Submitted to Econometrica

Nonparametric Instrumental Variables Estimation Under Misspecification

Ben Deaner

November 16, 2019

Submitted to Econometrica

NONPARAMETRIC INSTRUMENTAL VARIABLES ESTIMATION UNDER MISSPECIFICATION

BEN DEANER

We show that nonparametric instrumental variables (NPIV) estimators are highly sensitive to misspecification: an arbitrarily small deviation from instrumental validity can lead to large asymptotic bias for a broad class of estimators. One can mitigate the problem by placing strong restrictions on the structural function in estimation. However, if the true function does not obey the restrictions then imposing them imparts bias. Therefore, there is a trade-off between the sensitivity to invalid instruments and bias from imposing excessive restrictions. In light of this trade-off we propose a partial identification approach to estimation in NPIV models. We provide a point estimator that minimizes the worst-case asymptotic bias and error-bounds that explicitly account for some degree of misspecification. We apply our methods to the empirical setting of Blundell *et al.* (2007) and Horowitz (2011) to estimate shape-invariant Engel curves.

INTRODUCTION

In his 1799 work *The Vocation Of Man*, the German idealist philosopher Johann Gottlieb Fichte wrote that "you could not remove a single grain of sand from its place without thereby [...] changing something throughout

[1]PhD candidate, Massachusetts Institute of Technology. Email at bdeaner@mit.edu. The author thanks Whitney Newey, Anna Mikusheva and Jerry Hausman who provided indispensable advice and countless helpful comments throughout the creation of this project. We are also grateful to Isaiah Andrews and Florian Gunsilius for useful comments and advice and to Sophie Sun, Claire Lazar and Katharine Lauderdale for their feedback on drafts of this paper. We also thank those who attended the sessions in which this paper was presented at the 2019 North American and European Summer Meetings of the Econometric Society and those who attended lunch seminars at MIT for numerous useful comments. The author gratefully acknowledges the financial support of the MIT economics department's graduate program.

all parts of the immeasurable whole".[1] Fichte, to his great misfortune, died almost a century before the invention of instrumental variables (IV) regression, but his quote is of considerable relevance to IV estimation. Suppose Fichte is correct: subtle and serpentine causal channels connect all things. Then instrumental exogeneity is at best a close approximation of the truth. If we agree with Fichte we must hope that a small deviation from instrumental validity imparts only a small asymptotic bias in our IV estimates.

Similar arguments motivate analyses of parametric IV estimation when instruments may be mildly invalid. Conley *et al.* (2008) propose (among other things) a partial identification approach to estimation in linear IV models that is valid even if instrumental validity fails. Andrews *et al.* (2017) provide methods to analyze the sensitivity of GMM estimates to misspecification of the moment conditions. Recent work by Armstrong & Kolesar (2018) explores optimal estimation in the GMM framework under misspecified moment conditions.

Nonparametric instrumental variables (NPIV) estimation (Newey & Powell (2003), Ai & Chen (2003) and others) is a flexible alternative to linear IV. NPIV models relax the assumption of a linear causal relationship between regressors and outcomes. In NPIV estimation the 'structural function', which describes this causal relationship, is treated nonparametrically. We show that NPIV estimators of the structural function are more sensitive to invalid instruments than parametric IV estimators and standard nonparametric regression estimators. For a broad class of NPIV estimators, an arbitrarily small deviation from instrumental validity can impart a large asymptotic bias. In some cases arbitrarily large. This non-robustness is an inherent feature of NPIV estimation. Any NPIV estimator that is robust requires strong restrictions on the structural function for consistency.

The non-robustness of NPIV estimators is closely linked with the 'ill-

[1]Quote is from the English translation Fichte & Smith (1848).

posedness' of NPIV estimation. NPIV estimation is ill-posed in that a tiny change to the 'reduced-form' components of the NPIV estimating equation can induce a large jump in the solution (see, e.g., Darolles *et al.* (2011)). The reduced-form components are estimated empirically, and therefore subject to error. To limit the sensitivity to the estimation error one must 'regularize' the estimating equation. However, regularization generally imparts bias. As the sample size grows, the reduced-form is estimated with greater precision, and so the degree of regularization is reduced.

The presence of invalid instruments is akin to error in the reduced-form. Misspecification perturbs the reduced-form away from the shape it would take under instrumental validity. As the degree of regularization is reduced, the influence of error due to misspecification grows. Unlike the error due to sampling noise, the error from misspecification does not decrease with the sample size. If the researcher has access to a large sample and in response regularizes only weakly, then even a small deviation from instrumental validity can impart a large bias. To the best of our knowledge this observation is neither made nor addressed in any of the existing literature.

In addition to our non-robustness results, we specify two important cases in which misspecification-robust estimation is possible. Firstly, suppose that the researcher is interested in estimation of a continuous functional of the structural function (Ai & Chen (2003), Severini & Tripathi (2012), Ichimura & Newey (2017)) rather than the structural function itself. This estimation problem may not be ill-posed and our sensitivity results may not apply. We provide necessary and sufficient conditions under which a continuous linear functional can be estimated robustly.

Secondly, suppose it is known that the structural function obeys some shape restrictions, say a smoothness condition. If the restrictions are sufficiently strong then an estimator that imposes them on the structural function may be robust to misspecification, without sacrificing consistency. Two

NPIV methods which impose strong smoothness conditions are the sieve-type procedures of Newey & Powell (2003) and Blundell *et al.* (2007).[2] However, smoothness conditions (and other shape restrictions) are absent from a number of prominent NPIV methods.[3]

Unfortunately, even with the imposition of nonparametric smoothness, we show that NPIV estimators are more sensitive to misspecification than parametric IV estimators or standard nonparametric regression estimators. Moreover, if the structural function violates some smoothness assumptions, then a procedure that imposes those assumptions cannot be consistent, even if instruments are valid.

In sum, NPIV estimation under misspecification involves a trade-off. Imposing strong restrictions on the structural function reduces the sensitivity to a failure of instrumental validity but risks additional bias. Ideally, a researcher would make the trade-off optimally and evaluate a point-estimator with minimal worst-case asymptotic bias. Moreover, the researcher would present error bands that directly account for some degree of misspecification.

To this end we propose a new approach to estimation and empirical sensitivity analysis in NPIV. Rather than assume correct specification, we propose a method based on partial identification (Manski (1989) , Horowitz & Manski (1995)).[4]

[2] Chetverikov & Wilhelm (2017) impose monotonicity on the structural function in order to tackle the problem of ill-posedness. However, their analysis also assumes monotonicity in the reduced-form relationship between the instruments and endogenous regressors. This fall outside the scope of our analysis.

[3] The NPIV methods described in Chen & Christensen (2018), Darolles *et al.* (2011), Hall & Horowitz (2005) and Horowitz (2011) to name a few. Many analyses of NPIV estimation fall into a third category in that they are general enough to incorporate both estimators that do and do not impose smoothness, for example Chen & Pouzo (2012).

[4] Santos (2012) and Freyberger & Horowitz (2015) also propose partial identification approaches to NPIV. However, their analyses assume instrumental validity.

We replace the assumption of instrumental validity with a weaker assumption that the deviation from instrumental validity is bounded in the supremum norm. We also place a priori restrictions on the structural function (e.g., a bound on its second derivatives). This yields a linear conditional moment inequality model (Andrews & Shi (2013)) in which the parameter of interest is a function.

We provide a procedure to estimate the envelopes of the identified set and to evaluate a point estimator with minimal worst-case asymptotic bias. Our method is simple and computationally light. The first stage amounts to standard non-parametric regression and the second stage consists of linear programming. We derive uniform rates of convergence in probability under both high-level and primitive conditions.

The estimation problem in our partially identified model is not ill-posed and we show that our estimators can achieve the same uniform rate of convergence as in standard series regression.

We apply our methods to the empirical setting shared by Blundell *et al.* (2007) and Horowitz (2011) and replicate the results of the latter. Blundell *et al.* (2007) and Horowitz (2011) use NPIV methods to estimate shape-invariant Engel curves using data from The British Family Expenditure Survey. We use our methodology to assess which features of the structural Engel curve for food can be inferred robustly.

The rest of the paper is structured as follows. In Section 1 we provide an overview of NPIV models and estimators in the context of full instrumental validity. In Section 2 we consider the case in which instrumental validity fails and analyze the asymptotic implications for NPIV estimators. In Section 3 we present our partial identification approach to NPIV estimation. We provide conditions for the uniform consistency and convergence rate of the set estimator. In Section 4 we apply our methods to the empirical setting of Horowitz (2011). Supplementary material can be found in Appendix A and proofs in Appendix B.

1. NPIV ESTIMATION UNDER CORRECT SPECIFICATION

Newey & Powell (2003) present the first detailed, published analysis of nonparametric instrumental variables (NPIV) methods and their asymptotic properties. [5] They provide high-level conditions for identification of the structural function in an NPIV model and describe a nonparametric two-stage procedure for estimation of the structural function. In the years following their foundational work many competing NPIV estimators have been introduced and their asymptotic properties analyzed (e.g., Ai & Chen (2003), Chen & Pouzo (2012), Darolles *et al.* (2011), Hall & Horowitz (2005), Horowitz (2011)).

NPIV analyses identify and estimate the 'structural function', denoted by h_0, from a conditional moment restriction of the following form:

$$E[Y - h_0(X)|Z] = 0 \tag{1.1}$$

Y is a scalar dependent variable with finite first absolute moment, X is a vector of possibly endogenous regressors, Z is a vector of instruments. It should be understood that the equality holds 'almost surely' (i.e., with probability 1). We assume throughout that draws of the triple (Y, X, Z) are independent and identically distributed. Throughout we denote the support of X and of Z by $\mathcal{X}$ and $\mathcal{Z}$ respectively.

The structural function is treated nonparametrically. It is assumed to lie in an infinite-dimensional set of functions $\mathcal{H}$ which is in turn a subset of a Banach space $\mathcal{B}_X$ with norm $||\cdot||_{\mathcal{B}_X}$. We assume that for any $h \in \mathcal{B}_X$ the first absolute moment of $h(X)$ is finite.

It is useful to rewrite the moment condition in terms of functions and linear operators. Let g_0 denote the reduced-form function, that is $g_0(Z) =$

[5] A brief account of NPIV and its asymptotic properties appears as an example application in Newey (1991).

$E[Y|Z]$. Again, the equality should be understood to hold almost surely. We assume that g_0 lies in a Banach space $\mathcal{B}_Z$ with norm $||\cdot||_{\mathcal{B}_Z}$.

Let A be the bounded linear operator that maps from a function $h \in \mathcal{B}_X$ to the element of $\mathcal{B}_Z$ that is almost surely equal to $E[h(X)|Z]$. The conditional moment restriction (1.1) can then be expressed as a linear operator equation $A[h_0] = g_0$

1.1. *Standard Assumptions on the Joint Distribution of X and Z*

Below we state two properties of the joint distribution of the instruments and regressors, both of which are imposed throughout the NPIV literature.

The first of these assumptions, 'completeness', is key to identification of the structural function from the NPIV moment condition. Completeness is a topic of intense discussion in the NPIV literature. For some recent work see Andrews (2017), Canay *et al.* (2013), Chen *et al.* (2014), D'Haultfoeuille (2011), Freyberger (2017), Hu & Shiu (2018).

The second assumption is known as 'compactness' of the operator A defined above. Many useful results from functional analysis apply to operators that are compact, and so the compactness assumption simplifies analysis of the NPIV estimation problem. For some discussion of compactness in NPIV (including primitive conditions that imply this property) see, e.g., Florens (2011) and Horowitz (2011).

Assumption 1.1 ($\mathcal{H}$-Completeness) For any $h \in \mathcal{H}$, $E[h(X)|Z] = 0 \iff h(X) = 0$

Assumption 1.2 (Compactness) The linear operator A is a compact operator from $\mathcal{B}_X$ into $\mathcal{B}_Z$.

1.2. *Ill-posedness and Regularization*

Under Assumption 1.1 the structural function is the unique solution to the estimating equation (1.1) in the parameter space $\mathcal{H}$. For now let us assume that $\mathcal{H} = \mathcal{B}_X$, then the operator A is invertible on its range.

Denoting the inverse by A^{-1} we have $h_0 = A^{-1}[g_0]$. The objects on the right-hand side are known functionals of the joint distribution of observables. Thus this expression shows the structural function is identified.

If A is an infinite-dimensional and compact operator then the problem $A[h] = g_0$ is 'ill-posed'. In particular, the operator A does not have a closed range and the inverse A^{-1} is discontinuous everywhere on its domain. Let Assumption 1.2 hold, then:

$$\sup_{g \in R(A):\, ||g||_{\mathcal{B}_Z} \leq 1} ||A^{-1}[g]||_{\mathcal{B}_X} = \infty$$

Where $R(A) \subset \mathcal{B}_Z$ is the range of A.

Suppose A is known but g_0 is replaced with a consistent empirical estimate $\hat{g}_n$. Because A^{-1} is discontinuous, an estimate $\hat{h}_n = A^{-1}[\hat{g}_n]$ need not converge in probability to h_0. For this reason one employs a 'regularization scheme'. The researcher specifies a sequence of continuous functions $\{Q_k\}_{k=1}^{\infty}$ that converges pointwise to the discontinuous operator A^{-1}.[6] For a discussion of regularization in the context of NPIV see for example Darolles *et al.* (2011).

In economic applications the linear operator A is not a priori known and must be estimated from the data, correspondingly a regularized inverse must also be estimated empirically. For each k let $\hat{Q}_{n,k}$ estimate Q_k. Let $\hat{g}_n$ be an estimator of g_0. A typical NPIV estimator $\hat{h}_n$ takes the following form:

$$\hat{h}_n = \hat{Q}_{n,K_n}[\hat{g}_n] \tag{1.2}$$

[6]That is, for any fixed $g \in R(A)$, $||Q_k[g] - A^{-1}[g]||_{\mathcal{B}_X} \to 0$.

Where K_n is a sequence of natural numbers that grows to infinity with the sample size.

The choice of K_n controls the degree of regularization. If K_n is large then the estimator is highly sensitive to error in $\hat{g}_n$. Therefore K_n must grow sufficiently slowly so that the increased sensitivity is balanced by the increased precision in the estimate $\hat{g}_n$. However, K_n must grow to infinity because the regularization itself may impart bias.

For high level conditions for consistency of an estimator of the form above see Theorem A.1 in Appendix A.

2. NPIV ESTIMATION UNDER MISSPECIFICATION

We now allow for the possibility that the moment condition (1.1) is misspecified, i.e., that $E[Y - h_0(X)|Z] \neq 0$. Define a function $u_0 \in \mathcal{B}_Z$ by:

$$u_0(Z) = E[Y - h_0(X)|Z]$$

In terms of the notation developed in the previous section we have that:

$$g_0 = A[h_0] + u_0 \tag{2.1}$$

u_0 measures the deviation of the NPIV conditional moment from zero. In-keeping with our interpretation of misspecification as endogeneity of the instruments, we sometimes refer to u_0 as the 'instrumental endogeneity'. It is natural to measure the degree of misspecification by the norm of the function u_0. If the norm of u_0 is small, then the NPIV conditional moment is close to zero with respect to the norm. In the previous section we introduced a norm $||\cdot||_{\mathcal{B}_Z}$ on the function space that contains the reduced-form function, we use this same norm to define the degree of misspecification.

2.1. *Asymptotic Bias With Endogenous Instruments*

To measure the sensitivity of an NPIV estimator to instrumental endogeneity, we consider the largest possible asymptotic bias when the model is perturbed away from correct specification. We consider perturbations so that the degree of misspecification is bounded and parameters of the model that are not directly related to misspecification are fixed. Keeping these parameters fixed prevents us from perturbing the model in such a way that the instruments become weak or the moments of the reduced-form error become large. We call our measure of sensitivity the 'worst-case asymptotic bias'. It is a special case of the 'maximum bias' discussed in Huber (2011), albeit extended to estimators whose values are functions rather than scalars.

We fix the true structural function h_0 and the joint probability distribution of the regressors X, instruments Z and the reduced-form error η defined by $\eta = Y - E[Y|Z]$. We denote this joint probability by '$\mu_{XZ\eta}$'. Note that $\mu_{XZ\eta}$, h_0 and u_0 together determine the joint distribution of the observables Y, X and Z and therefore the distribution of any NPIV estimator at any sample size (recall Y, X and Z are iid).

Let $\hat{h}_n$ be an estimator of h_0, the estimation error of $\hat{h}_n$ is $||\hat{h}_n - h_0||_{\mathcal{B}_X}$. If $\hat{h}_n$ converges in probability then we define the asymptotic bias to be the probability limit of the estimation error.

With the structural function h_0 fixed, the asymptotic bias is fully determined by $\mu_{XZ\eta}$ and the degree of misspecification u_0. So we fix $\mu_{XZ\eta}$ and define the 'worst-case asymptotic bias' of $\hat{h}_n$ to be the largest asymptotic bias given $u_0 \in R(A)$ and $||u_0||_{\mathcal{B}_X} \leq b$. As a function of b this is:

$$bias_{\hat{h}_n}(b) = \sup_{u_0 \in R(A):\, ||u_0||_{\mathcal{B}_Z} \leq b} \text{plim}_{n\to\infty} ||\hat{h}_n - h_0||_{\mathcal{B}_X}$$

The worst-case asymptotic bias of an estimator captures the sensitivity of the estimator to misspecification in the form of instrumental endogeneity. If $bias_{\hat{h}_n}(b)$ is small when the argument b is small, then a tight bound on the magnitude of the misspecification implies that any asymptotic bias in the estimator $\hat{h}_n$ must be small. Thus the behavior of $bias_{\hat{h}_n}(b)$ around $b = 0$ captures the robustness/non-robustness of the estimator to a small amount of misspecification. If $bias_{\hat{h}_n}(b)$ converges to zero as the argument b goes to zero, we describe the estimator $\hat{h}_n$ as 'robust' to misspecification in the form of invalid instruments.

Theorem 2.1 applies to any estimator that is consistent whenever instruments are valid and the structural function lies in the interior of the parameter space $\mathcal{H}$. It states that if the structural function is indeed in the interior of the parameter space $\mathcal{H}$, then the worst-case asymptotic bias must be greater than some strictly positive constant, no matter how small the degree of misspecification. In fact, if the parameter space is the whole function space $\mathcal{B}_X$, then the worst-case asymptotic bias must be infinite.

THEOREM 2.1 *Fix $\mu_{XZ\eta}$ so that Assumptions 1.1 and 1.2 hold. Let $\hat{h}_n$ be an NPIV estimator that has a probability limit and is consistent under instrumental validity whenever $h_0 \in int(\mathcal{H})$. That is, for any $h_0 \in int(\mathcal{H})$, if $u_0 = 0$ then $\underset{n\to\infty}{plim}||\hat{h}_n - h_0||_{\mathcal{B}_X} = 0$.*

Then for any $h_0 \in int(\mathcal{H})$ the estimator $\hat{h}_n$ is not robust. More precisely, if h_0 is at the center of an open ball $\mathcal{V} \subseteq \mathcal{H}$ with radius r, then for any $b > 0$, $bias_{\hat{h}_n}(b) \geq r$. Furthermore, if $\mathcal{H} = \mathcal{B}_X$ then for any $b > 0$, $bias_{\hat{h}_n}(b) = \infty$.

To get an idea of the finite-sample effect of misspecification, suppose for simplicity that $\mathcal{H} = \mathcal{B}_X$ and consider an estimator of the form (1.2) discussed in Section 1. Below we give a lower bound on the error of such an estimator. The bound (which follows by the reverse triangle inequality) contains three parts: a) the error due to misspecification, b) the error due to regularization (i.e, due to replacing A^{-1} with Q_{K_n}), and c) The estimation error in $\hat{g}_n$ and $\hat{Q}_{n,K_n}$:

$$
\begin{aligned}
||\hat{h}_{n,K_n} - h_0||_{\mathcal{B}_X} \geq & ||Q_{K_n}[u_0]||_{\mathcal{B}_X} \\
& - ||(A^{-1} - Q_{K_n})A[h_0]||_{\mathcal{B}_X} \\
& - ||\hat{Q}_{n,K_n}[\hat{g}_n] - Q_{K_n}[g_0]||_{\mathcal{B}_X}
\end{aligned}
$$

For the worst-case asymptotic bias we take the supremum over all $u_0 \in R(A)$ such that $||u_0||_{\mathcal{B}_Z} \leq b$. The first term then becomes $||Q_{K_n}||_{op} b$.

Where '$||\cdot||_{op}$' denotes the operator norm.[7]

For a given k, the operator norm of Q_k is finite. However, as we discuss in Section 1, consistency with valid instruments generally requires Q_{K_n} converge pointwise to A^{-1}. Under Assumptions 1.1 and 1.2 this necessarily implies that $||Q_{K_n}||_{op} \to \infty$.

And so if $b > 0$ the first term in the worst-case asymptotic bias grows to infinity. Consistency also typically requires that the remaining two terms in the expansion above go to zero. So we see the worst-case asymptotic bias is divergent.

2.2. *The Role of Smoothness*

The parameter space $\mathcal{H}$ plays a key role in Theorem 2.1. The theorem only applies when the structural function lies in the interior of the parameter space $\mathcal{H}$. If the parameter space has an empty interior then this condition is trivially false. We consider two particular classes of sets in Assumption 2.1. If $\mathcal{H}$ belongs to either class it must have an empty interior. If the structural function is known to belong to a set in either class then one can generally construct estimators that are consistent under instrumental validity and are robust to the failure of instrumental validity.

The two classes of sets that we consider are given below.

Assumption 2.1 i. $\mathcal{H}$ is a compact infinite-dimensional subset of $\mathcal{B}_X$. ii. $\mathcal{H}$ is a finite-dimensional linear subspace of $\mathcal{B}_X$.

Assumptions of the infinite-dimensional type are employed extensively in the literature (e.g., in Newey & Powell (2003), Ai & Chen (2003), Blundell *et al.* (2007), Freyberger (2017), Santos (2012)).

[7]For an operator $\mathbb{T} : \mathcal{B}_Z \to \mathcal{B}_X$, $||\mathbb{T}||_{op} = \sup_{g \in \mathcal{B}_Z :\, ||g||_{\mathcal{B}_Z}=1} ||\mathbb{T}[g]||_{\mathcal{B}_X}$.

The compact function spaces employed in the NPIV literature generally correspond to sets of smooth functions, e.g., functions in a given Hölder ball. The finite-dimensional linear case corresponds to an IV model with a parametric and linear-in-parameters second stage.

In either case we can replace the NPIV estimating equation with an alternative estimating equation $A[h_0] = P_Z[g_0]$.

P_Z denotes a projection from $\mathcal{B}_Z$ to $A[\mathcal{H}]$ (the image of $\mathcal{H}$ under A). A projection onto $A[\mathcal{H}]$ is a function that maps elements of $\mathcal{B}_Z$ to elements of $A[\mathcal{H}]$ and when applied to elements of $A[\mathcal{H}]$ leaves them unchanged.

If the instruments are valid and the structural function is in $\mathcal{H}$, then the reduced-form g_0 lies in $A[\mathcal{H}]$. Consequently, $P_Z[g_0] = g_0$. So if the structural function lies in $\mathcal{H}$ and instruments are valid then the alternative estimating equation holds.

If Assumption 1.1 holds then under correct specification there is a unique element h_0 of $\mathcal{H}$ that satisfies this equation. Let $A_{\mathcal{H}}$ denote the restriction of the operator A to $A[\mathcal{H}]$. Then the unique solution in $\mathcal{H}$ to the alternative estimating equation can be written as $A_{\mathcal{H}}^{-1} P_Z[g_0]$.

If $\mathcal{H}$ satisfies either Assumption 2.1.i or 2.1.ii, then any estimator that converges to a unique solution in $\mathcal{H}$ to the alternative estimating equation is robust to instrumental endogeneity.

THEOREM 2.2 *Fix $\mu_{XZ\eta}$ so that Assumption 1.1 holds and suppose that either of Assumptions 2.1.i or 2.1.ii holds. Suppose P_Z is a uniformly continuous projection onto $A[\mathcal{H}]$ and the estimator $\hat{h}_n$ satisfies:*

$$||\hat{h}_n - A_{\mathcal{H}}^{-1} P_Z[g_0]||_{\mathcal{B}_X} \to^p 0$$

Then, if $h_0 \in \mathcal{H}$ the estimator is robust. That is $\lim_{b\to 0} bias_{\hat{h}_n}(b) = 0$.

Theorem 2.2 shows that the worst-case asymptotic bias goes to zero with the tightness of the bound on u_0, but it does not give a rate. We now show that Assumptions 2.1.i and 2.1.ii have rather different implications. In the finite-dimensional linear case the asymptotic bias goes to zero at the same rate as the bound b. However, under weak additional conditions, in the compact and infinite-dimensional case the rate is strictly slower. In short, even if nonparametric smoothness is imposed in estimation, NPIV estimators are still less robust than parametric linear IV estimators or standard nonparametric regression estimators.

ASSUMPTION 2.2 $\mathcal{H}$ is convex and symmetric.[8] Furthermore there exists $\alpha \in (0,1)$ so that $\frac{1}{\alpha}h_0 \in \mathcal{H}$.

The conditions Assumption 2.2 places on $\mathcal{H}$ hold for the compact infinite-dimensional spaces typically used in the NPIV literature including all of those considered by Freyberger & Masten (2019). The assumption that $\frac{1}{\alpha}h_0 \in \mathcal{H}$ for some $\alpha \in (0,1)$ is, loosely speaking, a requirement that the structural function does not lie at the edge of the parameter space.

THEOREM 2.3 *Fix $\mu_{XZ\eta}$ so that Assumptions 1.1 and 1.2 hold.*

a. Suppose Assumption 2.1.ii holds and suppose P_Z is a bounded linear projection operator onto $A[\mathcal{H}]$ (such a projection has to exist). Suppose that for any $g_0 \in A[\mathcal{H}]$ the estimator $\hat{h}_n$ satisfies:

$$||\hat{h}_n - A_{\mathcal{H}}^{-1}P_Z[g_0]||_{\mathcal{B}_X} \to^p 0$$

Then for any $b > 0$ there exists a constant C (not dependent on h_0) so that for any $h_0 \in \mathcal{H}$, $bias_{\hat{h}_n}(b)/b \leq C$.

b. Suppose Assumptions 2.1.i and 2.2 hold. Suppose the estimator $\hat{h}_n$ is consistent for h_0 whenever $h_0 \in \mathcal{H}$ and $u_0 = 0$. Then $\lim_{b\to 0} bias_{\hat{h}_n}(b)/b = \infty$.

[8] A subset $\mathcal{H}$ of a vector space is symmetric if $h \in \mathcal{H}$ implies $-h \in \mathcal{H}$.

2.3. *Continuous Functionals*

The sensitivity results in Subsection 2.1 apply to estimation of the structural function itself. If the object of interest is instead a functional of the structural function then it may be possible to construct estimates that a) are consistent under instrumental validity without any a priori restrictions on the true structural function (i.e., for any $h_0 \in \mathcal{B}_X$), and also b) have asymptotic bias under instrumental endogeneity that is at most proportional to the magnitude of the endogeneity.

The estimation of functionals of the structural function in NPIV models is analyzed extensively in the literature, for example in Ai & Chen (2003). Severini & Tripathi (2012) provide efficiency bounds for a class of linear functionals of the structural function in some statistical inverse problems, Ichimura & Newey (2017) expand upon their results. Following Severini & Tripathi (2012), we let the underlying function space $\mathcal{B}_X$ be $L_2(\mu_X)$. The function space $\mathcal{H}$ is the whole of $L_2(\mu_X)$. 'μ_X' denotes the distribution of the regressors.

Severini & Tripathi (2012) consider linear functionals of the form $\gamma_0 = E[w(X)h_0(X)]$

Where $w \in L_2(\mu_X)$ is a known weighting function. By the Reisz representation theorem, any continuous linear functional of the structural function (continuous in the sense of $L_2(\mu_X)$) can be written in the form above for some w.

Severini & Tripathi (2012) show that a linear functional of the form above is estimable at rate $\sqrt{n}$ only if there exists a function $\alpha \in L_2(\mu_Z)$ (where μ_Z is the probability measure of Z) so that:

$$w(X) = E[\alpha(Z)|X] \tag{2.2}$$

Under this same condition, robust estimation of γ_0 is achievable. If the condition fails, robust estimation is (under Assumptions 1.1 and 1.2) impossible without further restrictions to the parameter space.

Let $\hat{\gamma}_n$ be an estimator of γ_0 that has some probability limit. Fix h_0 and $\mu_{XZ\eta}$. The worst-case asymptotic bias of $\hat{\gamma}_n$ given $u_0 \in R(A)$ and $||u_0||_{L_2(\mu_Z)} \leq b$ is:

$$bias_{\hat{\gamma}_n}(b) = \sup_{u_0 \in R(A):\, ||u_0||_{L_2(\mu_Z)} \leq b} \lim_{n \to \infty} |\hat{\gamma}_n - \gamma_0|$$

In Theorem 2.4 we fully characterize those continuous linear functionals that can be estimated robustly. As stated above, the key condition is the existence of a function α that satisfies (2.2).

THEOREM 2.4 *Fix $\mu_{XZ\eta}$ so that Assumptions 1.1 and 1.2 hold. Suppose $\hat{\gamma}_n$ is consistent under instrumental validity whenever $h_0 \in L_2(\mu_X)$. That is, if $u_0 = 0$ then $|\hat{\gamma}_n - \gamma_0| \to^p 0$. Now fix the true structural function $h_0 \in L_2(\mu_X)$.*

a. Suppose there exists $\alpha \in L_2(\mu_Z)$ so that $w(X) = E[\alpha(Z)|X]$. Then the estimator is robust. In particular:

$$bias_{\hat{\gamma}_n}(b) = b \inf_{\substack{\alpha \in L_2(\mu_Z) \\ w(X) = E[\alpha(Z)|X]}} ||\alpha||_{L_2(\mu_Z)}$$

b. If there is no $\alpha \in L_2(\mu_Z)$ so that $w(X) = E[\alpha(Z)|X]$ then the estimator is non-robust and in fact for all $b > 0$, $bias_{\hat{\gamma}_n}(b) = \infty$.

2.4. *Discussion*

Theorems 2.1 and 2.3 may worry an empirical researcher. Theorem 2.1 shows that NPIV estimation methods that do not impose strong restrictions on the structural function can exhibit highly aberrant asymptotic behavior, no matter how small the degree of misspecification. If the researcher accepts that NPIV models are always at least mildly misspecified, then imposing at least some smoothness in estimation is paramount.

Theorem 2.2 shows that smoothness allows for robust estimation. However, Theorem 2.3 shows that imposing nonparametric smoothness still results in estimators that are less robust than those that impose parametric restrictions.

Even if the researcher accepts a priori that say, the true structural function lies in a particular Hölder ball, nonetheless it may be optimal to impose even stronger smoothness restrictions in order to reduce the sensitivity to a failure of instrumental validity. However, if the stronger restrictions do not hold then an estimator that imposes them is asymptotically biased even if instruments are valid.

In short, the researcher faces a trade-off between the sensitivity to instrumental endogeneity and the asymptotic bias that results from imposing overly strong smoothness restrictions. The methods we present in the next section are motivated in part by this trade-off.

3. THE PARTIAL IDENTIFICATION APPROACH

The results in the previous section show that point estimation in an NPIV model entails a trade-off between two different sources of asymptotic bias. A researcher must impose some restrictions (say, smoothness assumptions) on the structural function in order to reduce the sensitivity of the estimates to the failure of instrumental validity. But if the restrictions are too strong then imposing them imparts asymptotic bias.

In this section we propose a partial identification approach to NPIV estimation which explicitly accounts for both sources of bias. This set identification strategy allows us to evaluate error bands that account for all misspecification and allows us to achieve smallest possible worst-case asymptotic bias in point estimation. We use a priori bounds on the deviation from instrumental validity and some restrictions on the structural function (e.g., a bound on its second derivatives). Our approach is similar in spirit to that of Conley *et al.* (2008) who perform partial identification in the linear IV model allowing for some failure of instrumental validity.

The set estimation problem is (under weak conditions) well-posed. This contrasts with the case of standard point estimation in NPIV models. In our approach, the NPIV moment condition is replaced with an inequality constraint. Our convergence rate results depend crucially on the existence of functions in the parameter space for which this inequality does not bind. As such, our results do not extend to the point identified case and differ fundamentally from those in, for example, Chen & Christensen (2018).

We assume that the structural function h_0 lies in $\mathcal{B}_X$, the Banach space of bounded functions on the support of X. To achieve partial identification we assume that h_0 satisfies condition **a.** below, which is expressed as a condition on an element $h \in \mathcal{B}_X$:

a. $|E[Y - h(X)|Z]| \leq b$

The inequality is understood to hold with probability 1 and the bound b is treated as a priori known. In the case of correct specification the structural function satisfies **a.** with $b = 0$, thus condition **a.** weakens the NPIV moment restriction to allow for a limited degree misspecification.[9]

In addition, we assume that the structural function h_0 lies in a set of functions $\mathcal{H} \subseteq \mathcal{B}_X$. The results in Section 2 suggest that the space $\mathcal{H}$ should be sufficiently restrictive for the identified set to be meaningful. We assume that $\mathcal{H}$ can be expressed in terms of inequality constraints as follows. Let $\mathbb{T}$ be a linear functional from $\mathcal{B}_X$ to $(\mathcal{B}_X)^d$. That is, for any $h \in \mathcal{B}_X$, $\mathbb{T}[h](x)$ is a length-d column vector and each coordinate of $\mathbb{T}[h]$ is a function in $\mathcal{B}_X$. Let c be a function in $(\mathcal{B}_X)^d$. Then $\mathcal{H}$ is the set of functions in $\mathcal{B}_X$ that

[9] To map the constraint **a.** on h_0 into the functional notation developed in previous sections, let $\mathcal{B}_Z$ be the space of essential-supremum bounded functions on the support of Z equipped with the essential-supremum norm. Then the almost sure inequality can be written as $||g_0 - A[h_0]||_{\mathcal{B}_Z} \leq b$.

satisfy:

b. $\mathbb{T}[h](x) \leq c(x)\, \forall x \in \mathcal{X}$

Thus if $\mathbb{T}[h](x)$ is a vector of length d, then $c(x)$ is a column vector of the same length and the inequality is assumed to hold component-wise. Technically, the constraint **b.** should also require that h is in the domain of $\mathbb{T}$ (which could be a proper subset of $\mathcal{B}_X$), we leave this implicit for ease of exposition.

In the empirical application in Section 4 the regressors are one-dimensional and we take $\mathcal{H}$ to be the set of functions that map to the unit interval and have second derivatives bounded by a given constant. In a slight abuse of notation we simply denote the constant by c. In this case $c(x) = (1, 0, c, c)'$ and the operator $\mathbb{T}$ is given by:

$$\mathbb{T}[h](x) = \big(h(x), -h(x), \frac{\partial^2}{\partial x^2}h(x), -\frac{\partial^2}{\partial x^2}h(x)\big)'$$

The conditions **a.** and **b.** define a linear conditional moment inequality model. The moment inequality in condition **a.** may not appear linear, but note that for any scalars y and b, the inequality $|y| \leq b$ is equivalent to the two linear inequalities $y \leq b$ and $-y \leq b$.

We denote by Θ the set of functions in $\mathcal{H}$ that satisfy the moment inequality in **a.**, or equivalently the functions in $\mathcal{B}_X$ that satisfy constraints **a.** and **b.**. We refer to Θ as the 'identified set of functions'. Let $\underline{\theta}$ and $\bar{\theta}$ denote the lower and upper envelopes of the identified set of functions Θ. Our goal is to estimate these envelopes. For a given x in the support of X, let 'Θ_x' denote the identified set for the value of the structural function at x. A value θ is in Θ_x if and only if $\theta = h(x)$ for some function $h \in \Theta$. Proposition 3.1 stated and proven in Appendix B shows that Θ_x is an interval with end points $\underline{\theta}(x)$ and $\bar{\theta}(x)$. This motivates our focus on estimation of the envelopes: consistent estimation of the envelopes is equivalent to consistent estimation of the identified set for $h_0(x)$ at each point x in the support of

X.

Let $\hat{h}_n$ be a point estimator of the structural function that converges pointwise in probability to a limit h_∞. Under the assumption that h_0 satisfies **a.** and **b.** the pointwise worst-case asymptotic bias of $\hat{h}_n$ at some $x \in \mathcal{X}$ is given by:

$$\sup_{h_0 \in \Theta} \operatorname*{plim}_{n \to \infty} |\hat{h}_n(x) - h_0(x)| = \max\left\{|h_\infty(x) - \underline{\theta}(x)|, |h_\infty(x) - \bar{\theta}(x)|\right\}$$

If $\hat{h}_n$ converges uniformly in probability to h_∞ then the uniform worst-case asymptotic bias is given by:

$$\sup_{h_0 \in \Theta} \operatorname*{plim}_{n \to \infty} \sup_{x \in \mathcal{X}} |\hat{h}_n(x) - h_0(x)| = \sup_{x \in \mathcal{X}} \max\left\{|h_\infty(x) - \underline{\theta}(x)|, |h_\infty(x) - \bar{\theta}(x)|\right\}$$

Thus an estimator $\hat{h}_n$ that converges uniformly in probability achieves minimal pointwise and uniform worst-case asymptotic bias if and only if it converges uniformly to $\frac{1}{2}(\underline{\theta} + \bar{\theta})$.

We can define $\underline{\theta}$ and $\bar{\theta}$ formally as follows:

$$\begin{aligned}
\underline{\theta}(x) = &\inf_{h \in \mathcal{B}_X} h(x) \\
&\text{subject to conditions } \textbf{a.} \text{ and } \textbf{b.} \\
\bar{\theta}(x) = &\sup_{h \in \mathcal{B}_X} h(x) \\
&\text{subject to conditions } \textbf{a.} \text{ and } \textbf{b.}
\end{aligned}$$

The problems above cannot be solved in practice and therefore we refer to these as the 'infeasible' problems. Constraint **a.** involves a conditional expectation and therefore depends on the distribution of the data which is not a priori known. Furthermore, the optimization is over the space $\mathcal{B}_X$ of bounded functions on $\mathcal{X}$ and so if X is continuously distributed then $\mathcal{B}_X$ is infinite-dimensional. Finally, if X and Z are continuously distributed then the inequalities in constraints **a.** and **b.** must be enforced at an infinite set of values.

We describe a method for estimating the envelopes $\underline{\theta}$ and $\bar{\theta}$. Estimation requires that we replace the infeasible problems above with feasible ones. Instead of optimizing over $\mathcal{B}_X$ we optimize over a finite-dimensional subspace whose dimension grows with the sample size. We replace the conditional expectation in **a.** by an empirical analogue constructed using non-parametric regression in a first stage. We enforce the inequalities in each constraint only on finite grids that become increasingly fine as the sample size grows.

3.1. *An Estimator of the Identified Set*

Let Φ_n be a length-K_n column vector of basis functions defined on $\mathcal{X}$, where the dimension K_n grows with the sample size. We assume that each component of Φ_n is in the domain of $\mathbb{T}$. For example, in the case above of $\mathcal{H}$ a set of functions with bounded second derivatives, Φ_n must be twice differentiable. Let $\mathbb{T}[\Phi_n']$ denote the d-by-K_n matrix whose k^{th} column is $\mathbb{T}$ applied to the k^{th} component of Φ_n. Because $\mathbb{T}$ is linear $\mathbb{T}[\Phi_n'\beta] = \mathbb{T}[\Phi_n']\beta$.

The researcher estimates the reduced-form function $g_0(Z) = E[Y|Z]$ using standard non-parametric regression methods. The researcher also estimates the length-K_n column vector of regression functions $\Pi_n(Z) = E[\Phi_n(X)|Z]$. Let $\hat{g}_n$ denote the estimate of g_0 and $\hat{\Pi}_n$ the estimate of Π_n.

In our empirical application we use series regression for the first stage. Let Ψ_n be a length-L_n column vector of basis functions defined on $\mathcal{Z}$ with $L_n \to \infty$. Let $\hat{Q}_n = \frac{1}{n}\sum_{i=1}^n \Psi_n(Z_i)\Psi_n(Z_i)'$. Then (assuming $\hat{Q}_n$ is non-singular) the series first-stage regression functions are defined by:

$$\hat{g}_n(z) = \Psi_n(z)'\hat{Q}_n^{-1}\frac{1}{n}\sum \Psi_n(Z_i)Y_i \tag{3.1}$$

$$\hat{\Pi}_n(z) = \Psi_n(z)'\hat{Q}_n^{-1}\frac{1}{n}\sum \Psi_n(Z_i)\Phi_n(X_i)' \tag{3.2}$$

Let $\mathcal{X}_n$ be a finite grid of points in the support of X and let $\mathcal{Z}_n$ be a grid of points in the support of Z. The conditions **a.'** and **b.'** below are constraints on a vector $\beta \in \mathbb{R}^{K_n}$:

a.' $|\hat{g}_n(z) - \hat{\Pi}_n'(z)\beta| \leq b$, $\forall z \in \mathcal{Z}_n$

b.' $\mathbb{T}[\Phi_n](x)'\beta \leq c(x)$, $\forall x \in \mathcal{X}_n$

The estimators of $\underline{\theta}(x)$ and $\bar{\theta}(x)$ for a given x in the support of X are $\hat{\underline{\theta}}_n(x)$ and $\hat{\bar{\theta}}_n(x)$ respectively. These are defined as the solutions to the following linear programming problems:

$$\hat{\bar{\theta}}_n(x) = \max_{\beta \in \mathbb{R}^{K_n}} \Phi_n(x)'\beta$$

$$\text{subject to conditions } \textbf{a.'} \text{ and } \textbf{b.'}$$

$$\hat{\underline{\theta}}_n(x) = \min_{\beta \in \mathbb{R}^{K_n}} \Phi_n(x)'\beta$$

$$\text{subject to conditions } \textbf{a.'} \text{ and } \textbf{b.'}$$

Unlike the problems that define $\underline{\theta}(x)$ and $\bar{\theta}(x)$, the problems above are feasible: they can be solved in practice. They are linear programming problems each with K_n scalar parameters and $2|\mathcal{Z}_n| + d|\mathcal{X}_n|$ linear constraints (where $|\mathcal{X}_n|$ is the number of points in the grid $\mathcal{X}_n$ and $|\mathcal{Z}_n|$ is the number of points in $\mathcal{Z}_n$, recall d is the dimension of $\mathbb{T}[h](x)$).

Using the envelope estimators above we can evaluate a central estimator $\hat{h}_n$ by setting $\hat{h}_n(x) = \frac{1}{2}\big(\hat{\underline{\theta}}_n(x) + \hat{\bar{\theta}}_n(x)\big)$. If the envelope estimators are uniformly consistent then this estimator converges uniformly to $\frac{1}{2}(\underline{\theta} + \bar{\theta})$. Thus it achieves minimal worst-case uniform and pointwise worst-case asymptotic bias under our assumptions on h_0. The envelopes can be understood as error bounds on $\hat{h}_n$ which account for the possibility of misspecification.

3.2. *Consistency and Convergence Rates*

Let us introduce some additional notation. '$||\cdot||_2$' denotes the Euclidean norm. For any $h \in \mathcal{B}_X$, $|h|_\infty$ is the supremum norm of h, that is $|h|_\infty = \sup_{x \in \mathcal{X}} |h(x)|$. In a slight abuse of notation, for any $g \in \mathcal{B}_Z$, $|g|_\infty$ is the essential supremum norm of g, i.e., the infimum of the real numbers that exceed $|g(Z)|$ with probability 1. We say that a vector-valued function f

with domain $\mathcal{W} \subseteq \mathbb{R}^k$ is Lipschitz continuous with Lipschitz constant ξ if and only if:

$$\sup_{w_1, w_2 \in \mathcal{W}: w_1 \neq w_2} \frac{||f(w_1) - f(w_2)||_2}{||w_1 - w_2||_2} = \xi$$

Let $D_{1,n}$ denote the upper bound on the distance between any point in $\mathcal{X}$ and the nearest gridpoint in $\mathcal{X}_n$. That is, $D_{1,n} = \sup_{x_1 \in \mathcal{X}} \min_{x_2 \in \mathcal{X}_n} ||x_1 - x_2||_2$. Similarly, define the sequence $D_{2,n} = \sup_{z_1 \in \mathcal{Z}} \min_{z_2 \in \mathcal{Z}_n} ||z_1 - z_2||_2$. Define a sequence $C_n = \sup_{\beta \in \mathbb{R}^{K_n}} \frac{||\beta||_2}{|\Phi_n' \beta|_\infty}$.

The following assumptions provide high-level conditions for uniform consistency and particular uniform convergence rates for our estimated envelopes. We provide more primitive conditions later in this section.

ASSUMPTION 3.1 $\mathbb{T} : \mathcal{B}_X \to (\mathcal{B}_X)^d$ is linear, $\mathbb{T}[h](x) \leq c(x)$ implies $|h(x)| \leq \bar{c}$ for some $0 < \bar{c} < \infty$, and for some $\underline{c} > 0$, $c(x) \geq \underline{c}$ for all $x \in \mathcal{X}$.

ASSUMPTION 3.2 There is a sequence of positive scalars $a_n \to 0$ so that:

$$|\hat{g}_n - g_0|_\infty + \sup_{\beta \in \mathbb{R}^{K_n} : \Phi_n' \beta \in \mathcal{H}} |(\hat{\Pi}_n - \Pi_n)' \beta|_\infty = O_p(a_n)$$

ASSUMPTION 3.3 There is a sequence of positive scalars $\kappa_n \to 0$ so that for any $h \in \mathcal{H}$ there exists $\beta_n \in \mathbb{R}^{K_n}$ with $|\Phi_n' \beta_n - h|_\infty \leq \kappa_n$ and:

$$\mathbb{T}[\Phi_n'](x)\beta_n \leq \mathbb{T}[h](x), \forall x \in \mathcal{X}$$

ASSUMPTION 3.3 i. Both Φ_n and $\mathbb{T}[\Phi_n]$ are Lipschitz continuous with constant at most ξ_n and c is Lipschitz continuous with some constant. ii. With probability approaching 1 both $\hat{g}_n$ and $\hat{\Pi}_n$ are Lipschitz continuous with constant at most G_n. iii. $D_{1,n}, D_{2,n} \to 0$, $C_n \xi_n D_{1,n} \to 0$ and $C_n G_n D_{2,n} \to 0$.

Assumption 3.1 places conditions on $\mathbb{T}$ and therefore on $\mathcal{H}$. It implies that elements of $\mathcal{H}$ are bounded and that $\mathcal{H}$ is convex.

Assumption 3.2 allows us to control the effect of first-stage estimation error in the constraints of the feasible problem. In Theorem 3.2 we establish a rate for a_n when $\hat{g}_n$ and $\hat{\Pi}_n$ are series regression estimators and some primitive conditions hold. The rate in Theorem 3.2 is uniform over all choices of the sequence $\{K_n\}_{n=1}^{\infty}$.

Assumption 3.3 allows us to control the error from the replacement of the space $\mathcal{B}_X$ with the finite-dimensional space of functions of form $\Phi_n'\beta$ in the feasible problem. In Theorem 3.3 below we provide a rate for κ_n for the setting in Section 4.

Assumption 3.4 allows us to control the error from the use of finite grids $\mathcal{X}_n$ and $\mathcal{Z}_n$ in the constraints of the feasible problem. If Assumption 3.4 fails then the estimated envelopes may be too loose in large samples and so the set estimates may be too conservative in the limit.

Theorem 3.1 *Suppose Assumptions 3.1, 3.2, 3.3 and 3.4 hold and there exists $h \in \mathcal{H}$ with $|E[Y - h(X)|Z]| < b$. Then:*

$$\begin{aligned}|\hat{\underline{\theta}} - \underline{\theta}|_\infty &= O_p(a_n + \kappa_n + C_n\xi_n D_{1,n} + C_n G_n D_{2,n}) \\ &= o_p(1)\end{aligned}$$

$$\begin{aligned}|\hat{\bar{\theta}} - \bar{\theta}|_\infty &= O_p(a_n + \kappa_n + C_n\xi_n D_{1,n} + C_n G_n D_{2,n}) \\ &= o_p(1)\end{aligned}$$

Note that along with the Assumptions 3.1 to 3.4 we also require that there exists $h \in \mathcal{H}$ with $|E[Y - h(X)|Z]| < b$. If the identified set is non-empty then failure of the condition is knife-edge: if b were even slightly larger then the condition would hold, if b were even slightly smaller then the identified set would be empty.

Theorem 3.1 demonstrates the well-posedness of the set estimation problem. The first-stage rate a_n is not premultiplied by some growing factor like a 'sieve measure of ill-posedness' (Blundell *et al.* (2007)).

The final two terms in each rate in Theorem 3.1 depend on $D_{1,n}$ and $D_{2,n}$ which capture the density of the grids $\mathcal{X}_n$ and $\mathcal{Z}_n$. The terms a_n and κ_n do not depend on the grids, and so if the grids becomes dense sufficiently quickly then the rates in Theorem 3.1 simplify to $a_n + \kappa_n$. In practice the grid densities are limited by computational considerations.

If K_n grows quickly to infinity then the approximation error κ_n converges rapidly to zero. However, the first-stage error rate a_n may depend on K_n and so a faster rate for κ_n could result in a slower rate a_n. Below we provide primitive conditions for a first-stage rate a_n which is independent of K_n. Therefore, under these conditions a faster rate for K_n must lead to at least a weakly faster rate of convergence for the estimates. If K_n grows sufficiently quickly the term κ_n is dominated by a_n. Again, in practice K_n is restricted by computational limitations.

The following primitive conditions allow us to derive a first-stage rate a_n. Our analysis builds heavily on Belloni *et al.* (2015), we apply their results directly to get a rate for $|g_0 - \hat{g}_n|_\infty$ and adapt steps in their proof to allow for uniformity over a set of series regressions.

In the Assumptions below, we say a function $f : \mathcal{Z} \to \mathbb{R}$ is of Hölder smoothness class $s \in (0, 1]$ with constant ξ if and only if:

$$\sup_{z_1, z_2 \in \mathcal{Z} : z_1 \neq z_2} \frac{|f(z_1) - f(z_2)|}{||z_1 - z_2||_2^s} = \xi$$

Let $\lfloor s \rfloor$ denote the largest integer less than s. We say f is of Hölder smoothness class $s > 1$ with constant ξ if and only if all the derivatives of f of order weakly less than $\lfloor s \rfloor$ are uniformly bounded by ξ and all the derivatives of order exactly $\lfloor s \rfloor$ are of Hölder smoothness class $s - \lfloor s \rfloor$ with constant ξ. If we wish to leave the constant unspecified we simply say say f is Hölder of smoothness class s.

Further, for any $\delta > 0$, let $\mathcal{N}(\mathcal{H}, |\cdot|_\infty, \delta)$ denote the smallest number of $|\cdot|_\infty$-balls of radius δ that can cover $\mathcal{H}$.

Let $\dim(Z)$ denote the dimension of Z.

ASSUMPTION 3.5 i. The eigenvalues of the matrix $Q_n = E[\Psi_n(Z_i)\Psi_n(Z_i)']$ are bounded uniformly above and away from zero. ii. $\mathcal{Z}$ is bounded and the distribution of X given Z admits a conditional density $f_{X|Z}$ so that for any $x \in \mathcal{X}$ the function $f_{X|Z}(x, \cdot)$ is of Hölder smoothness class $s > 0$ with constant at most $\bar{\ell}$.

ASSUMPTION 3.6 For any $s > 0$ there is a sequence $R_n(s) \to 0$ so for any $g \in \mathcal{B}_Z$ that is of Hölder smoothness class s with constant ξ:

$$\sup_{z\in\mathcal{Z}} |g(z) - \Psi_n(z)' Q_n^{-1} E[\Psi_n(Z)g(z)]| \leq \xi R_n(s)$$

ii. For all $z \in \mathcal{Z}$, $||\Psi_n(z)||_2 \leq \bar{\xi}_n$. The function α_n defined by $\alpha_n(z) = \frac{\Psi_n(z)}{||\Psi_n(z)||}$ is Lipschitz continuous with constant ℓ_n. iii. $\mathcal{H}$ has finite Dudley entropy integral:

$$\int_0^1 \sqrt{log\mathcal{N}(\mathcal{H}, |\cdot|_\infty, u)}du < \infty$$

iv. $\frac{\bar{\xi}_n^2 log(L_n)}{n} \to 0$

ASSUMPTION 3.7 The function $g_0(Z) = E[Y|Z]$ is of Hölder smoothness class $s > 0$. For $m > 2$, $E\big[|Y - E[Y]|^m \big| Z\big] < \infty$, $\bar{\xi}_n^{2m/(m-2)} log(L_n)/n = O(1)$, $L_n log(L_n)/(n^{1-2/m}) = O(1)$ and $L_n^{2-2/\dim(Z)}/n = O(1)$. ii. $log(\ell_n) = O\big(log(L_n)\big)$, $\bar{\xi}_n = O(\sqrt{L_n})$ and $R_n(s) = O(L_n^{-s_0(s)/\dim(Z)})$ for some function $s_0 : \mathbb{R}_{++} \to \mathbb{R}_{++}$ and $R_n(s) = O(L_n^{-1/2})$.

Assumption 3.5 restricts the joint distribution of the regressors X and instruments Z. The assumption on the eigenvalues of Q_n is standard in the series estimation literature. Smoothness of the conditional density ensures that for any $h \in \mathcal{B}_X$, $A[h]$ is smooth.

Assumptions 3.6.i and 3.6.ii. can be verified for commonly used basis functions. 3.6.iii is a condition on the metric entropy of $\mathcal{H}$. Loosely speaking, it states that $\mathcal{H}$ is sufficiently restrictive. Conditions on metric entropy are commonplace in the sieve estimation literature (see Chen (2007)). Spaces of sufficiently smooth functions typically obey the condition (see e.g., Wainwright (2019) Chapter 5). In our empirical application, the set $\mathcal{H}$ can be shown to contain functions on an interval that are uniformly Lipschitz, such a set of functions must satisfy the assumption.[10] 3.6.iv ensures the empirical analogue of Q_n converges in an appropriate sense by Rudelson's law of large numbers for matrices (Rudelson (1999)).

Assumption 3.7.i provides conditions on the joint distribution of Y and Z that allow us to apply results from Belloni *et al.* (2015) to derive a convergence rate for $\hat{g}_n$. In our empirical application in Section 4 the dependent variable is bounded and so we can set m arbitrarily large. 3.7.ii specifies rates for some of the sequences mentioned in the other assumptions. It can be verified for a given choice of basis functions. If $s \geq 1/2$, then the Assumption holds for the one-dimensional B-spline case in Section 4. In particular, for B-splines, $s_0(s)$ is equal to the minimum of the smoothness class s and the order of the splines (see e.g., DeVore & Lorentz (1993)). To derive a rate without assuming 3.7.ii, one can directly apply Lemma 3.3 (stated in the Appendix) and Theorem 4.3 of Belloni *et al.* (2015).

The following theorem gives a rate a_n in terms of L_n and n. The key steps in the proof are Lemma 3.3 stated in Appendix B, which builds on ideas in Belloni *et al.* (2015) and may be of some independent interest, and Theorem 4.3 in Belloni *et al.* (2015).

THEOREM 3.2 *Suppose Assumptions 3.5, 3.6 and 3.7 hold. Let $\hat{g}_n$ and $\hat{\Pi}_n$*

[10] Note that whether or not a set of functions obeys a condition like 3.6.iii is closely related to whether the set is a universal Donsker class (see, Dudley (1987)).

be the series estimators defined in (3.1) and (3.2). Then uniformly over sequences $\{K_n\}_{n=1}^{\infty}$*:*

$$|\hat{g}_n - g_0|_\infty + \sup_{\beta \in \mathbb{R}^{K_n}:\, \Phi_n'\beta \in \mathcal{H}} |(\hat{\Pi}_n - \Pi_n)'\beta|_\infty$$
$$= O_p\left(\sqrt{\frac{L_n log(L_n)}{n}} + L_n^{-s_0(s)/\dim(Z)}\right) = o_p(1)$$

If the conditions of the theorem hold, then setting L_n optimally we can achieve first-stage rate $a_n = \left(\frac{log(n)}{n}\right)^{s_0(s)/(2s_0(s)+\dim(Z))}$. In the case of $s_0(s) = s$ (which holds for B-spline bases of order greater than s) this is the best uniform rate possible (see Belloni *et al.* (2015)). The rate given in Theorem 3.2 is independent of the sequence $\{K_n\}_{i=1}^{\infty}$. Therefore, if the conditions for Theorem 3.2 hold, the optimal rate in Theorem 3.1 is achieved by letting K_n grow as fast as possible.

Finally, we provide a rate for κ_n in Assumption 3.3 for the setting in our empirical application.

THEOREM 3.3 *Let* $\mathcal{H}$ *contain functions that map from a closed interval* $[a, b]$ *to* $[0, 1]$ *so that any* $h \in \mathcal{H}$ *is twice-differentiable with* $|\frac{\partial^2}{\partial x^2} h|_\infty \leq c$*. Let* Φ_n *be a vector of* s_0*-order B-spline basis functions with* K_n *knot points evenly spaced between* $[a, b]$*. If* $s_0 \geq 3$ *then Assumption 3.3 holds with* $\kappa_n = O(K_n^{-\frac{1}{2}})$*.*

4. AN EMPIRICAL APPLICATION

To demonstrate the usefulness of our partial identification approach we revisit an existing application of NPIV methods. In particular, we replicate the NPIV estimation results in Section 5.1 of Horowitz (2011) which estimates a shape-invariant Engel curve for food using data from the British Family Expenditure Survey .[11] Horowitz's application is in turn based on

[11]We made use of the data file that accompanies Horowitz (2011) and adapted the accompanying code in order to evaluate Horowitz's estimator and B-spline bases for our

Blundell *et al.* (2007) who also carry out NPIV estimation of shape-invariant Engel curves and use the same data.

From Horowitz (2011): "The data are 1655 household-level observations from the British Family Expenditure Survey. The households consist of married couples with an employed head-of-household between the ages of 25 and 55 years."

4.1. *Shape-Invariant Engel Curves and the Case for Mild Misspecification*

Blundell *et al.* (2007) aim to estimate 'structural' Engel curves. A structural Engel curve measures the budget share that would be spent on a good if the household's total expenditure were set exogenously to some level. One can imagine an ideal randomized experiment in which the household's weekly expenditure on nondurable goods is decided at random (i.e., exogenously) by a researcher. The household then decides how to allocate this budget across different classes of goods. The relationship between total expenditure and the budget share allocated to a good in such an experiment is a structural Engel curve.

In observational settings, the share of household wealth allocated to expenditure on nondurables in a given period is decided by the household. Therefore, it is likely associated with the household's underlying preferences. In short, total household expenditure on nondurable goods is endogenous.

Blundell *et al.* (2007) and Horowitz (2011) hope to recover structural shape-invariant Engel curves using household income as an instrument for total expenditure. Suppose one controls for fixed household characteristics like household size and socio-cultural make-up. Any remaining variation in income likely reflects outside economic shocks that are unrelated to variation in household tastes.

However, the household characteristics controlled for by Blundell *et al.*

own methods.

(2007) and Horowitz (2011) are limited to a small selection of coarsely measured demographic variables.[12] There is certainly some remaining variation in household features like the ages, ethnicities and education levels of each household's constituents. If the remaining variation is small or is only weakly associated with income or tastes, then income is only mildly endogenous. Therefore, in this setting it is of interest to see what can be inferred under some small failure of instrumental validity.

4.2. *Estimation*

Below we describe an application of our methods to the estimation of h_0 the structural, shape-invariant Engel curve for food. The dependent variable Y is the share of total expenditure on non-durables that a household spends on food. The endogenous variable X is the logarithm of the household's total expenditure and Z is the logarithm of household income.

Our methodology requires we select a vector of basis functions Φ_n. We follow Horowitz (2011) and use spaces of fourth-order (cubic) B-splines with evenly-spaced knot points.[13] Suppose we set K_n equal to some $k > 4$, the length-k vector of basis functions can then be defined as follows. Let $\{l_{k,j}\}_{j=1}^{k-4}$ be a sequence of scalars known as 'knot points'. Then:

$$\Phi_n(x) = M_k(1, x, x^2, x^3, |x - l_{k,1}|_+^3, |x - l_{k,2}|_+^3, ..., |x - l_{k,k-4}|_+^3)'$$

For a particular non-singular k-by-k matrix M_k. The function $|\cdot|_+$ returns the positive part of its argument, that is $|y|_+ = 1\{y \geq 0\}|y|$. For the knot points we set:

$$l_{k,j} = \frac{j}{k-3} x_{max} + \frac{k-3-j}{k-3} x_{min}$$

[12] In both papers, to control for the demographic information only a sub-sample with homogeneous characteristics is analyzed. Blundell *et al.* (2007) incorporate additional dummy variable controls when performing estimation on the sub-sample.

[13] See de Boor (2014) for a practical introduction to B-splines.

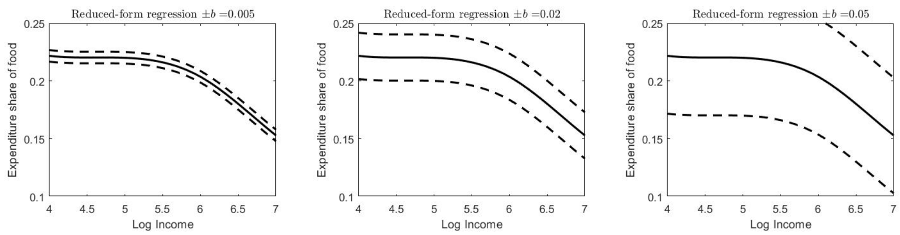
Where x_{max} and x_{min} are respectively the largest and smallest observed values of the regressor X in the data.

We carry out our first-stage estimates using series regression onto cubic B-splines. If we set $L_n = k$ then the vector of basis functions Ψ_n is defined exactly as Φ_n above albeit with domain $\mathcal{Z}$ rather than $\mathcal{X}$ and x_{max} and x_{min} replaced by the largest and smallest observed values of the instrument.

Our partial identification approach requires that we place an a priori bound b on the magnitude of $E[Y - h_0(X)|Z]$ and restrict the structural function h_0 to lie in some space $\mathcal{H}$. Here we take $\mathcal{H}$ to be the set of functions on $\mathcal{X}$ that take values in the unit interval and have second derivative bounded in magnitude by a constant c. An Engel curve is an expected budget share and must take values in $[0, 1]$ by definition, however a bound on the second derivative does not clearly follow from the setting. As such, we present results for a range of values both for the bound b on the deviation from instrumental validity and for the bound c on the second derivative.

To implement the methods detailed in Section 3, a researcher must choose K_n the number of basis functions and the grids $\mathcal{X}_n$ and $\mathcal{Z}_n$. Motivated by the results in Theorem 3.2 we set K_n to be large, specifically we let $K_n = 10$. The grid $\mathcal{X}_n$ consists of 100 evenly spaced points between the smallest and largest observed values of X (the same grid that we used to plot the curves in Figure 4.2). The grid $\mathcal{Z}_n$ consists of 100 evenly spaced points between the 0.005 and 0.995 quantiles of Z, we make this truncation because the first-stage regression functions are imprecisely estimated outside this region.

For our first-stage estimates we use nonparametric least squares regression on cubic B-spline basis functions defined over the log income. To estimate the reduced-form function g_0 we regress by least squares the dependent variable Y on the B-spline basis over the instruments Z described above with four interior knot points (this basis is six-dimensional). To estimate Π_{K_n} we regress $\Phi_{K_n}(X)$ on the same B-spline basis over Z used to estimate



(a) (b) (c)

Figure 4.1: The Reduced-Form Regression

The results of the reduced-form regression. The dark line in each of the sub-figures is the result of regressing the expenditure share of food on cubic B-spline basis functions with four evenly spaced interior knot points defined over log income. The dotted lines are the reduced-form regression plus or minus a value of the bound b.

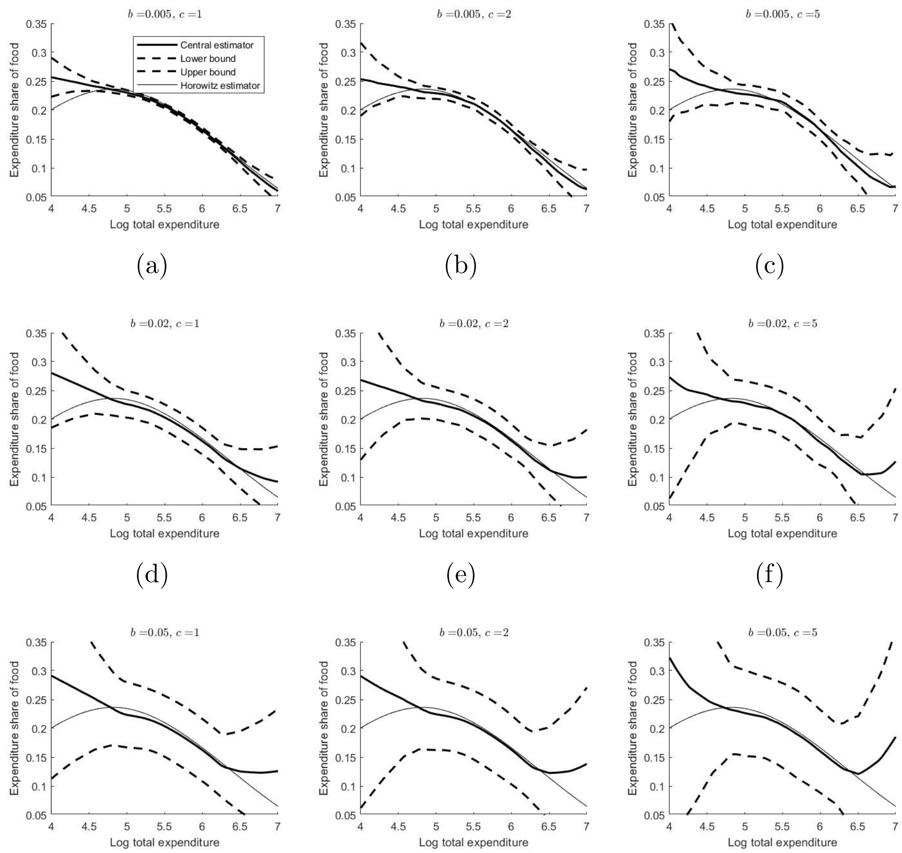
g_0.

The result of the reduced-form regression of the expenditure share for food on income (that is, the estimate of g_0) is given below in Figure 4.1. The dark line in each of the sub-figures is the estimated reduced-form function $\hat{g}_n$ and the dotted lines are the reduced-form regression plus or minus a value of the bound b. In Sub-Figure 4.1.a, b is equal to 0.005, which represents a tight bound on the deviation from instrumental validity. In Sub-Figure 4.1.b, b is set equal to 0.02 and in Sub-Figure 4.1.c it is set to 0.05. The units here are budget shares and so a deviation of 0.05 amounts to 5% of the total household budget on non-durable goods. Recall that the reduced-form is equal to the conditional expectation of the structural function plus the deviation of the NPIV moment condition from zero. That is $g_0(Z) = E[h_0(X)|Z] + u_0(Z)$. Therefore, if the essential supremum norm of u_0 is indeed bounded by b, then ignoring estimation error in $\hat{g}_n$, the dotted lines contain the hypothetical reduced-form function were there no deviation from instrumental validity (i.e., $u_0 = 0$) and everything else were held fixed.

Figure 4.2 below contains the results of our set estimation procedure. The figure contains nine sub-figures each corresponding to a different set of values for the bounds b and c. In each sub-figure the lower and upper dotted lines represent $\underline{\hat{\theta}}$ and $\hat{\bar{\theta}}$ the estimated upper and lower envelopes of the identified set. The thick black line represents the central estimator (the half-way point between the envelopes) described in the previous section. As noted in that section, if the envelope estimates are consistent and our assumptions on h_0 hold then this point-estimator achieves smallest possible worst-case asymptotic bias. The thin black line is the estimator evaluated by Horowitz (2011) which we include for comparison.

The sub-figures in the first row all correspond to the tight bound on the magnitude of instrumental endogeneity, $b = 0.005$. This is the same value of b shown in Sub-Figure 4.1.a above. The sub-figures on the second row correspond to the looser bound $b = 0.02$ as in Sub-Figure 4.1.b above and those in the final row correspond to $b = 0.05$ as in Sub-Figure 4.1.c. The three sub-figures in each column correspond to the same bound c on the second derivatives. The first column contains the sub-figures with bound 1 on the second derivatives, the second column contains those with bound $c = 2$ and the third column sub-figures correspond to the bound 5 on the second derivatives. As a benchmark, the second derivatives of Horowitz's estimated structural function are bounded in magnitude by 0.5. Thus in the sub-figures below the second derivatives are allowed to be either twice, four times or ten times the magnitude of those in Horowitz's estimates.

The results in Section 2 suggest that if the bound c on the second derivatives is too loose then the identified set will be large, even if the bound b on the failure of instrumental validity is small. Sub-Figure 4.2.c shows that the envelopes have non-negligible width when the bound c on the second derivatives is set to 5 even with the bound b set to the low value of 0.005, or 0.5% of total expenditure. More generally we see that for each b, the envelope estimates are looser the further right the sub-figure, i.e., the looser the bound c on the second derivatives.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 4.2: Set-Estimated Engel Curves

The results of our set estimation procedure. Each sub-figure corresponds to a different set of values for the bounds b and c. The values for these quantities are given above each sub-figure. In each sub-figure, the lower and upper dotted lines represent $\hat{\underline{\theta}}$ and $\hat{\bar{\theta}}$ the estimated upper and lower envelopes of the identified set. The thick black line represents the central estimator (which is the mean of $\hat{\underline{\theta}}$ and $\hat{\bar{\theta}}$). The thin black line is the estimate found in Horowitz (2011).

Sub-Figures in the first two rows all show a general downwards slope of the Engel curve for food at least for intermediate values of total expenditure. That is, the estimated envelopes in these sub-figures are fairly tight around a downward sloping central estimator. This suggests food is a necessary good, which conforms to conventional economic wisdom. We conclude therefore, that the finding that the Engel curve for food has a general downward slope is fairly robust to misspecification. The data support the finding even allowing for a failure of validity that amounts to 2% of the total expenditure on non-durables.

For sub-figures in the first row of Figure 4.2 (i.e., with the tight 1% bound on the failure of instrumental validity) the estimated envelopes of the identified set are tight enough to discern some non-linearity in the Engel curve (in the sense that the envelopes do not contain a straight line). The sub-figures seem to show an increasing downward slope for higher values of the log total expenditure. It is clear then that one must believe that income is only a weakly endogenous instrument in order to infer from the envelopes that the Engel curve demonstrates some non-linear trend.

Note that none of the results in Figure 4.2 provide evidence in favor of an upward sloping Engel curve for low values of total expenditure as found by Horowitz (2011). However, only in Sub-Figure 4.2.a do the estimated envelopes exclude Horowitz's estimates, and only by a small margin and for a narrow set of values for the log total expenditure. The envelopes are, in all sub-figures, wide for low values of total expenditure, which suggests that the Engel curve is poorly identified in this region. It seems then that the data do not provide meaningful evidence either for or against some positive slope in the Engel curve for low values of the expenditure on nondurables.

CONCLUSIONS

We demonstrate that NPIV estimates of the structural function are highly sensitive to misspecification in the form of invalid instruments. We show that the imposition of strong restrictions on the structural function can mitigate this problem, but can impart approximation bias. This motivates a partial identification approach to NPIV that allows a researcher achieve point estimation with minimal worst-case asymptotic bias and to evaluate error bounds (envelopes of the identified set) that explicitly account for possible misspecification.

The development of simple uniform confidence bands for envelopes of the identified set in conditional moment inequality models of the kind we study is an open question and is beyond the scope of this paper. Our model has an infinite-dimensional parameter space and (to the best of our knowledge) the only general analysis of inference in models of this kind is a working paper by Chernozhukov *et al.* (2016).[14] Their work may provide analytical tools for deriving valid confidence bands in our setting.

Future research may generalize our sensitivity results to a broader class of conditional moment restriction models. The non-robustness of NPIV estimators is tied to the ill-posedness of the NPIV estimating equation. In fact, a range of other nonparametric conditional moment restriction models are ill-posed. It seems likely then that estimation in these models is also non-robust. It may be useful to characterize precisely the class of nonparametric moment condition models associated with non-robust estimation and to extend our partial identification approach to these models.

REFERENCES

Ai, Chunrong, & Chen, Xiaohong. 2003. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, **71**(6), 1795–1843.

Andrews, Donald W. K., & Shi, Xiaoxia. 2013. Inference Based on Conditional Moment Inequalities. *Econometrica*.

Andrews, Donald WK. 2017. Examples of L2-complete and boundedly-complete distributions. *Journal of Econometrics*, **199**(2), 213–220.

Andrews, Isaiah, Gentzkow, Matthew, & Shapiro, Jesse M. 2017. Measuring the Sensitivity of Parameter Estimates to Estimation Moments. *The Quarterly Journal of Economics*, **132**, 1553–1592.

[14] Chernozhukov *et al.* (2013) considers some conditional moment inequality models with infinite-dimensional parameter spaces. However, these models can be rewritten as a set of conditional moment inequalities each with a finite-dimensional parameter space.

Armstrong, Timothy B., & Kolesar, Michal. 2018. Sensitivity Analysis using Approximate Moment Condition Models. *Cowles Foundation Discussion Papers.*

Belloni, Alexandre, Chernozhukov, Victor, Chetverikov, Denis, & Kato, Kengo. 2015. Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, **186**, 345–366.

Blundell, Richard, Chen, Xiaohong, & Kristensen, Dennis. 2007. Semi-nonparametric IV estimation of shape-invariant Engel curves. *Econometrica. Journal of the Econometric Society*, **75**(6), 1613–1669.

Canay, Ivan A., Santos, Andres, & Shaikh, Azeem M. 2013. On the testability of identification in some nonparametric models with endogeneity. *Econometrica. Journal of the Econometric Society*, **81**(6), 2535–2559.

Chen, Xiaohong. 2007. *Large Sample Sieve Estimation of Semi-nonparametric Models. The Handbook of Econometrics, JJ Heckman and EE Leamer (eds.), 6B.*

Chen, Xiaohong, & Christensen, Timothy M. 2018. Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression. *Quantitative Economics*, **9**, 39–84.

Chen, Xiaohong, & Pouzo, Demian. 2012. Estimation of Nonparametric Conditional Moment Models with Possibly Nonsmooth Generalized Residuals. *Econometrica.*

Chen, Xiaohong, Chernozhukov, Victor, Lee, Sokbae, & Newey, Whitney K. 2014. Local Identification of Nonparametric and Semiparametric Models. *Econometrica.*

Chernozhukov, Victor, Lee, Sokbae (Simon), & Rosen, Adam. 2013. Intersection bounds: estimation and inference. *Econometrica.*

Chernozhukov, Victor, Newey, Whitney K., & Santos, Andres. 2016. *Constrained Conditional Moment Restriction Models.*

Chetverikov, Denis, & Wilhelm, Daniel. 2017. Nonparametric instrumental variable estimation under monotonicity. *Econometrica.*

Conley, Timothy G., Hansen, Christian B., McCulloch, Robert E., & Rossi, Peter E. 2008. A semi-parametric Bayesian approach to the instrumental variable problem. *Journal of Econometrics*, **144**(1), 276–305.

Darolles, Serge, Fan, Yanqin, Florens, Jean-Pierre, & Renault, Eric. 2011. Nonparametric Instrumental Regression. *Econometrica.*

de Boor, Carl. 2014. *(B)asic-Spline Basics.*

DeVore, Ronald A., & Lorentz, George G. 1993. *Constructive Approximation.* Springer-Verlag.

D'Haultfoeuille, Xavier. 2011. On the completeness condition in nonparametric instrumental problems. *Econometric Theory*, **27**(3), 460–471.

Dudley, R. M. 1987. Universal Donsker classes and metric entropy. *The Annals of Probability*, **15**(4), 1306–1326.

Fichte, Johann, & Smith, William. 1848. *The Vocation of Man, translation by W. Smith.*

Florens, Jean-Pierre. 2011. *Non-parametric Models with Instrumental Variables.*

Freyberger, Joachim. 2017. On completeness and consistency in nonparametric instrumental variable models. *Econometrica. Journal of the Econometric Society*, **85**(5), 1629–1644.

Freyberger, Joachim, & Horowitz, Joel L. 2015. Identification and shape restrictions in nonparametric instrumental variables estimation. *Journal of Econometrics*, **189**(1), 41–53.

Freyberger, Joachim, & Masten, Matthew A. 2019. A Practical Guide to Compact Infinite Dimensional Parameter Spaces. *Econometric Reviews.*

Hall, Peter, & Horowitz, Joel L. 2005. Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics*, **33**, 2904–2929.

Horowitz, Joel L. 2011. Applied nonparametric instrumental variables estimation. *Econometrica. Journal of the Econometric Society*, **79**(2), 347–394.

Horowitz, Joel L., & Manski, Charles F. 1995. Identification and Robustness with Contaminated and Corrupted Data. *Econometrica*, **63**, 281.

Hu, Yingyao, & Shiu, Ji-Liang. 2018. Nonparametric identification using instrumental variables: sufficient conditions for completeness. *Econometric Theory*, **34**(3), 659–693.

Huber, Peter J. 2011. *Robust Statistics.*

Ichimura, Hidehiko, & Newey, Whitney K. 2017. *The influence function of semiparametric estimators.*

Manski, Charles F. 1989. Anatomy of the Selection Problem. *The Journal of Human Resources*, **24**, 343.

Newey, Whitney K. 1991. Uniform convergence in probability and stochastic equicontinuity. *Econometrica. Journal of the Econometric Society*, **59**(4), 1161–1167.

Newey, Whitney K., & Powell, James L. 2003. Instrumental Variable Estimation of Nonparametric Models. *Econometrica*, **71**, 1565–1578.

Rudelson, M. 1999. Random Vectors in the Isotropic Position. *Journal of Functional Analysis*, **164**, 60–72.

Santos, Andres. 2012. Inference in nonparametric instrumental variables with partial identification. *Econometrica. Journal of the Econometric Society*, **80**(1), 213–275.

Severini, Thomas A., & Tripathi, Gautam. 2012. Efficiency bounds for estimating linear functionals of nonparametric regression models with endogenous regressors. *Journal of Econometrics*, **170**, 491–498.

Wainwright, Martin J. 2019. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press.