28}$$

and as a result $\|P_{\mathcal{G}}(\nabla_\tau(f))\|_2 = \sqrt{\sum_{i=1}^d \beta_i^2}$. We can see that in this case sensitivity is nothing more than a particular measure of imbalance with respect to directions $\{g_1, \dots, g_d\}$. This suggests that experimental studies in which covariates are well balance across the groups should have low sensitivity.

3.3.2 Constant treatment effects with one sided compliance

In this example functional is just an IV-type projection coefficient:

$$\tau(f,\mu) = \frac{\mathbb{E}_\mu\left[(Y - f(X))(Z - \mathbb{E}_\mu[Z])\right]}{\mathbb{E}_\mu\left[T(Z - \mathbb{E}_\mu[Z])\right]} \tag{29}$$

which is linear as in the previous example and thus we have the following expression for the derivative:

$$\nabla_\tau(f) = \frac{(Z - \mathbb{E}_\mu[Z])}{\mathbb{E}_\mu\left[T(Z - \mathbb{E}_\mu[Z])\right]} \tag{30}$$

We can see that in this example sensitivity will be large if we can predict Z well enough using function from $\mathcal{G}$. As before $\nabla_\tau(\hat{f})$ is observable up to two unknown constants. In denominator we have a covariance between Z and T and the smaller it is, the larger is sensitivity. Small covariance is an indicator of a weak instrument.

3.4 Sensitivity in GMM framework

In this case our functional is given implicitly by the following moment condition:

$$\mathbb{E}_\mu[\psi_1(\tau(f,\mu), f(X), H)] = 0 \tag{31}$$

Consider a variation of f in direction g with step λ. Proceeding heuristically we have the following:

$$\mathbb{E}\left[\frac{\partial\psi_1(\theta,f,H)}{\partial\theta}|_{\theta=\tau(f,\mu)}d\tau + \frac{\partial\psi_1(\tau(f,\mu),\theta,H)}{\partial\theta}|_{\theta=f(X)}gd\lambda\right] = 0 \Rightarrow$$

$$\frac{d\tau(f+\lambda g,\mu)}{d\lambda} = -\frac{\mathbb{E}\left[\left(\frac{\partial\psi_1(\tau(f,\mu),\theta,H)}{\partial\theta}|_{\theta=f(X)}\right)g\right]}{\mathbb{E}\left[\frac{\partial\psi_1(\theta,f,H)}{\partial\theta}|_{\theta=\tau(f,\mu)}\right]} \tag{32}$$

This derivation suggests that in this case $\nabla_\tau(f)$ is given by the following expression:

$$\nabla_\tau(f) = -\frac{\left(\frac{\partial\psi_1(\tau(f,\mu),\theta,H)}{\partial\theta}|_{\theta=f(X)}\right)}{\mathbb{E}\left[\frac{\partial\psi_1(\theta,f,H)}{\partial\theta}|_{\theta=\tau(f,\mu)}\right]} \tag{33}$$

This intuition is correct if we assume some further smoothness conditions. Recall that $\Phi(\theta,f,\mu) := \mathbb{E}_\mu[\psi_1(\theta,f(X),H)]$.

Theorem 3.7. *Assume that Φ is Fréchet differentiable with respect to its two arguments at point $(\tau(\hat{f},\mu),\hat{f},\mu)$; fix μ, assume that for some $\varepsilon > 0$ the equation $\Phi(\theta,f,\mu) = t$ has as unique solution $\theta(f,t)$, where $|t| \le \varepsilon$; assume that the ratio $\frac{\theta(f+h,t)-\theta(f,0)}{|t|+\|h\|_2}$ is bounded as a function of h and t for (t,h) in some bounded subset of $\mathbb{R}\times\mathcal{S}$. Then the functional τ is Fréchet differentiable in the neighborhood of $\hat{f}$ and its derivative is given by (33).*

Proof. The proof is given in (Fernholz 2012, page 22, theorem 3.2.4).[5] □

Both our examples fit in the GMM framework and it is instructive to check that Fréchet derivatives are in fact given by the expression above. This is indeed the case and is left as an exercise for a reader.

3.5 Constrained sensitivity

Our local measure depends on three things: functional τ, suggested function $\hat{f}$ and $\mathcal{G}$. Among these three, the first two are essentially given, while the third one should be specified. Since linearized sensitivity is a length of a projection on linear subspace $\mathcal{G}$ it is clear that it crucially depends on this subspace. Sensitivity increases once we consider larger and larger $\mathcal{G}$ and the maximal sensitivity is achieved if we use $\mathcal{G} = \mathcal{S}$.

At the same time, it is clear that it is unreasonable to consider some directions. Indeed, typically while we were constructing $\hat{f}$ we optimized over some directions and thus there are no reasons to consider them once again. In order to take this into account we suggest the following procedure: we take $\mathcal{G} = \mathcal{S}$ but instead of unconditional supremum in definition of sensitivity we consider a constrained one.

Natural constraint to put on the functions is to assume that they should satisfy moment conditions that were used in constructing $\hat{f}$. In our notation we suggest the following restriction:

$$g \in \mathcal{S} : \mathbb{E}_\mu[\tilde{\psi}_2(\hat{f} + g, \hat{\theta}, X, H, \hat{\gamma}, \hat{\eta})] = 0 \tag{34}$$

Constrain (34) is a natural one but it is global, while our operational definition of sensitivity is a local one. Moreover, due to non-linearity constrain (34) might contain only $g = 0$. Thus it makes sense to substitute it for the following first oder constraint:

$$g \in \mathcal{S} : \mathbb{E}_\mu \left[\left(\frac{\partial \tilde{\psi}_2(\lambda, \hat{\theta}, X, H, \hat{\gamma}, \hat{\eta})}{\partial \lambda} \right)_{\lambda = \hat{f}} g \right] = 0 \tag{35}$$

Constraint (35) is an extremely easy one: it just describes an orthogonal complement to a m-dimensional space spanned by the derivatives of $\tilde{\psi}_2$. This leads us to the following definition:

Definition 3.8. *Let* $\mathcal{C} = \left\{ g \in \mathcal{S} : g = \nu^T \left(\frac{\partial \tilde{\psi}_2(\lambda, \hat{\theta}, X, H, \hat{\gamma}, \hat{\eta})}{\partial \lambda} \right)_{\lambda = \hat{f}}, \text{ where } \lambda = \mathbb{R}^m \right\}$. *Then the constrained sensitivity* $\Delta^{\mathcal{C}}$ *is defined as the following:* $\Delta^{\mathcal{C}}(\tau, \hat{f}) := \Delta(\tau, \hat{f}, \mathcal{S} \setminus \mathcal{C})$.

Restriction (35) might look artificial or unintuitive but as we show below it is very natural in our examples.

[5] The proof is for Hadamard differentiability but the refinement is straightforward.

Constant treatment effects under unconfoundedness: In this example function $\tilde{\psi}_2$ is given by the equation (12) and as a result the derivatives in (35) are given by the following expression:

$$\left(\frac{\partial\tilde{\psi}_2(\lambda,\hat{\theta},X,H,\hat{\gamma},\hat{\eta})}{\partial\lambda}\right)_{\lambda=\hat{f}} = \begin{pmatrix}\left(\frac{\partial f(\eta,X)}{\partial\eta}\right)_{\eta=\hat{\eta}}\\ 1\end{pmatrix} \tag{36}$$

As a result we can see that in this case we are interested in the orthogonal complement of space spanned by the quasi-regressors evaluated at $\hat{\eta}$.

Constant treatment effects with one sided compliance: In this example function $\tilde{\psi}_2$ is given by the equation (19) and the derivatives are given by the following expression:

$$\left(\frac{\partial\tilde{\psi}_2(\lambda,\hat{\theta},X,H,\hat{\gamma},\hat{\eta})}{\partial\lambda}\right)_{\lambda=\hat{f}} = \begin{pmatrix}g_i \text{ for } i=1,\dots,m\\ 1\end{pmatrix} \tag{37}$$

where $\{g_1,\dots,g_m\}$ are the first m basis vectors. As a result, restriction in (35) boils down to a subspace orthogonal to one spanned by $\{1,g_1,\dots,g_m\}$.

3.6 Sharpness of the bound

Our sensitivity measures the worst case performance and it is instructive to compare it with the actual performance. Actual performance is impossible to compute for the general case because of non-linearities, but it can be done for the case of linear functional τ. We show that the worst-case can indeed be achieved but in reality this is unlikely to happen.

Any linear functional can be written in the following form:

$$\tau(f,\mu) = \tau(0,\mu) + \mathbb{E}_\mu[\nabla_\tau f] \tag{38}$$

where $\nabla_\tau \in \mathcal{F}$. Given this representation we have the following:

$$\tau(\hat{f},\mu) - \tau(f,\mu) = \mathbb{E}_\mu[\nabla_\tau(\hat{f}-f)] \tag{39}$$

Let $\{g_1,\dots,g_m,\dots\}$ be a particular basis in $\mathcal{S}$. We use the following representations for $\hat{f}$ and f:

$$\begin{cases}\hat{f} = \sum_{i=1}^{m}\beta_i g_i\\ f = \sum_{i=1}^{\infty}\beta_i g_i\end{cases} \tag{40}$$

This representation assumes that $\hat{f}$ is correct as far as first m basis directions are concerned. As a result we have the following expression for the difference:

$$\tau(\hat{f},\mu) - \tau(f,\mu) = \mathbb{E}_\mu\left[\nabla_\tau\left(\sum_{i=m+1}^{\infty}\beta_i g_i\right)\right] \tag{41}$$

Next we expand ∇_τ with respect to $\{g_1,\dots,\}$:

$$\nabla_\tau = \sum_{i=1}^{\infty}\gamma_i g_i + \varepsilon \tag{42}$$

where ε is orthogonal to $\mathcal{S}$.

Substituting (42) into (41) we get the following:

$$\tau(\hat{f},\mu) - \tau(f,\mu) = \mathbb{E}_\mu\left[\left(\sum_{i=1}^{\infty}\gamma_i g_i + \varepsilon\right)\left(\sum_{i=m+1}^{\infty}\beta_i g_i\right)\right] =$$
$$= \sum_{i=m+1}^{\infty}\gamma_i\beta_i \leq \sqrt{\sum_{j=m+1}^{\infty}\beta_j^2}\sqrt{\sum_{i=m+1}^{\infty}\gamma_i^2} = \|\hat{f} - f\|_2\Delta(\tau,\hat{f},\mathcal{G}) \quad (43)$$

where the inequality is just Cauchy-Schwarz inequality. As a result equality will be achieved if $\gamma_i = a\beta_i$ for $i = m+1,\dots$. Clearly, this is unlikely to happen in practice. This result depends on the basis $\{g_1,\dots,g_m,\dots\}$ and adjusting the basis one can come closer to equality.[6]

3.7 Discussion

Our measure of sensitivity offers a particular way of answering to the following question: "How harmful can be a potential misspecification?". We use the worst case measure and as a result our answer is conservative: if sensitivity is small then we are in the good situation, while if it is large it doesn't necessarily mean that everything is wrong, it's just and indicator that things **might** go wrong.

Current research practice suggests a different way of answering this question. Typical empirical paper presents several specifications, typically showing that results don't vary much across specifications. There are many problems with this approach: this is a multiple comparison problem and thus is statistically tricky; presented specifications might not give a full understanding of variability of effect since they are selected at author's discretion; finally, it might be simply impossible to present different specifications because of computational complexity.

Given all these issues it is quite surprising that our measure can be easily computed and estimated (we discuss estimation in the next section) even for very complex problems. There are conceptually two reasons why this happens: first, we focus on estimation of sensitivity, abstracting from estimation of distance between f and $\hat{f}$. The second reason is that we linearize the functional and for linear functionals we can apply Cauchy-Schwartz inequality.

Most of our results are concentrated on study of local perturbations around $\hat{f}$. We justify it by the following logic: of course our $\hat{f}$ can be very wrong, but it is the best guess that we have at hand (otherwise we would have used a different $\hat{f}$) and as a result it is reasonable to explore around $\hat{f}$. Essentially, we are mimicking the following discussion: everybody agrees that $\hat{f}$ is a reasonable guess for true f, but still worried about potential problems with misspecification. Note that in linear case sensitivity doesn't depend on $\hat{f}$, but, of course the answer $\tau(\hat{f},\mu)$ depends on $\hat{f}$.

Linearization allows us to compute sensitivity even in nonlinear cases, but of course there is a price - our measure becomes inherently local. However, in light of the discussion that $\hat{f}$ is the best guess that we have for f we don't think that this is a huge problem. At

[6] One obvious way of achieving equality is to use as a first basis vector $\sum_{i=1}^{m}\gamma_i g_i$ (normalized). In this case both sides of the inequality are equal to zero.

the end, we are only interested in local perturbations, because global perturbations can lead to completely unreasonable f. At the same time, it is clear that for very non-linear functionals our approximation can be quite poor even for reasonably local perturbations (e.g., perturbations of size of 1% of $\|\hat{f}\|_2$). This problem can be addressed given a particular application at hand.

Functional τ depends on two arguments: f and μ. Our discussion is completely centered on the first argument. We always have a good understanding of μ by means of the empirical measure and thus we don't think that misspecification of μ presents a particular problem.

There is a different related question regarding relationship between τ and μ: in complex models τ might not have a closed form. One way of understanding this black box is to understand how the functional depends on some characteristics of μ (moments). This is an interesting computational problem that deserves separate treatment. One approach to this problem is presented in (Gentzkow and Shapiro 2015).

4 Statistical estimation

4.1 Statistical setup

We work in the GMM framework. Moment conditions for τ and $\hat{f}$ are given below:

$$\begin{cases} \mathbb{E}_\mu[\psi_1(\theta, f(\eta, X), H, \gamma)] = 0 \\ \mathbb{E}_\mu[\tilde{\psi_2}(f(\eta, X), \theta, X, H, \gamma, \eta] = 0 \\ \mathbb{E}_\mu[\tilde{\psi_3}(X, H, \gamma)] = 0 \end{cases} \tag{44}$$

We assume that we know solutions to the system (44) without any error. Essentially, we assume that we know (η, γ) and as a result $\hat{f}(x) = f(\eta, x)$ and $\tau(\hat{f}, \mu)$. This is obviously a simplifying assumption, since in fact all these are random quantities. However, in this section we focus on estimation of sensitivity and all these parameters are of second importance. Moreover, in most practical cases these are low-dimensional parameters which can be estimated very precisely anyway.

Recall that in this setup crucial quantity for sensitivity is $\nabla_\tau(\hat{f})$ which in this case is equal to the following:

$$\nabla_\tau(\hat{f}) = \frac{\left(\frac{\partial\psi_1(\tau(\hat{f},\mu),\theta,H,\gamma)}{\partial\theta}|_{\theta=\hat{f}(X)}\right)}{\mathbb{E}_\mu\left[\frac{\partial\psi_1(\theta,\hat{f}(X),H,\gamma)}{\partial\theta}|_{\theta=\tau(\hat{f},\mu)}\right]} \tag{45}$$

Since we assume that γ, $\hat{f}$ and $\tau(\hat{f}, \mu)$ are known without error the numerator in (45) is observable. We further assume that the denominator is known, which given all the assumptions above is quite reasonable.

Another important quantity that we require in order to estimate sensitivity is the space $\mathcal{C}$. Recall that it is the space spanned by the components of the following m-dimensional vector:

$$\left(\frac{\partial\tilde{\psi_2}(\lambda, \tau(\hat{f}, \mu), X, H, \gamma, \eta)}{\partial\lambda}\right)_{\lambda=\hat{f}} \tag{46}$$

We assume that we observe a random sample $\{(X_i, H_i)\}_{i=1}^n$ from distribution μ. Let $\nabla_\tau(\hat{f})_i = Y_i$. We use the following shortcuts in what follows:

$$\begin{cases} Y_i = \dfrac{\left(\frac{\partial \psi_1(\tau(\hat{f},\mu),\theta,H_i,\gamma)}{\partial \theta}|_{\theta=\hat{f}(X_i)}\right)}{\mathbb{E}_\mu\left[\frac{\partial \psi_1(\theta,\hat{f}(X),H,\gamma)}{\partial \theta}|_{\theta=\tau(\hat{f},\mu)}\right]} \\ C_{ki} = \left(\frac{\partial \tilde{\psi}_{2k}(\lambda,\tau(\hat{f},\mu),X_i,H_i,\gamma,\eta)}{\partial \lambda}\right)_{\lambda=\hat{f}(X_i)} \end{cases} \tag{47}$$

Given our assumptions about observable quantities both Y_i and C_{ki} are observable and i.i.d. random variables. Let C be a random $n \times m$ matrix with elements C_{ki}.

4.2 Important identities

In our statistical estimation we make use of several simple and well-known identities which are true for any Hilbert space. In order to make discussion more straightforward we summarize them below.

Let Y be an element of $\mathcal{F}$ – space of all square-integrable measurable functions of (X, H). Let $\{c_1, \ldots, c_m\}$ be a set of m linearly independent functions from $\mathcal{S}$. Let $\mathcal{G}$ be an orthogonal complement (in $\mathcal{S}$) of the linear space spanned by vectors in $\{c_1, \ldots, c_m\}$ and let $\{g_1, \ldots, g_k, \ldots\}$ be its countable basis. Then we have the following two identities:

$$\begin{cases} Y = \sum_{i=1}^m \beta_i c_i + u^Y \\ Y = \sum_{i=1}^m \beta_i c_i + \sum_{i=1}^\infty \gamma_i g_i + \varepsilon^Y \end{cases} \tag{48}$$

where u is orthogonal to $\{c_1, \ldots, c_m\}$ and ε is orthogonal to $\{c_1, \ldots, c_m\} \cup \{g_1, \ldots, g_k, \ldots\}$. Essentially, u^Y is a populational residual after projecting Y on the linear space spanned by $\{c_1, \ldots, c_m\}$ and ε^Y is a populational residual after projecting Y on the whole space $\mathcal{S}$.

Identities in (48) lead to the following "Pythagorean"-style results:

$$\begin{cases} \|Y\|_2^2 = \|\sum_{i=1}^m \beta_i c_i\|^2 + \|u^Y\|_2^2 \\ \|Y\|_2^2 = \|\sum_{i=1}^m \beta_i c_i\|^2 + \sum_{i=1}^\infty \gamma_i^2 + \|\varepsilon^Y\|_2^2 \end{cases} \tag{49}$$

and from here we in turn have the next identity:

$$\sum_{i=1}^\infty \gamma_i^2 = \|u\|_2^2 \; - \|\varepsilon\|_2^2 \tag{50}$$

Observe that $\sum_{i=1}^\infty \gamma_i^2 = \|P_\mathcal{G}(Y)\|_2^2$ – square length of projection of Y on $\mathcal{G}$. Also, observe that if $\hat{f} = \sum_{i=1}^m \beta_i c_i$ and $f = \sum_{i=1}^m \beta_i c_i + \sum_{i=1}^\infty \gamma_i g_i$ then we have the following identity:

$$\|\hat{f} - f\|_2^2 = \|u^Y\|_2^2 - \|\varepsilon^Y\|_2^2 \tag{51}$$

Identity (51) is important in the following sense: it quantifies the distance between the conditional expectation f and its approximation $\hat{f}$.

For the further use define the following two functions: $m(x) = \sum_{i=1}^m \beta_i c_i(x) + \sum_{i=1}^\infty \gamma_i g_i(x)$ – conditional expectation function and $\sigma^2(x) = \mathbb{E}[\varepsilon^2|X = x]$ – conditional variance function.

4.3 Estimation

In this section we discuss a particular estimation procedure. We defer analysis of the resulting estimator to the next section.

Given the discussion above, and in particular given the identity (50) we understand that the we need to estimate the following expression:

$$\Delta(\mathcal{G}, \tau, \hat{f}) = \sqrt{\|u^Y\|_2^2 - \|\varepsilon^Y\|_2^2} \tag{52}$$

In order to do this we use a plug-in estimator: we estimate separately $\|u^Y\|_2^2$ and $\|\varepsilon^Y\|_2^2$.

Estimation of $\|u^Y\|_2^2$ is straightforward. Recall that u is just a populational residual that is left after projecting Y on a space spanned by $\{c_1, \ldots, c_m\}$. As a result we can estimate $\|u^Y\|_2^2$ in the following OLS manner:

$$\widehat{\|u^Y\|_2^2} = \frac{1}{n} Y^T \left(\mathcal{I}_n - C\left(C^T C\right)^{-1} C^T\right) Y \tag{53}$$

Estimation of $\|\varepsilon\|$ is trickier. Recall that ε is a population residual left after projecting Y on the whole $\mathcal{S}$. One would think that in order to estimate it's norm we need to construct the aforementioned projection. There is, however, a different way, suggested by (Liitiäinen, Corona, and Lendasse 2008).

In order to describe the proposed estimator we need to introduce some additional notation. Let $d_A(X_j, X_k) = \sqrt{(X_j - X_k)^T A (X_j - X_k)}$, where A is some positive definite matrix. For each i define $Y_i[1]$ using the following algorithm:[7]

$$Y_i[1] = Y_j : j = \arg\min_{j \neq i} \{d_A(X_i, X_j)\} \tag{54}$$

that is $Y_i[1]$ is the 1-nearest neighbor of Y_i. Let $i[1] = j = \arg\min_{j \neq i}\{d_A(X_i, X_j)\}$. Given $Y_i[k-1]$ define $Y_i[k]$ iteratively using the following algorithm:

$$Y_i[k] = Y_j : j = \arg\min_{j \notin \{i[1], \ldots, i[k-1]\}} \{d_A(X_i, X_j)\} \tag{55}$$

As a result, $Y_i[k]$ is just a k-nearest neighbor of Y_i.

Given these definitions we can construct the following estimator:

$$\widehat{\|\varepsilon^Y\|_2^2} = \frac{\sum_{i=1}^n (Y_i - Y_i[1])(Y_i - Y_i[2])}{n} \tag{56}$$

Estimator in (56) is quite intuitive: both $Y_i[1]$ and $Y_i[2]$ serve as approximations for projection of Y on $\mathcal{S}$. One would think that it is better to use $Y_i[1]$ everywhere instead of $Y_i[2]$ but this intuition is incorrect: since approximation errors in $Y_i[1]$ and $Y_i[2]$ are independent using both of them decreases bias of the estimator.

As a result the proposed estimator for sensitivity is the following one:

$$\widehat{\Delta}(\mathcal{G}, \tau, \hat{f}) = \sqrt{\max\left\{\widehat{\|u^Y\|_2^2} - \widehat{\|\varepsilon^Y\|_2^2}, 0\right\}} \tag{57}$$

[7] We assume that ties occur with probability zero and thus can be broken arbitrary.

Note that we need to take non-negative part of the difference, since in sample it can be negative.

Finally, we summarize our estimation procedure in the following algorithm:

```
  Data: $\{(Y_i, C_i, X_i)\}_{i=1}^n$
1 Construct $C = (C_1, \dots, C_n)^T$;
2 Construct $\hat{u}^Y = Y - C\left(C^T C\right)^{-1} C^T Y$;
3 Define $\widehat{\|u^Y\|_2}^2 := \frac{1}{n}\sum_{i=1}^n (u_i^Y)^2$;
4 Construct $Y[1]$ and $Y[2]$;
5 Define $\widehat{\|\varepsilon^Y\|_2}^2 := \frac{1}{n}\sum_{i=1}^n (Y_i - Y_i[1])(Y_i - Y_i[2])$;
6 Return $\widehat{\Delta}(\tau, \hat{f}, \mathcal{S} \setminus \mathcal{C}) := \sqrt{\max\left\{\widehat{\|u^Y\|_2^2} - \widehat{\|\varepsilon^Y\|_2^2}, 0\right\}}$ ;
  Result: $\widehat{\Delta}(\tau, \hat{f}, \mathcal{S} \setminus \mathcal{C})$
```

Algorithm 1: Computation of sensitivity

Proposed algorithm is quite fast. The most computationally intensive steps are (2) and (3). There exist (and implemented in R) algorithms that solve (3) in $O(n\ln(n))$ steps, while (2) requires $O(m^2 n)$ steps. Since m is fixed in our framework we have that the complexity of the algorithm is of order $(n\ln(n))$ which is quite fast.

4.4 Risk estimation

Here we focus on bias and variance bounds of our estimator. In order to simplify things we work with square sensitivity, that is we calculate bias and variance for the following quantity:

$$\hat{\Delta}^2 = \max\left\{\widehat{\|u^Y\|_2^2} - \widehat{\|\varepsilon^Y\|_2^2}, 0\right\} \tag{58}$$

Note that both bias and variance of the proposed estimator are bounded above by the bias and variance of the following estimator:

$$\widehat{\|u^Y\|_2^2} - \widehat{\|\varepsilon^Y\|_2^2} \tag{59}$$

We start with a bias; recall that $m(x)$ is a conditional expectation function of Y, while $\sigma^2(x)$ is a conditional variance. Our result is summarized in the following theorem:

Theorem 4.1. *Assume that $|m(x) - m(y)| \leq K d_A(x, y)$, assume that $\sigma^2(x)$ is a bounded function of x. Then the conditional bias of estimator (59) is of order $O\left(n^{\frac{-2}{d_1}}\right)$.*

Proof. Let $h(X_i) = \sum_{j=1}^{\infty} \gamma_j g_j(X_i)$ and $h = (h(X_1), \dots, h(X_n))^T$, let $C = UDV^T$ be a singular-value decomposition of C, let $M = \mathcal{I}_n - C(C^T C)^{-1} C = \mathcal{I}_n - UU^T$. Let $r[k](X_i) = m(X_i) - m(X_i[k])$, $r[k] = (r[k](X_1), \dots, r[k](X_n))^T$ and let $\varepsilon[k] = (\varepsilon_1[k], \dots, \varepsilon_n[k])^T$

By construction we have the following equality for the bias:

$$\begin{aligned}
b = Y^TMY - (Y - Y[1])^T(Y - Y[2]) - h^Th = \\
(h + \varepsilon^T)M(h + \varepsilon) - (r[1] + \varepsilon - \varepsilon[1])(r[2] + \varepsilon - \varepsilon[2]) - h^Th = \\
\varepsilon^TUU^T\varepsilon + 2\varepsilon^TUU^Th + h^TUU^Th- \\
r[1]^Tr[2] - r[1]^T(\varepsilon - \varepsilon[2]) - r[2]^T(\varepsilon - \varepsilon[1]) + \varepsilon[1]^T\varepsilon + \varepsilon[2]^T\varepsilon - \varepsilon[1]^T\varepsilon[2] \quad (60)
\end{aligned}$$

Let $D = \mathbb{E}[\varepsilon\varepsilon^T|X] = \text{diag}\{\sigma^2(X_1), \dots, \sigma^2(X_n)\}$; taking conditional expectation of b in (60) we come to the following equation for the conditional bias:

$$\text{bias}(X) = \frac{1}{n}\mathbb{E}[b|X] = \frac{1}{n}\left(h^TUU^Th + \text{trace}(DUU^T) - r[1]^Tr[2]\right) \quad (61)$$

Applying Von Neumann's trace inequality to the second term we get the following:

$$\begin{aligned}
|\text{bias}(X)| \leq \frac{1}{n}\left(h^TUU^Th + \text{trace}(DUU^T) + |r[1]^Tr[2]|\right) \leq \\
\frac{1}{n}h^TUU^Th + \frac{m}{n}\max_{i=1}^{n}\{\sigma^2(x_i)\} + \frac{1}{n}|r[1]^Tr[2]| \quad (62)
\end{aligned}$$

Now, we use the Lipchitz continuity assumption: $|r[k](X_i)| \leq Kd_A(X_i, X_i[k])$:

$$|\text{bias}(X)| \leq \frac{1}{n}h^TUU^Th + \frac{m}{n}\max_{i=1}^{n}\{\sigma^2(x_i)\} + \frac{K^2}{n}\sum_{i=1}^{n} d_A(X_i, X_i[2])^{2\gamma} \quad (63)$$

Results of (Liitiäinen, Corona, and Lendasse 2008) show that $\frac{K^2}{n}\sum_{i=1}^{n} d_A(X_i, X_i[2])^{2\gamma} \leq M_1n^{-\frac{2}{d_1}}$ for some constant M_1 that doesn't depend on d_1. Compactness assumption guaranties that $\frac{1}{n}h^TUU^Th \leq \frac{mM_2}{n}$ for some constant M_2 as a result we have the following bound:

$$|\text{bias}(X)| \leq M_2\frac{m}{n} + M_1n^{-\frac{2}{d_1}} \quad (64)$$

Since we assumed that $d_1 \geq 4$ we can imply that $|\text{bias}(X)|$ is or order $O\left(n^{\frac{-2}{d_1}}\right)$. □

We do not analyze variance, because it is well known that variance is going to be of order at most $O(n^{-1})$ and as a result it would be dominated by the bias term. As a result we can conclude that MSE = bias^2 + variance is of order $O\left(n^{-\frac{4}{d_1}}\right)$.

4.5 Discussion

In this section we presented a particular algorithm for estimation of our sensitivity measure. This algorithm is far from being perfect and of course it is not the only one that can be used to estimate the sensitivity.

One can proceed directly and try to estimate sensitivity directly as a length of projection. This is essentially equivalent to estimation of conditional mean and thus can be done by various non-parametric methods. We think that this approach is challenging because to implement it we need to work with a space $\mathcal{S} \setminus \mathcal{C}$ and this might be hard to do. At the same time we don't need the resulting estimator to be good in the typical sense (consistent, efficient, etc.) since we only use it to estimate the norm.

Our study of statistical properties of our procedure is far from being complete. Most importantly we don't build confidence intervals for our measure. We explore this question in our future research. One particularly attractive way to build these intervals is to use modern concentration of measure techniques instead of traditional asymptotic analysis. Check (Boucheron, Lugosi, and Massart 2013) for introduction in this topic.

Another important limitation of our analysis is that we assume that all low-dimensional nuisance parameters (η, γ, etc) are known. Clearly, full analysis requires us treat them as random.

5 Empirical example

In our empirical example we focus on a dataset of (LaLonde 1986). In his influential paper Lalonde showed directly the problems that econometricians face when trying to replicate results of the experimental study using observational data. Our goal here is to test our sensitivity measure and see whether it will give reasonable answers in a situation which we understand quite clearly.

5.1 Description of data

Our dataset consists of three distinct groups: one treated group and two control groups. Treated and control group correspond to the experimental data from a job training program for men in in 1970s. In the experiment we have 185 treated individuals and 205 control individuals. Let D denote the treatment status. For each individual we observe 8 attributes (X-s): age, education, hispanic, marriage, black, degree indicator and earnings in 1974 and 75. We also observe an outcome variable Y – earnings in 1978.

The second control group is just a group randomly drawn from Current Population Survey (CPS). There are $15,592$ units in the second control group. For each one we observe the same 8 attributes and the outcome variable, as well as two additional covariates: unemployment indicators in 1974 and 75.

For a detailed description of the experiment and construction of the control group check (LaLonde 1986). Original goal in the experiment was to asses the effect of the program on the outcome variable. As such this goal fits in our first example – assessing treatment effect under unconfoundedness.

5.2 Analysis

Our goal here is to construct our sensitivity measure for two datasets: "experimental" dataset, using only experimental data and "observations" dataset, using experimental treatment group and general control group. First, we estimate treatment effect for both datasets, sing the following moment conditions:

$$\begin{cases} \mathbb{E}_\mu[(Y - \theta D - \sum_{i=1}^m \beta_i X_i)(D - \mathbb{E}[D])] = 0 \\ \mathbb{E}_\mu\left[(Y - \alpha - \theta D - \sum_{i=1}^m \beta_i X_i) X_i\right] = 0 \text{ for } i = 1, \dots, m \\ \mathbb{E}_\mu\left[(Y - \alpha - \theta D - \sum_{i=1}^m \beta_i X_i)\right] = 0 \end{cases} \tag{65}$$

For the experimental dataset we have $m = 8$, while for the observational dataset we have $m = 10$. Since we discussed this set of moment conditions in one of the examples we know that $\mathcal{C} = \text{span}\{X_1, \dots, X_m\}$.

To estimate sensitivity we proceed with the algorithm described in (4.3). First we run a regression of D on $(X_1, \dots, X_m)$ to get $\hat{u}^D$ and estimate $\|u^D\|_2^2$. Then we construct nearest neighbors $D[1]$ and $D[2]$. We standardize regressors (divide by the norm) which is equivalent to using diagonal weighting matrix A in definition of d_A. Using $D[1]$ and $D[2]$ we estimate $\|\varepsilon^D\|_2^2$.

In this setup function f has a particular interpretation: it is a conditional expectation of $Y(0)$ given $X_1, \dots, X_m$. We can view $\hat{f} = \hat{\alpha} + \sum_{i=1}^m \hat{\beta}_i X_i$ as its approximation and as a result use equality (51) to estimate not just sensitivity but also the distance $\|f - \hat{f}\|$. This distance has exactly same structure as sensitivity (although conceptually it is a very different object) and we can use straightforward adaptation of our algorithm to estimate it.

Namely, first we estimate $\hat{f}$ by OLS running regression of $Y(0)$ (in control groups) on $(X_1, \dots, X_m)$, construct $\hat{u}^Y$ and use it to estimate $\|u^Y\|_2^2$. Then we construct $Y[1]$ and $Y[2]$ and use them to estimate $\|\varepsilon^Y\|_2^2$.

5.3 Results

We summarize our results in the Table 1. We see that sensitivity is equal to 5.57 in the observational dataset and estimated at zero in the experimental one. Zero in the experimental dataset is the result of the small sample size – non-parametric noise turned out to be larger than the parametric one.

In the table we also report estimated distance to the truth and we can see that it is quite significant in both cases: it is estimated at 723 in the observational data and at 439 in the experimental data. This means that both in experimental and observational datasets simple linear model is far from the correct one. However, it has drastically different implications: this misspecification doesn't have any impact in the experimental framework, but in the observational framework. If we construct the worst case variation ($e \times \Delta$) it will be equal to 0 in the experimental dataset and to 4013 in the observational one.

6 Future research

The main goal of this paper is study of misspecification. Be believe that this study is far from being complete and discuss below several potential areas that we would like to explore in future research.

In this paper we focused on a particular kind of probability models – those defined by moment conditions. Specifically we were working in exactly identified framework. It is interesting to explore how important is misspecification in over-identified frameworks. In

Table 1: Sensitivity computation for Lalonde dataset

	Non-experimental	Experimental
$\widehat{\Vert u^Y \Vert}_2$	6978	5436
$\widehat{\Vert \varepsilon^Y \Vert}_2$	6940	5418
$\widehat{\Vert u^D \Vert}_2$	0.10	0.49
$\widehat{\Vert \varepsilon^D \Vert}_2$	0.08	0.49
$\hat{\Delta} = \frac{\sqrt{\widehat{\vert u^D \Vert}_2^2 - \widehat{\Vert \varepsilon^D \Vert}_2^2}}{\hat{\text{var}}(D)}$	**5.57**	**0**
$\hat{e} = \sqrt{\widehat{\Vert u^Y \Vert}_2^2 - \widehat{\Vert \varepsilon^Y \Vert}_2^2}$	723	439
$\hat{e} \times \hat{\Delta}$	**4013**	**0**
$\hat{\tau}$	1133	1784

these models the method used to reduce over-identified system to just-identified clearly affects the definition of parameter of interest. One important question is to disentangle two effects that misspecification has in these frameworks: change of question and change of f. Potentially different methods (GEL, GMM) can be compared on these grounds.

This paper is mostly focused on populational framework, abstracting away from many finite sample problems. It is important to provide further analysis of sample behavior of our measure, taking into account those uncertainties that we ignored in our analysis. We also need additional analysis in order to build confidence intervals for our measure.

Very different area that we would like to explore is connection between our analysis and well-established analysis of semiparametric efficiency bounds (see Bickel et al. 1993). We believe that potential connection between our measure and these bounds is of theoretical interest.

On more abstract level we think that it is important to understand the role of misspecification in frameworks that don't fit into hours: namely those in which definition of parameter changes with different models. We think that understanding relationship between changes in models and changes in question is crucial for empirical work.

References

[1] Alberto Abadie and Guido W Imbens. "Large sample properties of matching estimators for average treatment effects". In: *Econometrica* 74.1 (2006), pp. 235–267.

[2] Stanislav Anatolyev and Nikolay Gospodinov. *Methods for estimation and inference in modern econometrics*. CRC Press, 2011.

[3] Susan Athey and Guido Imbens. "A Measure of Robustness to Misspecification". In: *American Economic Review* 105.5 (2015), pp. 476–80.

[4] Peter J Bickel et al. *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press Baltimore, 1993.

[5] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. OUP Oxford, 2013.

[6] Rajeev H Dehejia and Sadek Wahba. “Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs”. In: *Journal of the American statistical Association* 94.448 (1999), pp. 1053–1062.

[7] Luisa Turrin Fernholz. *Von Mises calculus for statistical functionals*. Vol. 19. Springer Science & Business Media, 2012.

[8] Matthew Gentzkow and Jesse M Shapiro. *Measuring the sensitivity of parameter estimates to sample statistics*. Tech. rep. National Bureau of Economic Research, 2015.

[9] Lars Peter Hansen. “Large sample properties of generalized method of moments estimators”. In: *Econometrica: Journal of the Econometric Society* (1982), pp. 1029–1054.

[10] Guido W Imbens and Donald B Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

[11] Robert J LaLonde. “Evaluating the econometric evaluations of training programs with experimental data”. In: *The American economic review* (1986), pp. 604–620.

[12] Edward E Leamer. “Let’s take the con out of econometrics”. In: *The American Economic Review* (1983), pp. 31–43.

[13] Elia Liitiäinen, Francesco Corona, and Amaury Lendasse. “On nonparametric residual variance estimation”. In: *Neural Processing Letters* 28.3 (2008), pp. 155–167.