Worst-Case Regression with Interval-Observed and Missing Covariates*

Preliminary Draft

Martin Fankhauser†, Gil Peled‡

February 28, 2026

Abstract

Researchers often confront regressions where key covariates are missing or only interval-observed. Standard remedies, such as imputation or auxiliary modeling assumptions, resolve ambiguity at the expense of credibility. Instead, we derive the sharp identified set, the smallest parameter set consistent with both the incomplete data and the maintained model, for a class of conditional expectation models without auxiliary assumptions. A computationally tractable Hausdorff consistent estimator for this set is provided with associated asymptotic uniformly valid confidence regions for the true parameter. This is achieved by extending an existing random set framework to infinite-dimensional Polish spaces. We apply our method to a multiple price list survey design measuring discount rates and document substantial sensitivity even for relatively short observed intervals.

Keywords: Partial identification, set identification, conditional moments, missing covariates, interval data, random sets, support function

JEL Codes: C14, C18, C21, C51

*We thank Francesca Molinari, Arie Beresteanu, Federico Bugni, Toru Kitagawa, Ivan Canay, Charles Manski, Filip Obradovic, Pamela Giustinelli, Ismael Mourifié and Marc Henry for helpful discussions.

†Bocconi University

‡Northwestern University

1 Introduction

Conditional moment equality models are widely used in applied econometrics, often taking the form of mean independence assumptions or regression specifications. In practice, however, these models frequently condition on continuous covariates that are only incompletely observed. For example, key regressors may be missing due to non-random sample attrition, or surveys may allow respondents to report variables such as income in brackets rather than as exact values.[1] Standard applied solutions, like imputation methods, conditioning on the interval itself or ignoring incomplete data points, replace the maintained model with a different estimand or introduce strong auxiliary assumptions. As a result, even qualitative conclusions can be driven more by auxiliary model assumptions than by the data and the original model. This paper instead takes a worst case approach: we impose the maintained conditional moment restrictions and the observed incomplete data, and characterize all parameter values consistent with both.

Despite recent advances in the partial identification literature on inference for moment equality and inequality models, for the setting studied there is currently no estimator of the sharp identified set, that is, the smallest set of parameter values consistent with both the conditional moment equality restrictions and the observed incomplete data, without imposing auxiliary assumptions. To the best of our knowledge only two papers discuss similar settings: First, Aucejo, Bugni, and Hotz (2017) study conditional moment models with missing outcomes or covariates, however only derive results for outer sets, i.e. supersets of the sharp identified set. Second, C. Manski and Tamer (2002) which derive the sharp identified set for interval regression problems under an auxiliary model assumption discussed in detail below. We instead establish identification results for a broad class of conditional moment models and propose a Hausdorff-consistent estimator of the sharp identified set, along with uniformly valid confidence sets without auxiliary assumption for general incomplete data structures.

[1] see for instance the Health and Retirement Survey in the US.

Formally, we study conditional moment models of the form

$$\mathbf{E}_{F_0}[m(Y, X, \theta_0)|X] = 0 \qquad F_0 - a.s. \tag{1}$$

where $(\Omega, \mathcal{A}, F_0)$ is the underlying probability space with $(\Omega, \mathcal{A})$ being standard Borel and F_0 is a complete, non-atomic probability measure on $(\Omega, \mathcal{A}, F_0)$. $Y : (\Omega, \mathcal{A}) \to (\mathcal{Y}, \mathcal{A}_Y)$ and $X : (\Omega, \mathcal{A}) \to (\mathcal{X}, \mathcal{A}_X)$ with dimensions d_y and d_x respectively, are random variables defined therein. $m : \mathbb{R}^{d_y} \times \mathbb{R}^{d_x} \times \mathbb{R}^{d_\theta} \to \mathbb{R}^{d_m}$ is a known function and the moment condition is fulfilled for the true parameter vector in a finite-dimensional space $\theta_0 \in \Theta \subseteq \mathbb{R}^{d_\theta}$. The partial identification problem emerges not from an incomplete model but instead from incomplete data. In particular, we assume that the regressors X and/or the outcome Y are only interval-observed in the following sense:

Assumption 1. *Let (Y_l, Y_u, X_l, X_u, Z) be defined on the same probability space as (Y, X), where $z \in \mathcal{Z}$ is a deterministic function of observed variables $W^O = (X_l, X_u, Y_l, Y_u)$ that indicates the missingness pattern. $W = (Y, X)$ are unobservable (but bounded) by the data generating process. We assume for each $z \in \mathcal{Z}$ that $P(Y \in [Y_l, Y_u], X \in [X_l, X_u]|Z = z) = 1$. Without loss of generality define:*

Z=0, both observed:	$Y_l = Y_u$ *and* $X_l = X_u$	F_0-*a.s.*
Z=1, Y observed, X interval:	$Y_l = Y_u$ *and* $X_l < X_u$	F_0-*a.s.*
Z=2, Y interval, X observed:	$Y_l < Y_u$ *and* $X_l = X_u$	F_0-*a.s.*
Z=3, both interval:	$Y_l < Y_u$ *and* $X_l < X_u$	F_0-*a.s.*

We denote by $P_Y, P_X, P_{Y,X}$ the pushforward measures induced by $Y, X, (Y, X)$ respectively. The associated conditional distributions are denoted by e.g. $P_{Y|X}$.

This includes the case of missing data if $[Y_l, Y_u] = \mathcal{Y}$ and/or $[X_l, X_u] = \mathcal{X}$. Although, we also consider the case of missing or interval-valued outcomes, in the case for which only the

outcome is observed incompletely (i.e. $\mathcal{Z} = \{0, 2\}$) solutions to the problem exist for linear regressions and are addressed in Beresteanu and Molinari (2008) and Molinari (2020). This subproblem with all X observed is considerably easier as for instance the OLS estimator $\theta = (X'X)^{-1}(X'Y)$ is a linear function of the outcome, whereas it is non-linear in the covariates.

This paper proposes an estimator for the sharp identified set, i.e. the smallest set of parameter values consistent with the model in (1), Assumption 1 and the data without introducing auxiliary assumptions. The literature thus far only handled special cases of the above setting under additional assumptions or proposed outer sets, i.e. supersets of the sharp identified set. In particular, C. Manski and Tamer (2002) derive the sharp identified set for the regression problem for interval Y or X under the auxiliary assumption of mean independence.[2] Whether such an assumption is reasonable is an application specific question. Our results are complementary as they allow to compare the identified set under mean independence with the worst-case set of parameters without such an assumption. For the special case of missing data, Aucejo, Bugni, and Hotz (2017) point out that the mean independence assumption in C. Manski and Tamer (2002) is implying that the data is missing at random in their setting. Without auxiliary assumptions characterizing the sharp identified set in an operationalizable way is considerably more complex. Thus, Aucejo, Bugni, and Hotz (2017) resort to derive outer sets for their problem of missing covariates and outcomes. Below we will show that one of their outer sets is a special case of our identification theorem. Magnac and Maurin (2008) instead focus on a special case of binary choice models. Their results rely on a different set of assumptions.

Our identification result is based on the observation that the model in (1) implies infinitely many unconditional moment equalities. A natural idea is thus to expand the ideas in Beresteanu, I. Molchanov, and Molinari (2011), which characterizes the sharp identified set for the latter problem by constructing a finite dimensional random set. For the problem at

[2]Asher, Novosad, and Rafkin (2018) build on the results in C. Manski and Tamer (2002) tighting the identified sets by imposing stronger assumptions.

hand this requires setting up an infinite dimensional random set which's (Aumann) expectation can be traced out by using its support function. Hence, a sequence of directional vectors defines an outer set, converging to the true identified set in the limit. In other words, for each candidate parameter value, we replace the task of finding a proper probability measure which jointly fullfills the constraints imposed by model (1) and Assumption 1 , with a search over directional vectors. Formally, this defines an infinite collection of moment inequalities indexed by the latter vector. This structure allows us to prove consistency and construct uniformly valid confidence sets for the true parameter relying on the literature on moment inequality models (see for instance, Chernozhukov, Hong, and Tamer (2007), D. Andrews and Shi (2013), D. Andrews and Shi (2017), etc.).

The approach of this paper is following the worst-case and partial identification literature pioneered by Manski in a series of monographs and papers (C. F. Manski (2005), C. F. Manski (2008), etc) based on earlier contributions by Marschak and W. Andrews (1944) and Frechet (1951). For more recent reviews of the literature see Molinari (2020), Kline and Tamer (2023) or Canay and Shaikh (2017) for inference. However, it should be noted that there are also point identification approaches relying on stronger assumptions available in the literature. The contributions of Hsiao (1983) and Rigobon and Stoker (2009) highlight the potential pitfalls of commonly applied solutions to the problem: First, model (1) is frequently modified by conditioning on the observed intervals themselves and thus conducting inference on $\mathbf{E}_{F_0}[m(Y, X, \theta_0)|[X_l, X_u]]$ instead of $\mathbf{E}_{F_0}[m(Y, X, \theta_0)|X]$. This appears sensible, but is not equivalent to the original model, might contain a loss of information and may be problematic when either the categories have overlapping values, or the regressors are continuous.[3] Second, Hsiao (1983) discusses the implications of replacing the missing or interval valued observations with the midpoint of the observed intervals and Rigobon and Stoker (2009) treat the case of taking censored regressors at face value. Both of which can be seen as a naive imputation rules

[3]Hsiao (1983) also discusses a potential advantage of binning a continuous variable into discrete intervals in the case where the regression model is misspecified. For instance, if the true regression model is non-linear, using discrete bins of the regressor may better capture the true relation than a linear model using the continous variable.

and rest on strong and untestable assumptions on the underlying data generating process. Even more sophisticated imputation approaches, for a recent example see Duan and Pelger (2025), rely on ignorability or missingness at random assumptions. We view our contribution as complementary to these approaches as our results provide an assumption-free[4] background for which alternative methods, based on stronger assumptions or imputation strategies, can be framed. Conclusions based on a particular set of assumptions might be deemed more credible if the parameter values within the sharp identified set do not contain values which substantially change the qualitative and quantitative conclusions of the imputation estimator. Similarly our results are complementary to approaches utilizing additional information in the form of a second sample, see for instance Pacini (2018). These two sample identification approaches can often be interpreted as imputation methods using external information.

As an application, we revisit Castillo, Jordan, and Petrie (2018), who use a multiple price list (MPL) to elicit individual discount rates. An MPL is a sequence of binary choice questions that inherently pins down the elicited quantity only up to an interval. We show that even seemingly small observed intervals can matter substantially for empirical conclusions: in a simple probit regression of high-school graduation on the elicited discount rate, the implied marginal effects are highly sensitive once the interval nature of the regressor is respected rather than collapsed to a point (e.g., by midpoint imputation). Accounting only for ambiguity induced by the question design, which pins down the discount rate to brackets of 0.0735 width, yields an estimate of the sharp identified set of the marginal effect of the discount rate on on-time high-school graduation ranging from -21% to $+6\%$. In contrast to the original estimate of -9%, this sharp set cannot rule out a positive correlation of the discount rate with high-school graduation. Further accounting for inconsistent choice pattern and missing outcome data enlarges the estimated sharp identified set even further. As MPLs are widely used in applied research (see e.g. a discussion in Yu, Zhang, and Zuo (2021), we believe our empirical results are of interest beyond the specific application and caution against ignoring

[4] i.e. an approach which does not impose auxiliary assumptions to shrink the sharp identified set beyond the assumptions implied by model and data itself: assumptions 1-3

the interval-measurement in such settings.

The paper is structured as follows: First, Section 2 presents the main identification result for the sharp identified set. Section 3 proceeds by connecting our identification results with the constructing in Aucejo, Bugni, and Hotz (2017), before Section 4 propose an estimator for the sharp set alongside uniformly valid confidence sets. Section 5 presents a preliminary simulation study. Section 6 contains the empirical illustration with regard to the MPL design. Section 7 concludes.

2 Identification

Let us first define the object of interest. To this end using the law of total probability we can rewrite the expectation in (1) as

$$\mathbf{E}[m(Y,X,\theta)|X] = \sum_{z\in\mathcal{Z}} \mathbf{E}[m(Y,X,\theta)|X,Z=z]P(Z=z|X)$$

where $\mathbf{E}[m(Y,X,\theta)|X,Z=z] = \int_{\mathcal{Y}} m(y,X,\theta)dP(y|X,Z=z)$. Let the parameter space be Θ. Under Assumption 1, the sharp identified set Θ_I^* is then given by:

$$\begin{aligned}
\Theta_I^* \equiv \big\{\theta\in\Theta : \exists \text{ regular conditional laws } & P_{Y|X,Z}(\cdot \mid x,z),\ P_{Z|X}(\cdot \mid x) : \\
& \sum_{z\in\mathcal{Z}} \underbrace{\int_{\mathcal{Y}} m(y,x,\theta)\, P_{Y|X,Z}(dy \mid x,z)}_{=\ \mathbb{E}[m(Y,x,\theta)|X=x,Z=z]}\ P_{Z|X}(z \mid x) = 0 \quad \text{for } P_X\text{-a.e. } x, \\
& \int_{\mathcal{Y}} P_{Y|X,Z}(dy \mid x,z) = 1, \quad P_{Y|X,Z}(A \mid x,z) \geq 0 \ \ \forall A\in\mathcal{A}_Y \quad \text{for } P_X\text{-a.e. } x \\
& \sum_{z\in\mathcal{Z}} P_{Z|X}(z \mid x) = 1, \quad P_{Z|X}(z \mid x) \geq 0 \ \text{ for } P_X\text{-a.e. } x \\
& P\big(Y\in[Y_l,Y_u],\ X\in[X_l,X_u] \mid Z=z\big) = 1 \ \text{ for all } z\in\mathcal{Z} \text{ with } P(Z=z)>0\big\}.
\end{aligned}$$

Aucejo, Bugni, and Hotz (2017) point out that computing Θ_I^* is a complicated task because it requires, for each candidate value $\theta\in\Theta$, proving the existence or lack of existence

of a set of functions - specifically, the unknown conditional laws $P_{Y|X,Z}(\cdot \mid x, z)$ and $P_{Z|X}(\cdot \mid x)$ for $z \neq 0$ - that satisfy the (potentially) uncountable number of restrictions.

The first step to derive an computable result is to get rid of the unobserved conditioning event in the model. To that end, the following Lemma gives an alternative characterization of the sharp identified set, utilizing the well-known transformation of conditional moment equalities into unconditional ones with the help of instrument functions:

Lemma 1. *Let Assumption 1 hold. Define*

$$\begin{aligned}\Theta_I \equiv \Big\{\theta \in \Theta : \exists \text{ regular joint laws } & P_{Y,X|Z=z}(\cdot,\cdot \mid z) \text{ on } (\mathcal{Y}\times\mathcal{X}, \mathcal{A}_Y \otimes \mathcal{A}_X) : \\ & \sum_{z\in\mathcal{Z}} \underbrace{\int_{\mathcal{Y}\times\mathcal{X}} m(y,x,\theta)\, g(x)\, P_{Y,X|Z=z}(dy,dx)}_{= \mathbb{E}[m(Y,X,\theta)\, g(X)|Z=z]} P(Z=z) \;=\; \boldsymbol{0} \quad \forall\, g \in \mathcal{G}, \\ & \int_{\mathcal{Y}\times\mathcal{X}} P_{Y,X|Z=z}(dy,dx) = 1, \qquad P_{Y,X|Z=z}(A) \geq 0 \;\; \forall A \in \mathcal{A}_Y \otimes \mathcal{A}_X, \\ & P\big(Y \in [Y_l, Y_u],\; X \in [X_l, X_u] \mid Z = z\big) = 1 \\ & \text{for all } z \in \mathcal{Z} \text{ with } P(Z=z) > 0\Big\}.\end{aligned}$$

Then $\Theta_I = \Theta_I^*$ provided the instrument class $\mathcal{G}$ is comprehensively revealing (definition 3.2 in Stinchombe and White (1998)). In particular the result holds with $\mathcal{G}_{cube}$ defined later. For each $g_k \in \mathcal{G}_{cube}$ it can be multiplied by a positive weight w_k defining the class $\mathcal{G}_{wc-cube}$ with no change of results. See from the above that w_k can be pulled out of the integral.

Proof. Follows directly from Billingsley (1995), see also Dominguez and Lobato (2004) or D. Andrews and Shi (2013). □

A class $\mathcal{G}$ being comprehensively revealing is requiring that the class of functions has identifying power against fixed alternatives. That is if $\tilde{\theta} \notin \Theta_I^*$ then $\exists g \in \mathcal{G}$ and $\exists P_{Y,X}$ where $E[m(Y,X,\tilde{\theta})g(X)] \neq 0$. Θ_I may also be written as $\Theta_I(\mathcal{G})$, to make this dependence explicit. Ignoring technical conditions and notation D. Andrews and Shi (2013) and D. Andrews and Shi (2017) call the condition of $\mathcal{G}$ being comprehensively revealing Assumption CI, standing

for "conditionally identified." The assumption implies that $\Theta_I^* = \Theta_I(\mathcal{G})$.

Lemma 1 postulates that the identification problem can equivalently be stated as proving or disproving the existence of unconditional joint distributions $P_{Y,X|Z=z}(\cdot,\cdot \mid z)$, as opposed to the conditional distributions directly from model (1). The result turns our finite number (d_m) of conditional moment conditions into an infinite collection of unconditional moment conditions. This characterization of Θ_I, using instrument functions and unconditional moment conditions, allows us to apply the tools of random set theory, which was introduce to econometrics in the seminal work of Beresteanu and Molinari (2008). Excellent textbook treatments are Molchanov (2017) and I. Molchanov and Molinari (2018). For a review see Beresteanu and F. Molchanov I. M. (2012) and Molinari (2020).

Random sets are a generalization of random variables and can be seen as *bundles* of random variables (Aumann (1965)). Our identification setting is similar to that of Beresteanu, I. Molchanov, and Molinari (2011) in that we use a support function approach to trace out the (Aumann) expectation of a properly defined random closed set, which is defined by the moment conditions and the observed intervals. Relative to Beresteanu, I. Molchanov, and Molinari (2011), our main technical contribution is to expand their results to accommodate the infinite number of moment restrictions given in Lemma 1 and handling the discontinuous mapping from the observed intervals to the moment conditions. The latter emerges from the particular instrument class we use to transform conditional into unconditional moments.

Before presenting the main identification theorem, it is useful to discuss the intuition behind our approach and that of Beresteanu, I. Molchanov, and Molinari (2011). The closed intervals $[Y_l, Y_u]$ and $[X_l, X_u]$ can be viewed as random closed sets (Beresteanu and Molinari (2008)). Such sets are characterized by their selections, i.e., all random variables that are stochastically dominated by these intervals. We construct a new θ-dependent random set Q_θ, defined below, containing the infinitely many unconditional moment conditions stated in Lemma 1. Q_θ is built such that $\theta \in \Theta_I \Leftrightarrow \mathbf{0} \in \mathbb{E}_A[Q_\theta]$, where $\mathbb{E}_A[\cdot]$ refers to the Aumann expectation. The Aumann expectation of a random set is defined as the closure of the set

of expectations of all integrable random variable q of that set Q_θ. Hence, if zero belongs to this expectation, it follows that there exist random variables $(\tilde{Y}, \tilde{X})$ with joint distribution $P_{\tilde{Y},\tilde{X}|Z=z}(\cdot, \cdot \mid z)$ that satisfy the conditions in Lemma 1. In other words, establishing whether a valid joint law exists that fulfills the moment restrictions reduces to checking whether the (infinite dimensional) zero vector is contained in the Aumann expectation of a (infinite dimensional) random set. The Aumann expectation can also be characterized by its support function, yielding a variety of computationally tractable results.

Our results require the following set of regularity conditions ensuring that the objects we are working with are well behaved.

Assumption 2. *The parameter space Θ is a nonempty compact subset of $\mathbb{R}^d$.*

Assumption 3. *$m(y, x, \theta)$ is bounded, i.e. $\exists C_1 : \ |m(y, x, \theta)| \leq C_1$ and $m(y, x, \theta)$ is continuous in θ.*

Assumption 2 holds if and only if Θ is bounded, even if bounded by a very number. Assumption 3 is fulfilled by common examples like for instance linear mean regressions with bounded covariates, logit and probit models or quantile regressions.

The following Theorem formalizes the above intuition. Let $\text{Sel}^1(A)$ denotes the set of measurable selections of the random set A and $h(A, u) \equiv \sup_{a \in A} u'a$ be its support function.

Theorem 1. *Let Assumption 1-3 hold, then*

$$\Theta_I = \{\theta \in \Theta : \inf_{u \in B_1^\infty} \sum_{z \in \mathcal{Z}} \mathbf{E}[h(Q_{\theta,z}, u)|Z = z]P(Z = z) \geq 0\}$$

where we define the infinite dimensional random sets as

$$Q_{\theta,z} \equiv \left\{ \begin{bmatrix} m(y, x, \theta)\, g_1(x) \\ m(y, x, \theta)\, g_2(x) \\ \vdots \end{bmatrix} : \ (y, x) \in \text{Sel}^1\big([Y_l, Y_u] \times [X_l, X_u] \mid Z = z\big) \right\}.$$

with $B_1^\infty \equiv \{u \in \ell^\infty : \ ||u||_\infty \leq 1\}$ *and* $g \in \mathcal{G}^{wc,cube}$ *as defined below.*

Proof. See Appendix A. □

The class of weighted cubes is constructed based on the class of countable cubes in D. Andrews and Shi (2017). The following defines our class of cubes with side lengths that are $(2r)^{-1}$ for all large positive integers r. The cubes partition $[0,1]^{d_x}$ for each r:

$$\mathcal{G}_{\text{wc-cube}} \equiv \{g_{a,r} : g_{a,r}(x) = w(r)1\{x \in C_{a,r}\} \quad \text{for } C_{a,r} \in \mathcal{C}_{\text{c-cube}}\},$$

where $w(r) = (100 + r^\beta)^{-1}$ with $\beta > 1$ and

$$\begin{aligned}\mathcal{C}_{\text{c-cube}} \equiv \{C_{a,r} = &\prod_{j=1}^{d_x}((a_j - 1)/(2r),\, a_j/(2r)] \in [0,1]^{d_x} : \\ &a = (a_1, \ldots, a_{d_x})',\ a_j \in \{1, 2, \ldots, 2r\} \\ &\text{for } j = 1, \ldots, d_x \text{ and } r = r_0, r_0 + 1, \ldots\}\end{aligned}$$

Choosing weights this way enables us to show that the ambient space of the random set is ℓ_1. An unweighted class of $\mathcal{G}$ would imply an ambient spaceof ℓ_∞ which greatly complicates proof. The dual space of ℓ_1 is ℓ_∞. Thus justifying the search over random directions in B_1^∞. If we replace $g \in \mathcal{G}$ with $\mathcal{G} \neq \mathcal{G}^{w,cube}$ provided that instrument class $\mathcal{G}$ is still *comprehensively revealing* (definition 3.2 in Stinchombe and White (1998)).

Our characterization of Θ_I through the support function $h(\mathbb{E}_A[Q_\theta], u)$ is convenient computationally, as for any particular u the optimization problem is convex and can often be solved quickly using gradient methods. Intuitively, each u is defining a direction in an infinite space. Searching over all $u \in B_1^\infty$ and ensuring a positive support function is equivalent to assessing if the $\min_{u \in B_1^\infty} h(\mathbb{E}_A[Q_\theta], u) \geq 0$. In general this problem is nonconvex - since u and the selection $(x, y) \in Sel^1(X \times Y)$ interact in complicated ways - so the optimization problem implied by Theorem 1 is still computationally difficult. Nevertheless, for estimating the identified set it is sufficient to search for counterexamples to exclude values. Thus as soon

as one direction violates the moment condition for a particular parameter set, we can exclude it and move to the next candidate. We show in application and the simulation below that often even a moderate number of random directions is already approximating the identified set quite well. Further, simplifications may be possible for monotone index models like linear regressions or probit regressions.

It may be instructive to see how this works for a simple linear model defined by the moment condition $E[y - \beta x | X = x] = 0 \; \forall x \in \mathcal{X}$. The support function given an arbitrary direction u and a candidate parameter value θ is defined in this case as:

$$h(Q_\theta, u) = \max_{x \in [X_L, X_U], y \in [Y_L, Y_U]} \sum_{k=1}^{\infty} u_k (y - \beta x) g_k(x)$$

Finding a maximum $(x, y) \in Sel^1(X \times Y)$ is not difficult for a fixed u vector as this function is piecewise linear. The potential solutions are the partitions of the space defined by the instrument class within the feasible range defined by the observed interval $[X_L, X_U]$ and the interval boundaries itself for x. For y only Y_L and Y_U need to be checked depending on the sign of u and the candidate parameter value.

Theorem 1 provides the sharp identified set in the limit and is conceptually simple. We provide an associate (Hausdorff)-consistent estimator for it in the following section. However, let us first discuss several special cases arising from our result which yield outer sets instead. An outer set of the identified set is any $\tilde{\Theta} \supseteq \Theta_I$. Proving consistency and inference of such sets may be much simpler than the analogous tasks for the sharp identified set. We define two classes of these outer sets, $\Theta_O(\mathcal{U}) \supseteq \Theta_I$ and $\Theta_O^K(\mathcal{U}) \supseteq \Theta_I$ and show how these can be constructed and tightened gradually at only the cost of computational resources.

First, note that Theorem 1 defines an infinite collection of inequalities index by the directional vector $u \in B_1^\infty$. Any subset of those inequalities, i.e. replacing B_1^∞ with any

$\mathcal{U} \subset B_1^\infty$, defines an outer set, which we denote by $\Theta_O(\mathcal{U})$:

$$\Theta_O(\mathcal{U}) \equiv \Big\{ \theta : \mathbb{E}\big[h(Q_\theta^\infty, u)\big] \geq 0 \ \ \forall\, u \in \mathcal{U} \Big\},$$

Similarly, because $Q_\theta^\infty \subset \mathbb{R}^\infty$ is computationally unwieldy, a researcher may only look at $Q_\theta^K \subset \mathbb{R}^K$, i.e. the restriction to the first K coordinates. D. Andrews and Shi (2013) call this a truncated problem. We define the finite-dimensional outer set $\Theta_O^K(\mathcal{U})$, where $P_K(\mathcal{U})$ projects $\mathcal{U}$ onto the first K coordinates so that the support function is well-defined:

$$\Theta_O^K(\mathcal{U}) \equiv \Big\{ \theta : \mathbb{E}\big[h(Q_\theta^K, u)\big] \geq 0 \ \ \forall\, u \in P_K(\mathcal{U}) \Big\}.$$

The following is an immediate consequence of Theorem 1:

Corollary 1. *Let $\mathcal{U} \subseteq B_1^\infty$. If $\theta \in \Theta_I$, then $\theta \in \Theta_O(\mathcal{U})$. If $\theta \in \Theta_O(\mathcal{U})$, then $\theta \in \Theta_O^K(\mathcal{U})$. Consequently, $\Theta_I \subseteq \Theta_O(\mathcal{U}) \subseteq \Theta_O^K(\mathcal{U})$.*

Proof. We have $\theta \in \Theta_I \iff E[h(Q_\theta^\infty, u)] \geq 0 \quad \forall\, u \in B_1^\infty$, in particular for all $u \in \mathcal{U}$, thus $\Theta_I \subseteq \Theta_O(\mathcal{U})$. For the second implication, write $P_K(\mathcal{U})$ for the projection of $\mathcal{U}$ onto its first K coordinates and, for $\tilde{u} \in P_K(\mathcal{U})$, embed by $u = (\tilde{u}, 0, \dots)$. Then

$$\begin{aligned}
\theta \in \Theta_O(\mathcal{U}) &\iff E[h(Q_\theta^\infty, u)] \geq 0 \quad \forall\, u \in \mathcal{U} \\
&\Rightarrow E[h(Q_\theta^\infty, (\tilde{u}, 0, \dots))] \geq 0 \quad \forall\, \tilde{u} \in P_K(\mathcal{U}) \\
&\iff E\big[h\big(Q_\theta^K, \tilde{u}\big)\big] \geq 0 \quad \forall\, \tilde{u} \in P_K(\mathcal{U}) \\
&\iff \theta \in \Theta_O^K(\mathcal{U}),
\end{aligned}$$

. □

The implication is that outer sets can be constructed *loosely* with minimal computational power using a small $\mathcal{U}$ that can be handled in a short period of time, which can grow progressively more complex. Lipschitz continuity of the support function with respect to

u (given Q_θ is bounded), can be used to construct grid $\mathcal{U}_\varepsilon$ over B_1^∞ which will guarantee $|min_{u\in B_1} h(Q_\theta^\infty, u) - min_{u\in\mathcal{U}_\varepsilon} h(Q_\theta^\infty, u)| < \varepsilon$ (Rockafellar and Wets (2009)). Still this grid is also infinite dimensional and thus cannot be searched over. Theoretical guarantees are more tractable in finite dimensions. This is similar to D. Andrews and Shi (2013) who present the main theorems over an infinite dimensional space, but in their appendix justifies a use of a truncated number of instruments, noting that these are outer sets and thus coverage guarantees will hold. Asymptotically, the estimator we propose in Section 4 converges to the identified set by using $K(n) \to \infty$ at an appropriate rate as $n \to \infty$.

3 Relation to Aucejo, Bugni, and Hotz (2017)

Aucejo, Bugni, and Hotz (2017) are concerned with the same conditional moment model defined by equation (1) and Assumption 1 above. However, their focus is on the missing data problem, i.e. $\mathcal{Z} = \{0, 1, 2, 3\}$ but $[Y_l, Y_u] = \mathcal{Y}$ and $[X_l, X_u] = \mathcal{X}$. Nevertheless, comparing our results with their construction helps developing further intuition for the identification problem, thus in this section we discuss both approaches and their relation in detail. For simplicity, let us focus on the one moment case, thus $d_m = 1$, although both their paper and the present work can deal with the more general case. Similarly, let us only cover the missing covariate case, i.e. $\mathcal{Z} = \{0, 1\}$ and abstract from fully observed additional covariates.

As mentioned above, Aucejo, Bugni, and Hotz (2017) note that the sharp identified set, Θ_I^*, is in general a too complicated object to compute and thus they resort to constructing outer sets, i.e. supersets of the sharp identified set. They propose two outer sets.

The first step in Aucejo, Bugni, and Hotz (2017) transforms the conditional moment conditions into unconditional moments via instrument functions as we propose for our problem in Lemma 1. They use the class of boxes $\mathcal{G}_{\text{box}} \equiv \{\mathbb{I}[X \in G(x, v)] : \forall (x, v) \in \mathcal{X} \times \mathbb{R}_{++}^{d_x}\}$ with $G(x, v) \equiv \{\tilde{x} \in \mathcal{X} : \{x_j - v_j \leq \tilde{x}_j \leq x_j + v_j\}_{j=1}^{d_x}\}$ as instrument functions.[5] Let us denote the

[5]Note that in the definition of $B(x, v)$ we use $\tilde{x} \in \mathcal{X}$ directly, thus $B(x, v) \cap \mathcal{X} \neq \emptyset$ is always true and we can avoid carrying the intersection in the definition of M_k below.

collection of boxes by $\{G_k\}_{k\geq 0} \equiv \{G(x,v) \quad \forall (x,v) \in \mathcal{X} \times \mathbb{R}^{d_x}_{++}\}$. Note that the main point of the following conceptual comparison does not rely on the specific instrument class.

The first outer identified set in Aucejo, Bugni, and Hotz (2017) uses the following vector-valued auxiliary functions for each member of $\{G_k\}_{k\geq 0}$:

$$M_{k,1}(W^O) \equiv \begin{pmatrix} -\left(\inf_{y\in\mathcal{Y}, x\in G_k} m(y,x,\theta)\mathbb{I}[Z=1] + m(Y,X,\theta)\mathbb{I}[Z=0, X\in G_k]\right) \\ +\left(\sup_{y\in\mathcal{Y}, x\in G_k} m(y,x,\theta)\mathbb{I}[Z=1] + m(Y,X,\theta)\mathbb{I}[Z=0, X\in G_k]\right) \end{pmatrix}$$

They proceed by defining their first outer set as $\Theta_{I,1}^{ABH} \equiv \{\theta \in \Theta : \ \mathbf{E}[M_{k,1}(W^O)] \geq 0 \quad \forall k \geq 0\}$. This solves the moment problem for each box, or in general, each subspace defined by the instrument functions, but not simultaneously for all of them. As Aucejo, Bugni, and Hotz (2017) mention, the difficult part of characterizing the sharp identified set is to ensure that the moment conditions hold *simultaneously* for all boxes and not only individually, thus Θ_I^{ABH} is an outer set.

The second outer set defined in Aucejo, Bugni, and Hotz (2017) is based on moments integrating out the missing covariate. Their auxiliary moment conditions do not depend on k in the case of no additional fully observed covariates and are given by:

$$M_{k,2}(W^O) \equiv \begin{pmatrix} -\left(\inf_{x\in\mathcal{X}} m(Y,x,\theta)\mathbb{I}[Z=1] + m(Y,X,\theta)\mathbb{I}[Z=0]\right) \\ +\left(\sup_{x\in\mathcal{X}} m(Y,x,\theta)\mathbb{I}[Z=1] + m(Y,X,\theta)\mathbb{I}[Z=0]\right) \end{pmatrix}$$

The second outer identified set is the collection $\Theta_{I,2}^{ABH} \equiv \{\theta \in \Theta : \ \mathbf{E}[M_{k,2}(W^O)] \geq 0\}$.

Let us now show the relation between $\Theta_{I,1}^{ABH}$, $\Theta_{I,2}^{ABH}$ and our sharp identified set Θ_I. First note that the class of boxes is comprehensively revealing (see D. Andrews and Shi (2013)), thus the theorem applies, however, potentially looses sharpness. Further, by Corollary 1, we know that any subset $\mathcal{U} \subseteq B_1^\infty$ is defining an outer set. In particular, choose $\mathcal{U}_{point} := \{\pm e_j : j \geq 1\}$, where $e_j := \{0, .., 1, 0, \dots\}$ is the jth dimensional basis vector. The outer set implied by this

choice is

$$\Theta_O(\mathcal{U}_{point}) = \{\theta \in \Theta : \sum_{z \in \mathcal{Z}} \mathbf{E}[h(Q_{\theta,z}, u)|Z = z]P(Z = z) \geq 0 \quad \forall u \in \mathcal{U}_{point}\}$$

For the special case of $\mathcal{Z} = \{0, 1\}$ we have

$$\sum_{z \in \mathcal{Z}} \mathbf{E}[h(Q_{\theta,z}, u)|Z = z]P(Z = z) = \mathbf{E}[h(Q_{\theta,0}, u)\mathbb{I}[Z = 0]] + \mathbf{E}[h(Q_{\theta,1}, u)\mathbb{I}[Z = 1]]$$

Further, considering the class of boxes, the support function for the missing covariate case only is given by $h(Q_{\theta,z}, u) = \sup_{x \in [X_l, X_u]} \sum_{k \geq 0} u_k m(Y, x, \theta)\mathbb{I}[x \in G_k]$. Any choice of $\tilde{u} \in \mathcal{U}_{point}$ simplifies the support function considerably. In particular, chosing $\tilde{u} = e_k$ yields $h(Q_{\theta,z}, \tilde{u}) = \sup_{x \in [X_l, X_u]} m(Y, x, \theta)\mathbb{I}[x \in G_k]$. For the cases for which we fully observe the covariates, $Z = 0$, this further collapses to $h(Q_{\theta,0}, \tilde{u}) = m(Y, X, \theta)\mathbb{I}[X \in G_k]$ with $X = X_l = X_u$. Thus, the condition summed over missingness patterns for the specific choice e_k is

$$\begin{aligned} &\mathbf{E}[\sup_{x \in [X_l, X_u]} m(Y, x, \theta)\mathbb{I}[x \in G_k]\mathbb{I}[Z = 1]] + \mathbf{E}[m(Y, X, \theta)\mathbb{I}[X \in G_k]\mathbb{I}[Z = 0]] \geq 0 \\ &\iff \mathbf{E}\left[\sup_{x \in \mathcal{X}} m(Y, x, \theta)\mathbb{I}[x \in G_k]\mathbb{I}[Z = 1] + m(Y, X, \theta)\mathbb{I}[Z = 0, X \in G_k]\right] \geq 0 \end{aligned}$$

Where we used $[X_l, X_u] = \mathcal{X}$ for the missing data case. Simultaneously, using $\sup_{a \in A} f(a) = -\inf_{a \in A}(-f(a))$, and considering $\tilde{u} = -e_k$ we obtain:

$$\mathbf{E}\left[-\left(\inf_{x \in \mathcal{X}} m(Y, x, \theta)\mathbb{I}[x \in G_k]\mathbb{I}[Z = 1] + m(Y, X, \theta)\mathbb{I}[Z = 0, X \in G_k]\right)\right] \geq 0$$

These expressions are almost equivalent to the elements of the vector $M_{k,1}$ defined above, however whereas we take the supremum over $x \in \mathcal{X}$, Aucejo, Bugni, and Hotz (2017) are considering $y \in \mathcal{Y}, x \in G_k$. Additionally, our result includes the indicator $\mathbb{I}[x \in G_k]$ in the objective for the missing observations. Notice that taking the supremum over the entire support of Y even for observed values, is not using the entire information provided from the

data. For instance, assume that all observations with missing X, have a constant value $Y = c$ for some constant c. In a linear model the only slope rationalizing this data is $\theta = 0$, yet for unbounded $\mathcal{X}$, $\Theta_{I,1}^{ABH}$ would contain the entire real line. Although, this is still an outer set, using the observed values of Y, as it is done in our construction, does shrink the identified set and decrease the computational burden of the approach.

The second difference is that instead of maximizing over the full support as in Theorem 1 above, Aucejo, Bugni, and Hotz (2017) maximize over the individual boxes. In combination with also varying Y for the missing X observations this still yields an outer set. However, conceptually it is assuming that for each box, the value of the missing covariate must lie within the specific box. This seems to be stronger than needed. The model in (1) is only implying that the moment condition has to hold over all subspaces, not that it must hold under all subspaces assuming that all realized values are within that space. Instead, our construction maximizes $m(Y, x, \theta)\mathbb{I}[x \in G_k]$ over $x \in \mathcal{X}$, thus it allows for a specific observation not contributing to G_k.[6]

Taking all of the above into account, it appears that there is no clear ranking between $\Theta_{I,1}^{ABH}$ and $\Theta_O(\mathcal{U}_{point})$ for the general case. However, the main take-away is that our identification results considers arbitrary combination of boxes. As mentioned, it is the relation between boxes and the requirement of the moment condition to hold simultaneously for all boxes that renders the sharp identification set hard to compute. Our support function approach facilitates this and provides a sequence of outer sets which in the limit converge to the sharp identified set. So although, neither $\Theta_{I,1}^{ABH} \subseteq \Theta_O(\mathcal{U}_{point})$ nor $\Theta_{I,1}^{ABH} \supseteq \Theta_O(\mathcal{U}_{point})$ for this specific directions $\mathcal{U}_{point}$, considering further directions is guaranteed to yield a smaller set than $\Theta_{I,1}^{ABH}$ in the limit.

[6]Another way to see this point is to consider a simplier setting. Suppose $m(y, x, \theta) = x$ with $\mathcal{X} = [0, 1]$, thus the moment condition boils down to $\mathbb{E}[X|X] = 0$. If X is missing for all observations we cannot rule out that $P_x = \delta_0$, a mass point at zero. The latter is the only distribution of X fullfilling the moment condition. For any box in $[0, 1]$ and u vector a feasible solution to our support function approach is zero as we can either choose $x = 0$ or another value with associated $u = 0$. Instead, considering any box $[a, b]$ with $a \neq 0$, the first moment in Aucejo, Bugni, and Hotz (2017) first outer set is negative, thus violated. We thank Federico Bugni for bringing up this example.

For $\Theta_{I,2}^{ABH}$, now we will show that there is an equivalence result using a specific subvector between our result and Aucejo, Bugni, and Hotz (2017) in a special case. Consider $u : u_k = \bar{u}$ for all k. It follows that

$$h(Q_{\theta,1}, u) = \bar{u} \sup_{x \in \mathcal{X}} m(Y, x, \theta) n(x) \qquad \text{with} \qquad n(x) \equiv \sum_{k \geq 1} \mathbb{I}[x \in G_k]$$

and similarly $h(Q_{\theta,0}, u) = \bar{u}\ m(Y, X, \theta)\ n(X)$.

$$\begin{aligned} &\mathbf{E}[h(Q_{\theta,0}, u)\mathbb{I}[Z = 0]] + \mathbf{E}[h(Q_{\theta,1}, u)\mathbb{I}[Z = 1]] \\ &= \mathbf{E}[\bar{u} \sup_{x \in \mathcal{X}} m(Y, x, \theta) n(x) \mathbb{I}[Z = 1]] + \mathbf{E}[\bar{u}\ m(Y, X, \theta)\ n(X)\mathbb{I}[Z = 0]] \end{aligned}$$

For a disjoint partition of the space $n(x) = N(X) = 1$, i.e. if we consider the class of cubes defined below for a fixed resolution, the implied moment condition is

$$\mathbf{E}[\sup_{x \in \mathcal{X}} m(Y, x, \theta)\mathbb{I}[Z = 1] + m(Y, X, \theta)\mathbb{I}[Z = 0]] \geq 0$$

which is exactly the moment condition used to construct $\Theta_{I,2}^{ABH}$. Thus, for instrument classes partitioning the underlying space, $\Theta_{I,2}^{ABH}$ can be viewed as a special case of Theorem 1 only considering subdirections.

To further built some intuition, consider a simple example with, $Y = \theta X + \epsilon$, $\mathbf{E}[\epsilon|X] = 0$, $\mathcal{X} = [\underline{x}, \bar{x}]$ and $\mathcal{Y} = [\underline{y}, \bar{y}]$. For simplicity assume $0 < \underline{x} \leq \bar{x}$ and $0 < \underline{y} \leq \bar{y}$ so all variables have positive support. Assume that all X values are missing and denote the mean of the observed Y by μ_y. Before studying the differences between the outer sets and our construction for this case, notice that the bounded support already shrinks the set of feasible parameter values θ. For $\mathbf{E}[\epsilon|X] = 0$ to be feasible we need

$$0 \in [\underline{y} - \theta\bar{x}, \bar{y} - \theta\underline{x}] \quad \iff \quad \underline{y} - \theta\bar{x} < 0\ ,\ \bar{y} - \theta\underline{x} > 0 \quad \iff \quad \theta \in [\frac{\underline{y}}{\bar{x}}, \frac{\bar{y}}{\underline{x}}]$$

This is not the sharp identified set as it only uses the model assumptions, not the data information, so we will denote the set of values by $\Theta_{Assumption} = [\frac{\underline{y}}{\bar{x}}, \frac{\bar{y}}{\underline{x}}]$.

Being able to learn the mean of Y gives us the sharp identified set $\Theta_I = [\frac{\mu_y}{\bar{x}}, \frac{\mu_y}{\underline{x}}]$. In this case the moment function of the first outer set of Aucejo, Bugni, and Hotz (2017) becomes:

$$M_{k,1} = \begin{pmatrix} -\inf\limits_{y\in\mathcal{Y}, x\in G_k}(y - x\theta) \\ +\sup\limits_{y\in\mathcal{Y}, x\in G_k}(y - x\theta) \end{pmatrix} = \begin{pmatrix} -(\underline{y} - b_k\theta) \\ +(\bar{y} - a_k\theta) \end{pmatrix}$$

where we use $G_k = [a_k, b_k]$ for all $a_k, b_k \in \mathcal{X}$ as instrument functions. The condition $\mathbf{E}[M_{k,1}] \geq 0$ implies $\theta \in \bigcap\limits_k [\frac{\underline{y}}{b_k}, \frac{\bar{y}}{a_k}] = \Theta_{I,1}^{ABH} = \Theta_{Assumption}$. Because this outer set is not using the sample information of the outcome for the missing data, instead of the sharp identified set it only gives back all logical feasible values implied the model and support assumptions. Note that above implicitly assumed $\theta > 0$, it is straightforward to show that all $\theta < 0$ are violating the moment conditions.

The second outer set of Aucejo, Bugni, and Hotz (2017) instead uses the information in the observed Y. The respective moment conditions are

$$M_{k,2} = \begin{pmatrix} -\inf\limits_{x\in\mathcal{X}}(Y - x\theta) \\ +\sup\limits_{x\in\mathcal{X}}(Y - x\theta) \end{pmatrix} = \begin{pmatrix} -(Y - \bar{x}\theta) \\ +(Y - \underline{x}\theta) \end{pmatrix}$$

Taking expectations yields the inequalities $\theta \geq \frac{\mu_y}{\bar{x}}$ and $\theta \leq \frac{\mu_y}{\underline{x}}$, thus the sharp identified set, i.e. $\Theta_{I,2}^{ABH} = \Theta_I$. It follows immediately that our identified set is equal to $\Theta_{I,2}^{ABH} = \Theta_I$ in this case by Theorem 1.

We have seen above that $\Theta_{I,1}^{ABH}$ is not using the information about the location of Y and is thus not utilizing all of the available information. A similar loss of information can be demonstrated for $\Theta_{I,1}^{ABH}$ by enriching the setting slightly. We still assume that all X values are missing but divide the support of X into two distinct, obviously unobserved, boxes $G_l = [\underline{x}, a)$ and $G_u = [a, \bar{x}]$, where the share of the population within each box is denoted by q and $1 - q$.

Take a $\theta' \in \Theta_{I,2}^{ABH}$, thus we know that $-(\mu_y - \bar{x}\theta) \geq 0$ and $(\mu_y - \underline{x}\theta) \geq 0$ from before. Suppose the data is generated such that the support of Y depends on the X, i.e. $supp(Y|G_i) = [\underline{y}_i, \bar{y}_i]$. We assume for illustration

$$y - \bar{x}\theta' < y - \underline{x}\theta' < 0 \qquad \forall\ y \in [\underline{y}_l, \bar{y}_l]$$
$$y - \underline{x}\theta' > y - \bar{x}\theta' > 0 \qquad \forall\ y \in [\underline{y}_u, \bar{y}_u]$$

In particular this implies

$$\underline{\mu}_y - \bar{x}\theta' < \underline{\mu}_y - \underline{x}\theta' < 0$$
$$\bar{\mu}_y - \underline{x}\theta' > \bar{\mu}_y - \bar{x}\theta' > 0$$

where $\underline{\mu}_y = \mathbf{E}[Y|x \in G_l]$ and $\bar{\mu}_y = \mathbf{E}[Y|x \in G_u]$. Under these choices there exists a q satisfying $\mathbf{E}[M_{k,2}] \geq 0$ for θ', thus there is no contradiction with $\theta' \in \Theta_{I,2}^{ABH}$. Further, note that as long as the overall mean is unchanged $\mu_y = q\underline{\mu}_y + (1-q)\bar{\mu}_y$, the outer set $\Theta_{I,2}^{ABH}$ does not change.

Instead, consider the boxes $[\underline{x}, a), [a, \bar{x}]$, which are a subset of the instrument class of boxes we could use for Theorem 1. The support function conditions for θ' for these choices are

$$\mathbf{E}[\sup_{x \in \mathcal{X}} u_1(Y - \theta' x)\mathbf{1}\{x \in [\underline{x}, a)\} + u_2(Y - \theta' x)\mathbf{1}\{x \in [a, \bar{x}]\}] \geq 0 \qquad \forall u = (u_1, u_2)$$

by assumption the first summand in the supremum is equal to the direction u_1 times a negative number, whereas the second summand is equal to u_2 times a positive number. Thus, if we choose $u = (u_1, u_2) = (1, -1)$, the moment condition is violated and $\theta' \notin \Theta_I$. Finally note that $\Theta_{I,1}^{ABH}$ is also unchanged as although it uses the same boxes, it is not using the sample information on Y.

This illustrative toy example highlights the need to account for cross-box relations and how our infinite sequence of outer sets compares to the two outer sets proposed in Aucejo, Bugni,

and Hotz (2017) and tightens the identified set relative to their outer sets.

4 Estimator and Asymptotic Properties

The following theorem introduces an estimator associated to the identification results of Theorem 1 and prove its consistency.

Theorem 2. *Define the estimator*

$$\hat{\Theta}_I^{K(n)} = \{\theta \in \Theta : \min_{u \in B_1^{J(n)}} \sum_{z \in \mathcal{Z}} \frac{1}{n} \sum_{i=1}^{n} [h(Q_{\theta,z,i}, u)] \mathbb{I}\{Z_i = z\} \geq -\tau_n\}$$

where $B_1^{J(n)} \equiv \{u \in \mathbb{R}^{J(n)} : \|u\|_\infty \leq 1\}$ *and* $J(n) = \sum_{r=r_0}^{K(n)} (2r)^{d_x}$. *Under assumptions A1-A4 if* $K(n)/n \to 0$, $K(n) \to \infty$, $(\sqrt{\frac{n}{K(n)}}\tau_n \to \infty$ *and* $\tau_n \to 0$ *than* $\hat{\Theta}_I^{K(n)}$ *is consistent in Hausdorff metric for* Θ_I *, i.e.* $\rho_H(\hat{\Theta}_I^{K(n)}, \Theta_I) \xrightarrow{p} 0$. *Where* $K(n)$ *is the maximum resolution used in the class of weighted cubes* $\mathcal{G}_{wc\text{-}cube} \equiv \{g_{a,r} : g_{a,r}(x) = w(r) 1\{x \in C_{a,r}\} \quad for\ C_{a,r} \in \mathcal{C}_{c\text{-}cube}\}$ *where* $w(r) = (100 + r^\beta)^{-1}$ *with* $\beta > 1$ *and*

$$\begin{aligned} \mathcal{C}_{c\text{-}cube} \equiv \{C_{a,r} = &\prod_{j=1}^{d_x} ((a_j - 1)/(2r),\ a_j/(2r)] \in [0,1]^{d_x} : \\ &a = (a_1, \ldots, a_{d_x})',\ a_j \in \{1, 2, \ldots, 2r\} \\ &for\ j = 1, \ldots, d_x\ and\ r = r_0, r_0 + 1, \ldots, K(n)\} \end{aligned}$$

The rate conditions $K(n) \to \infty$ and $K(n)/n \to 0$ balance an approximation error of order $K(n)^{-\beta}$ from truncating the infinite system of moments against a sampling error of order $(K(n)/n)^{1/2}$ for the finite-dimensional random sets. The proof for Theorem 2 is given in Appendix B and relies on results in Chernozhukov, Hong, and Tamer (2007). The main complication in showing their conditions hold in our case is the growing dimension of the number and complexity of our moments with the sample size. The Appendix B introduces a law of large numbers for random sets for the latter case and several other necessary Lemmata

in order to show the result.

The class of weighted cubes $\mathcal{G}^{wc,cube}$ is a partition of the transformed space $[0,1]^{d_x}$ for each fixed resolution, e.g. with $d_x = 1$ and $r = 1$ we would have $\{(0, \frac{1}{2}], (\frac{1}{2}, 1]\}$. In order to use this class as it is, we transform the observed x values by considering $T(x) = \frac{x-\underline{x}}{\bar{x}-\underline{x}}$ with $\underline{x}$ and $\bar{x}$ being the boundaries of $\mathcal{X}$. This can be equivalently applied to observed intervals $[x_l, x_u]$, i.e. $[T(x_l), T(x_u)]$, to translate them to the unit interval.

Next, let us denote by $Q^{\mathcal{G}}_{\theta,z}$ the random sets constructed in Theorem 1 based on the instrument class $\mathcal{G}$, to be explicit on the dependence of the random sets on the latter. Being explicit about the individual indices as well we can define our moment function as

$$\tilde{m}(W_i, \theta, \tau) \equiv \sum_{z \in \mathcal{Z}} h(Q^{\mathcal{G}_{\text{wc-cube}}}_{\theta,z,i}, \tau) \mathbb{I}\{Z_i = z\}$$

Recall, the support function is given by:

$$h(Q_{\theta,z_i,i}(g), u) = \max_{y_i \in [\,]_i, x_i \in [\,]_i} \left(\sum_j m(y_i, x_i, \theta) g_j(x) u_j \right) = \max_{y_i \in [\,]_i, x_i \in [\,]_i} \left(m(y_i, x_i, \theta) \sum_j g_j(x) u_j \right)$$

Thus, for any $K(n)$ we can use $\tau(x) = \sum_{j=r_0}^{J(n)} u_k \cdot w(r) \mathbb{I}\{x \in C_{a,j}\}$, to define the class of step functions induced by the first $J(n)$ cube resolutions as

$$\mathcal{T}_n \equiv \left\{ \tau(x) = \sum_{j=1}^{J(n)} u_k \cdot w(r) \mathbb{I}\{x \in C_{a,j}\} : u \in \mathbb{R}^{J(n)},\ \|u\|_1 \leq 1 \right\}.$$

This collapses the two indices of the original moment into one which refers to the combined function class above, thus the moments are given by:

$$\mathbb{E}[\tilde{m}(W_i, \theta, \tau)] \geq 0 \qquad \forall \tau \in \mathcal{T}_n$$

which is the framework in D. Andrews and Shi (2017) except that the moments are uncon-

ditional thus we do not have to transform them again.[7] and crucially, the function class complexity is growing as $n \to \infty$. However, for any fixed $K(n)$ manageability of $\mathcal{T}_n$ holds trivially and thus the conservative coverage results in D. Andrews and Shi (2017) hold. Define the sample mean and variance as:

$$\bar{m}_n(W_i, \theta, \tau) \equiv \frac{1}{n} \sum_{i=1}^{n} \tilde{m}(W_i, \theta, \tau)$$
$$\Sigma_n(W_i, \theta, \tau) \equiv \frac{1}{n} \sum_{i=1}^{n} \left(\tilde{m}(W_i, \theta, \tau) - \bar{m}_n(W_i, \theta, \tau)\right)^2 + \epsilon\sigma_n^2(\theta)$$

Similarly, define the CvM test statistic as

$$T_n(\theta) \equiv \sup_{\tau \in \mathcal{T}_n} S(\sqrt{n}\bar{m}_n(W_i, \theta, \tau), \Sigma_n(W_i, \theta, \tau))$$

where $S(\cdot, \cdot)$ is one of the functions $S_1, \ldots, S_4$ defined in equation 3.9 of D. Andrews and Shi (2017). Let us denote by $\hat{c}_{n,1-\alpha}(\theta)$, the GMS critical value constructed as in D. Andrews and Shi (2017) and define the associated confidence set as

$$C_n(\alpha) \equiv \{\theta \in \Theta : T_n(\theta) \leq \hat{c}_{n,1-\alpha}(\theta)\}$$

With these definitions we can state the following uniform coverage result for our estimator.

Proposition 1. *Let $\mathcal{F}$ be the set of data generating processes F on W such that assumptions A1-A4 are fullfilled and $\mathbb{E}[\tilde{m}(W_i, \theta, \tau)] \geq 0 \ \forall \tau \in \mathcal{T}_n$ whenever $\theta \in \Theta_I(F)$. Then, for any $\alpha \in (0, 1)$, we have $\liminf_{n\to\infty} \inf_{F\in\mathcal{F}} \inf_{\theta\in\Theta_I(F)} P_F(\theta \in C_n) \geq 1 - \alpha$.*

Appendix D outlines the setting in more detail. The proof of the result is omitted as it follows directly from the managability arguments used in D. Andrews and Shi (2017) in Lemma 9.1 which corresponds to their Beresteanu, I. Molchanov, and Molinari (2011) example.

[7] see note in D. Andrews and Shi (2017) page 276, first paragraph.

5 Monte Carlo Simulation

In this section we present a preliminary simulation study focusing on the identification results. We simulate $n = 500$ and $n = 1000$ observations, respectively, from $y_i = x_i + \epsilon_i$, where $\epsilon_i \sim N(0, 1)$ and $x_i \sim U[0, 1]$. We generate interval data according to

$$x_{i,l} = x_i - \mathbf{1}\{\text{censored}_i\} \times \text{width} \times s_i \times pos,$$
$$x_{i,h} = x_i + \mathbf{1}\{\text{censored}_i\} \times \text{width} \times s_i \times (1 - pos),$$

with width $= 0.2$ (20% of the regressor support), $pos = 0.90$, and $s_i \sim U[0.5, 1.5]$. Thus, for censored observations the regressor is interval-measured with an average interval length of 0.2, and the true value tends to lie near the upper end of the interval.

We consider two censoring mechanisms. In the first design, censoring is random: each observation is censored independently with probability p. In the second design, censoring depends on the outcome: for the 50% smallest values of y, the censoring probability is $2p$, whereas for the remaining observations it is $p/2$. For the random design we set $p = 0.10$, and for the outcome-dependent design we set $p = 0.30$, implying a higher overall fraction of interval-measured data in the second design.

Figure 1 reports results averaged over 100 Monte Carlo replications. The blue region corresponds to our estimator of the sharp identified set, implemented using an extensive random search over directional vectors. The green region implements the same estimator, however, by only checking a few random directions. The potentially higher inaccuracy of the second rough random search comes with the benefit of reducing the computational burden, see Table 2 below. The comparison is included to understand how well a random grid search solves the minimization problem within our estimator. Finally, the red region implements an outer-set estimator that searches only over the extreme directions that are used in the approach of Aucejo, Bugni, and Hotz (2017). This is not identical to their original implementation, however can be viewed as an adaptation of their method to interval data.

For comparability, we use the same class of cubes as in our estimator rather than the boxes considered in Aucejo, Bugni, and Hotz (2017). The black asterisk and black plus symbol correspond to ordinary least squares estimates obtained by (i) dropping censored observations ("ignore") and (ii) imputing midpoints of the observed intervals, respectively. The four panels show the results from the sample of size 500 in the first row and the one for $n = 1000$ in the second. The first column displays results from the random censoring whereas the second considers the non-random censoring setting.

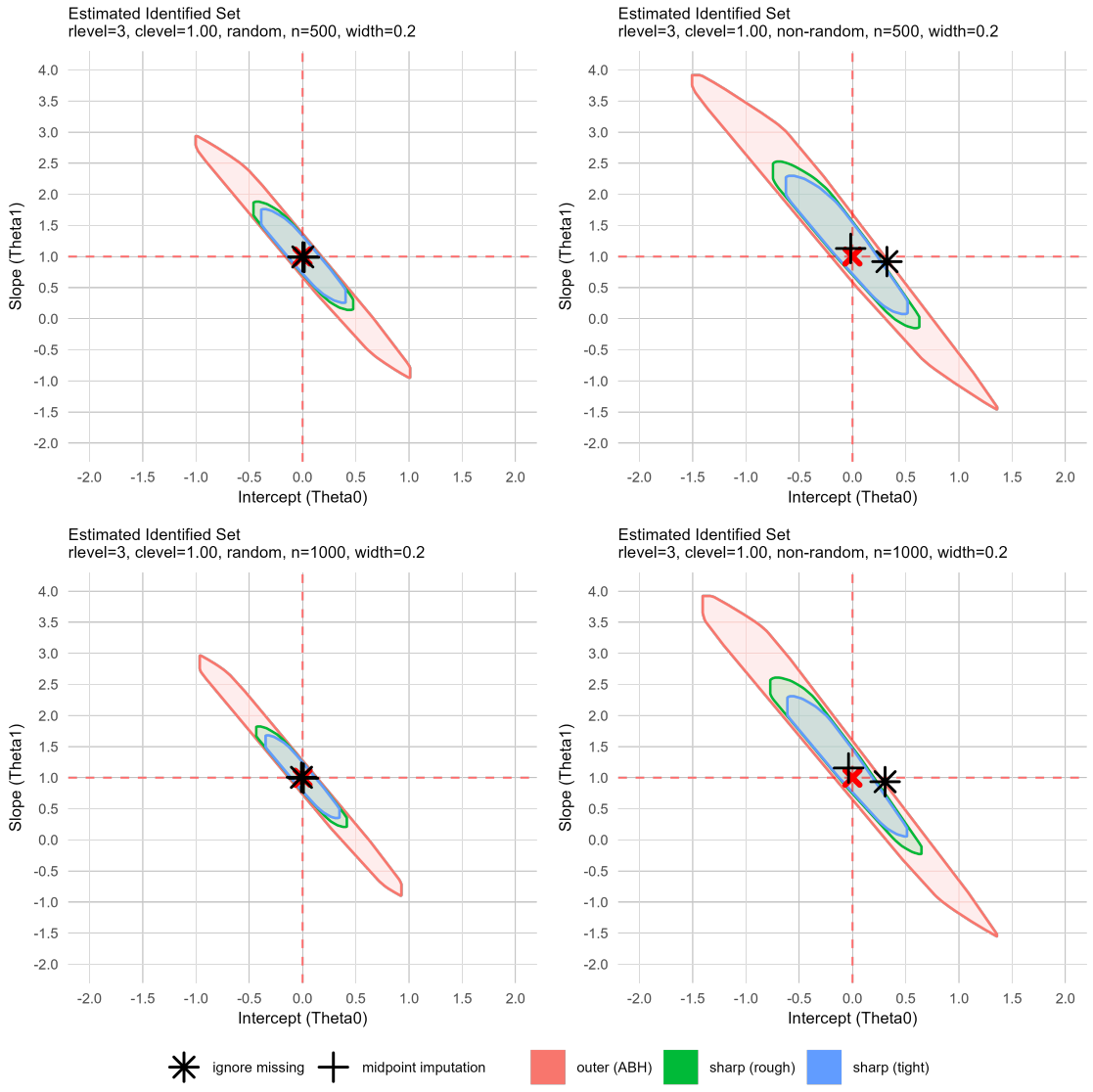


Figure 1: Average Estimated Sharp Sets

Overall, enforcing the moment restrictions jointly over the instrument class substantially

shrinks the estimated set relative to outer sets based only on extreme directions. As expected, outcome-dependent censoring leads to larger identified sets, reflecting the more severe loss of information. Under random censoring, both midpoint imputation and dropping censored observations perform reasonably well. This no longer holds when censoring depends on y: Ignoring the censored observations overestimates the intercept heavily. Note that this estimate lies outside of the identified set, which is possible as the identified uses more information in the form of the observed outcome data. Thus, other than the midpoint estimator, ignoring censored datapoints is not guaranteed to generate estimators which are strictly included in the identified set. This last point shows that our estimator does not necessarily weaken conclusions based on other approaches, as in some cases it allows researchers to use more available information in the data.

Table 2 summarizes the computational times in seconds of the current random search implementation of our estimator. For only a small fraction of missing data and smaller sample sizes estimation is performed under one minute. However, this computational time scales up with the number of censored observations. For a 1000 observations and a censored percentage of 30% the estimation takes more than 5 minutes. Although, these times are not an issue for the implementation of our set estimator, performing the bootstrap procedure described in Section 4, requires further development of our algorithms, which we currently revise for an efficient estimation of associated confidence sets.

Table 1: Computational Time in Seconds

Missingness	n	Cens. (\%)	outer (ABH)	sharp (rough)	sharp (tight)	Tuning parameters
random	500	10	0.11	1.06	16.47	r=3, c=1.00
non-random	500	30	0.29	6.43	101.35	r=3, c=1.00
random	1000	10	0.21	2.86	44.05	r=3, c=1.00
non-random	1000	30	0.91	21.97	336.78	r=3, c=1.00

6 Application

As an empirical illustration, we revisit Castillo, Jordan, and Petrie (2018), who elicit children's time preferences using a multiple price list (20 binary choices between \$49 in one month and a larger amount in seven months) and study how the resulting discount-rate measure predicts on-time high school graduation. In a multiple price list, respondents face a sequence of binary intertemporal choices in which one option is held fixed while the alternative varies monotonically across rows (here, the later payment increases). Under standard monotonicity, choices should exhibit a single cutoff: the respondent selects the sooner option for lower later amounts and then switches once to the later option when the offered premium becomes sufficiently large. This structure maps each switching point into a bracket for the respondent's discount rate, because all later amounts below the switch are rejected while all later amounts above are accepted.

The setting is attractive for our purposes for three reasons. First, multiple price lists are a workhorse design in experimental and applied micro work, so the example speaks to a broad class of studies. The design intrinsically delivers interval-valued covariates: even with a single switch, the discount rate is only pinned down to a fixed-width bracket determined by two adjacent rows. Second, one common problem for multiple price list elicitation is that some respondents answer inconsistently, i.e. they switch multiple times between accepting and rejecting the later payment. For instance, such a respondent might state to accept \$49 in a month over \$60 in seven month but prefers \$50 in seven month over \$49 in one month, which is inconsistent with standard preference relations. These respondents are reported to differ systematically in observable characteristics (see e.g. the discussion and potential solutions targeting the study design in Yu, Zhang, and Zuo (2021)). Treating such responses as random error or collapsing them to the interval between the first and last switch generates heterogeneous interval lengths that can be correlated with unobservables, raising concerns for identification strategies that rely on mean independence in the presence of interval measurement (C. Manski and Tamer 2002). Third, the main outcome suffers from substantial

selective missingness: roughly 30% of participants have unknown graduation status because they leave the district or otherwise attrit, motivating inverse-probability weighting in the original analysis and providing a natural testbed for methods designed to remain valid under missing outcomes alongside interval measurement.

In the main specification of the original study the authors use an algorithm which finds for all inconsistent decision makers an associated single switch pattern which minimizes the cost to a consistent decision. As the used algorithm produces potential multiplicity, Castillo, Jordan, and Petrie (2018) are using either the average of these artificial single switching patterns, the first switch, the last switch and a cost minimizing third alternative. Thus, for all observation a single switch pattern is either imputed or observed and the midpoint of the associated discount rate interval is used as regressor.

For the moment let us abstract from additional controls and focus on univariate probit models, to see how sensitive the results are to different assumptions. We can weaken their assumptions in several ways. First, we take the single switch rational implementation and only account for the intervals resulting from the multiplicity of their algorithm ("single switch"). Second, we use the same single switch values, but account for the fact that even for them the discount is only known to be within an interval of width 0.0735 ("single switch - interval"). To test the impact of the choice consistency algorithm employed, our third specification keeps the midpoint for the consistent respondents but uses an interval spanning from the first and last switch for the inconsistent ones ("raw switching"). Finally, the last specification ignores the interval measurement but includes the individuals for which high school graduation is not observed ("missing Y"). We use a probit model regressing the outcome on an intercept and the potentially interval-measured discount rate.

Table 3 below contains the estimated sharp identified set of the marginal effect at the mean discount rate for the models described above as well as the original point estimate of the study. We abstract from standard errors and control variables at the moment. Note, however, that the point estimate of -0.09 below for the replication is essentially the same as the

one estimated by the baseline specification in Castillo, Jordan, and Petrie (2018). A 100% percentage point increase decreases the probability of graduating from high school by 9.1%. Without control variables the conclusion of the original study weakens significantly even if we only account for the intervals emerged from the multiplicity of imputed single switching solutions. The estimated identified set for the marginal effect at the mean ranges from -21% to $+6\%$. Although, in this case the number of censored observations is small, the interval width is relatively wide. Ignoring the multiplicity and instead only account for the interval-measured inherent in the multiple price list design yields a similar identified set. Thus, even the small design interval width of 0.0735 yields an ambiguous sign of the effect. This finding should be viewed as a general caution against treating midpoint imputations as harmless if the interval width is small. Even in this case, if the overall variation in covariates is small the estimates can be heavily influenced by it.

Relaxing the assumption further produces even wider identified sets. If we treat the observed responses for inconsistent individuals at face value, even when ignoring the interval-measurement of the consistent respondents, the effect ranges from -42% to $+18\%$. Including the students who could not be tracked and for which the outcome data thus is missing, further increases ambiguity.

The last column presents the computational time for each of the set estimators. We conduct a fairly extensive random search to make sure to reduce the approximation error. As a result the computational time needed for estimation ranges between 33 minutes to 20 hours. As delineated in the simulation section we currently work on a more efficient implementation of our approach in order to make bootstrapping critical values feasible.

	estimated sharp set	nobs	missing (%)	duration (min)
baseline	-0.09	547		
single switch	$[-0.21, 0.06]$	547	0.08	33.00
single switch - interval	$[-0.23, 0.05]$	547	all	197.00
raw switch	$[-0.42, 0.18]$	547	0.25	623.00
missing Y	$[-0.58, 0.42]$	790	0.31 + 0.08	1212.00

Table 2: Marginal Effects at Means

7 Conclusion

Conditional moment models are ubiquitous in economics. We study the common setting in which some continuous conditioning variables are missing or interval-observed. To the best of our knowledge we provide the first operational characterization of the sharp identified set for this problem without imposing auxiliary assumptions. Our results complement imputation-based approaches (e.g. Duan and Pelger (2025)), methods that rely on additional restrictions (e.g. C. Manski and Tamer (2002)), and estimators that leverage external data (e.g. Pacini (2018)). By construction, the worst-case identified set contains the estimates delivered by these approaches (up to finite-sample deviations), making it a natural tool for assessing the informational content of added assumptions.

This paper builds upon Beresteanu, I. Molchanov, and Molinari (2011) in using random set theory to look at sets defined by moment conditions with missing data. We extend their results to accommodate the infinite dimensional setting arising from our conditional moment restrictions as well as the discontinuous mapping implied by the instrument functions used. To the best of our knowledge Aucejo, Bugni, and Hotz (2017) is the only other paper that looks at the identified set of model (1) subject to missing data.

We illustrate the practical value of the approach in an application to multiple price lists, a widely used survey design that inherently generates interval-valued measurements and, for inconsistent respondents, cannot plausibly be treated as random interval noise. The results caution against midpoint imputation in such designs, even when all respondents are consistent.

In our application, the set of marginal effects that are jointly compatible with the interval data and the regression model includes values with signs opposite to the midpoint-based point estimates.

References

Andrews, D. and X. Shi (2013). "Inference based on Conditional Moment Inequalities". In: Econometrica 81.2, pp. 609–666.

— (2017). "Inference based on many conditional moment inequalities". In: Journal of Econometrics 196, pp. 275–287.

Asher, S., P. Novosad, and C. Rafkin (2018). "Partial identification of expectations with interval data". In: Permanent Working Paper.

Aucejo, E., F. Bugni, and J. Hotz (2017). "Identification and Inference on Regressions with Missing Covariate Data". In: Econometric Theory 33, pp. 196–241.

Aumann, R. (1965). "Integrals of Set-Valued Functions". In: Journal of Mathematical Analysis and Applicat 12.1, pp. 1–12.

Beresteanu, A. and F. Molchanov I.and Molinari (2012). "Partial Identification Using Random Set Theory". In: Journal of Econometrics 166.1, pp. 17–32.

Beresteanu, A., I. Molchanov, and F. Molinari (2011). "Sharp Identification Regions in Models with convex Moment Predictions". In: Econometrica 79.6.

Beresteanu, A. and F. Molinari (2008). "Asymptotic Properties for a Class of Partially Identified Models". In: Econometrica 76.4, pp. 763–814.

Billingsley, P. (1995). Probability and Measure. New York: Wiley and Sons.

Canay, I. and A. Shaikh (2017). "Practical and theoretical advances in inference for partially identified models". In: Advances in Economics and Econometrics 2, pp. 271–306.

Castillo, M., J. Jordan, and R. Petrie (2018). "Discount Rates of Children and High School Graduation". In: The Economic Journal 129, pp. 1153–1181.

Chernozhukov, V., H. Hong, and E. Tamer (2007). "Parameter set inference in a class of econometric models". In: Econometrica 75.5, pp. 1243–1284.

Dominguez, M. and I. Lobato (2004). "Consistent Estimation of Models definded by Conditional Moment Restrictions". In: Econometrica 72, pp. 1601–1615.

Duan, J. and M. Pelger (2025). "Imputation-powered inference for missing covariates". In: NBER Working Paper.

Frechet, M. (1951). "Sur les tableaux de corrélation dout les marges sont données". In: Annales de l'Université de Lyon 111, pp. 53–77.

Hsiao, C. (1983). "Regression Analysis with a Categorized Explanatory Variable". In: Studies in Econometrics, Time Series and Multivariate Statistics, pp. 93–129.

Kline, B. and E. Tamer (2023). "Recent Developments in Partial Identification". In: Annual Review of Econor 15.

Magnac, T. and E. Maurin (2008). "Partial Identification in Monotone Binary Models: Discrete Regressors and Interval Data". In: The Review of Economic Studies 75, pp. 835–864.

Manski, C. and E. Tamer (2002). "Inference on Regressions with Interval Data on a Regressor or Outcome". In: Econometrica 70, pp. 519–546.

Manski, C. F. (2005). "Partial identification with missing data: Concepts and findings". In: International Journal of Approximate Reasoning 39.2-3, pp. 151–165.

— (2008). "Partial Identification in Econometrics". In: The New Palgrave Dictionary of Economics. Ed. by S. N. Durlauf and L. E. Blume. Palgrave Macmillan.

Marschak, J. and W.H. Andrews (1944). "Random Simultaneous Equations and the Theory of Production". In: Econometrica 12, pp. 143–205.

Molchanov, I (2017). Theory of Random Sets.

Molchanov, I. and F. Molinari (2018). Random Sets in Econometrics.

Molinari, F. (2020). "Microeconometrics with partial identification". In: Handbook of Econometrics 7A.

Pacini, David (2018). "The two-sample linear regression model with interval-censored covariates". In: Journal of Applied Econometrics 33, pp. 1–16.

Rigobon, R. and T. Stoker (2009). "Bias from censored regressors". In: Journal of Business and Economic Sta 27, pp. 340–353.

Rockafellar, T. and R. Wets (2009). Variational Analysis. Springer.

Stinchombe, M. and H. White (1998). "Consitent Specification Testing with Nuisance Parameters Present Only Under The Alternative". In: Econometric Theory 14, pp. 295–325.

Yu, C., J. Zhang, and S. Zuo (2021). "Multiple Switching and Data Quality in the Multiple Price List". In: The Review of Economics and Statistics 103.1, pp. 136–150.

A Proof of Theorem 1

Fix $z \in \mathcal{Z}$ and work on $(\Omega, \mathcal{A}, P_z)$ where $P_z(\cdot) = P(\cdot|Z = z)$. Under A1, the interval rectangles $D_z(\omega) = ([Y_l(\omega), Y_u(\omega)] \times [X_l(\omega), X_u(\omega)])$ are nonempty compact subsets of $\mathbb{R}^{d_y+d_x}$ for P_z a.e. ω. Let $I = \{(a, r) : \ r \geq r_0, \ a \in \{1, \ldots, 2r\}^{d_x}\}$. Define the map $M_\theta(y, x) \equiv (m(y, x, \theta)g_{a,r}(x))_{(a,r)\in I} \in \ell_1$. Further define the (possible non-closed) random set and its closure as

$$Q_{z,\theta}(\omega) \equiv M_\theta(D_z(\omega)) \subset \ell_1 \qquad \bar{Q}_{z,\theta}(\omega) \equiv \mathrm{cl}\,(Q_{z,\theta}(\omega))$$

Before we prove the Theorem, the following Lemma ensure that the random sets defined above (or in different notation in Theorem 1) are well-behaved.

Lemma 2. *Let assumptions A1-A3 hold, each $\bar{Q}_{z,\theta}$, i.e. the closure of $Q_{z,\theta}$, is non-empty, closed, bounded, absolute integrable and Effros measurable and lives in ℓ_1.*

Proof. Fix $z \in \mathcal{Z}$ and $\theta \in \Theta$ and abbreviate $Q(\omega) \equiv Q_{z,\theta}(\omega)$ and work on $(\Omega, \mathcal{A}, P_z)$ where $P_z(\cdot) = P(\cdot|Z = z)$. For any $x \in [0, 1]^{d_x}$ and fixed r, the cubes $\{C_{a,r}\}_a$ partition the space,

so $\sum_a \mathbf{1}\{x \in C_{a,r}\} = 1$ and hence

$$\sum_{(a,r)\in I} |g_{a,r}(x)| = \sum_{r=r_0}^{\infty} \sum_a w(r)\mathbf{1}\{x \in C_{a,r}\} = \sum_{r=r_0}^{\infty} w(r) \equiv W < \infty.$$

where the last equality follows from $\beta > 1$ in the definition of the weights. By Assumption A3, $|m(y, x, \theta)| \leq C_1$. Therefore, for any admissible (y, x),

$$\left\| \big(m(y, x, \theta) g_{a,r}(x)\big)_{(a,r)\in I} \right\|_{\ell_1} \leq \sum_{(a,r)\in I} C_1 |g_{a,r}(x)| = C_1 W.$$

Thus every element $q(\omega)$ lies in ℓ_1 and $Q(\omega) \subset \ell_1$ is bounded by the deterministic radius $C_1 W$. Non-emptiness follows because the interval-valued sets $[Y_l, Y_u]$ and $[X_l, X_u]$ are non-empty a.s. under A1. Further, by definition $\bar{Q}(\omega)$ is closed-valued and $\bar{Q}$ is absolute integrable since $\|\bar{Q}(\omega)\| = \sup_{q\in\bar{Q}(\omega)} \|q\|_{\ell_1} \leq C_1 W$ a.s., so $\mathbb{E}\|\bar{Q}\| < \infty$.

It remains to show Effros measurability of $\bar{Q}(\omega)$. Note that ℓ_1 is Polish (separable and complete). Since taking the closure does not change whether a set hits an open set (Proposition 1.3.29 in Molchanov (2017)), we have $\{\omega : \bar{Q}(\omega) \cap G \neq \emptyset\} = \{\omega : M_\theta(D_z(\omega)) \cap G \neq \emptyset\}$.

We next verify that M_θ is Borel measurable. Write a generic element of ℓ_1 as $q = (q_{a,r})_{(a,r)\in I}$ and let $\pi_{a,r} : \ell_1 \to \mathbb{R}$ denote the coordinate projection $\pi_{a,r}(q) = q_{a,r}$. Since the index set I is countable and ℓ_1 is separable, the Borel σ-field $\mathcal{B}(\ell_1)$ is generated by $\{\pi_{a,r} : (a, r) \in I\}$. For each $(a, r) \in I$, $(\pi_{a,r} \circ M_\theta)(y, x) = m(y, x, \theta)w(r)\mathbb{I}\{x \in C_{a,r}\}$ which is Borel because $m(\cdot, \cdot, \theta)$ is continuous and $C_{a,r}$ is a Borel subset of $[0, 1]^{d_x}$. Hence M_θ is Borel measurable. Therefore, Proposition 1.3.30 in Molchanov (2017) implies that the random set $\omega \mapsto M_\theta(D_z(\omega))$ is weakly measurable. Combining this with the display above yields $\{\omega : \bar{Q}(\omega) \cap G \neq \emptyset\} \in \mathcal{A}$ for every open $G \subset \ell_1$, so $\bar{Q}$ is Effros measurable. Since $\bar{Q}$ is closed-valued by construction, it is a random closed set (Definition 1.3.1 in Molchanov (2017)). □

Lemma 3. *Under A1-A3, $Q_{\theta,z}$ is graph measurable, hence $\mathbb{E}[Q_{\theta,z}] = \mathbb{E}[\bar{Q}_{\theta,z}]$. Moreover, $\mathbb{E}[Q_{\theta,z}]$ is compact.*

Proof. Without loss of generality, we work on the completion of $(\Omega, \mathcal{A}, P_z)$ and define graph measurability with respect to the completed product σ-algebra. For the first part, we only need to show graph measurability as given that, the result follows from Proposition 2.1.25 in Molchanov (2017). To that end, consider the measurable set

$$A \equiv \left\{(\omega, (y,x), q) : \ (y,x) \in D_z(\omega), q = M_\theta(y,x)\right\} \subset \Omega \times \mathbb{R}^{d_y+d_x} \times \ell_1$$

$D_z(\omega)$ is a random closed set, thus graph-measurable under A1 and by Lemma A.2 in Beresteanu and Molinari (2008). Further, M_θ is Borel measurable as established in the previous Lemma, thus A is measurable in $\mathcal{A} \otimes \mathcal{B}(\mathbb{R}^{d_y+d_x}) \otimes \mathcal{B}(\ell_1)$. The graph of the random set $Q_{z,\theta}$ is defined as

$$\text{Gr}(Q_{z,\theta}) = \left\{(w, q) : \ \exists \ (y,x) \text{ such that } (\omega, (y,x), q) \in A\right\}$$

which is the projection of A onto $\Omega \times \ell_1$. The domain of this map, i.e. what is projected out, is $\mathbb{R}^{d_y+d_x}$, thus a locally compact Hausdorff space. By Theorem F.7 in Molchanov (2017) we can conclude that $Q_{z,\theta}$ is graph measurability.

Next we show compactness. Note that $\mathbb{E}[Q_{z,\theta}]$ is closed by definition, thus it suffices to show that $\mathbb{E}[Q_{z,\theta}]$ is relatively compact. To that end, let $A_r \equiv \{1, \ldots, 2r\}^{d_x}$ and define

$$K \equiv \left\{q = (q_{a,r})_{(a,r)\in I} \in \ell_1 \ : \ \sum_{a \in A_r} |q_{a,r}| \leq C_1 \, w(r), \quad \forall r \geq r_0\right\}.$$

Because ℓ_1 is a metric space is suffices to show that every sequence in K has a convergent subsequence with limit in K. For any $q \in K$, as established in Lemma 2 we have $||q|| \leq C_1 W$, thus K is bounded. Moreover, for any $R \geq r_0$

$$\sum_{r>R} \sum_{a \in A_r} |q_{a,r}| \leq C_1 \sum_{r>R} w(r) \underset{R\to\infty}{\to} 0$$

which does not depend on q thus the tail bound is uniform, i.e. $\sup_{q\in K}\sum_{r>R}\sum_{a\in A_r}|q_{a,r}|\to 0$. Now define the truncation at R as $T_R:\ell_1\to\ell_1$

$$(T_Rq)_{a,r}\equiv\begin{cases}q_{a,r} & r_0\le r\le R\\ 0 & r>R\end{cases}$$

Let $E_R\equiv\{q\in\ell_1 : q_{a,r}=0\ \forall r>R\}$. Then E_R is finite dimensional because it is determined bz finitely many coordinates. The set $T_R(K)$ is a closed and bounded subset of E_R, thus it is compact by Heine-Borel. Hence, for each R we can extract a subsequence along which T_Rqn converges in E_R and using a diagonal argument, there exists a subsequence (q^{n_k}) such that for every fixed R, $R_Rq^{n_k}$ converges in E_R as $k\to\infty$. Now define $q=(q_{a,r})$ coordinate-wise as $q_{a,r}\equiv\lim_{k\to\infty}q^{n_k}_{a,k}$, which exists as we established coordinate-wise convergence. Because A_r is finite

$$\sum_{a\in A_r}|q_{a,r}|=\lim_{k\to\infty}\sum_{a\in A_r}|q^{n_k}_{a,r}|\le C_1w(r)$$

thus $q\in K$. It remains to show that $q^{n_k}\to q$, i.e. convergence in full ℓ_1-norm and not only truncation and coordinate-wise. As $C_1\sum_{r>R}w(r)\underset{R\to\infty}{\to}0$ we can choose R such that $C_1\sum_{r>R}w(r)<\epsilon/3$ for some $\epsilon>0$. Then for every $k\ge k_0$ and $q^{n_k}\in K$ as well as $q\in K$ we have

$$||q^{n_k}-T_Rq^{n_k}||_{\ell_1}=\sum_{r>R}\sum_{a\in A_r}|q^{n_k}_{a,r}|<\epsilon/3$$
$$||q-T_Rq||_{\ell_1}=\sum_{r>R}\sum_{a\in A_r}|q_{a,r}|<\epsilon/3$$

By construction of $T_Rq^{n_k}$ converges to T_Rq in E_R, i.e. $||T_Rq^{n_k}-T_Rq||_{\ell_1}<\epsilon/3$. Thus by the triangle inequality $||q^{n_k}-q||_{\ell_1}<\epsilon$. Therefore $q^{n_k}\to q$ in ℓ_1, so every sequence in K has an ℓ_1 convergent subsequence with limit in K, thus K is compact. By construction

$Q_{z,\theta} \subseteq K \subset \text{co}(K)$, where $\bar{\text{co}}$ denotes the convex hull of K. By standard arguments also $\mathbb{E}[Q_{z,\theta}] \subset \bar{\text{co}}(K)$ concluding the proof. □

With the above Lemmata we are able to prove Theorem 1:

Proof. The proof consists of two parts. First, we present a natural condition in terms of Minkowksi sums and show it is equivalent to the support function representation in the main text. Then, we verify that the condition is indeed providing the sharp identified set.
The natural analog to the point identified case is to consider the weighted Minkowski sum of the Aumann expectations of all random sets:

$$\mathbf{0} \in \underset{z \in \mathcal{Z}}{\oplus} \mathbb{E}[Q_{\theta,z}]P(Z = z) = \underset{z \in \mathcal{Z}}{\oplus} \mathbb{E}[\bar{Q}_{\theta,z}]P(Z = z)$$

where the equality follows from Lemma 3. Under mild assumptions (no atoms), each $\mathbb{E}[\bar{Q}_{\theta,z}]$ is convex (Theorem 2.1.30 in Molchanov (2017)). Further the (weighted) Minkowski sum of convex sets is convex (Proposition 2.23 in Rockafellar and Wets (2009)), thus $\underset{z \in \mathcal{Z}}{\oplus} \mathbb{E}[\bar{Q}_{\theta,z}]P(Z = z)$ is convex. The Minkowski sum above is not necessarily closed unless one of the summands is compact (see e.g. page 229 in Molchanov (2017)), which is ensured by Lemma 3. Thus, by standard convex analysis results and because the Minkowski sum of random sets lives in ℓ_1 and is closed we have

$$\mathbf{0} \in \underset{z \in \mathcal{Z}}{\oplus} \mathbb{E}[\bar{Q}_{\theta,z}]P(Z = z) \qquad \Longleftrightarrow \qquad h(\underset{z \in \mathcal{Z}}{\oplus} \mathbb{E}[\bar{Q}_{\theta,z}]P(Z = z), u) \geq 0 \quad \forall u \in L$$

with $L = B_1^\infty$. Further, by the linearization property of the support function with respect to Minkowski sums (Corollary 11.24 in Rockafellar and Wets (2009))

$$h(\underset{z \in \mathcal{Z}}{\oplus} \mathbb{E}[\bar{Q}_{\theta,z}]P(Z = z), u) \geq 0 \quad \forall u \in L \qquad \Longleftrightarrow \qquad \sum_{z \in \mathcal{Z}} h(\mathbb{E}[\bar{Q}_{\theta,z}]P(Z = z), u) \geq 0 \quad \forall u \in L$$

Scaling the set by $P(Z = z)$ translates into scaling the support function. Further, by

Theorem 2.1.35 in Molchanov (2017) the expectation of the support function equals the support function of the Aumann expectation under the conditions shown in Lemma 1 and a non-atomic probability space. Further, the support function of the closure is equal to the support of the set itself (see. e.g. proof of Proposition 8.29 in Rockafellar and Wets (2009)), we can replace $\bar{Q}_{\theta,z}$ with $Q_{\theta,z}$ in the last statement.

$$\mathbf{0} \in \underset{z\in\mathcal{Z}}{\oplus} \mathbb{E}[\bar{Q}_{\theta,z}]P(Z=z) \qquad \Longleftrightarrow \qquad \inf_{u\in L} \sum_{z\in\mathcal{Z}} \mathbf{E}[h(\bar{Q}_{\theta,z},u)]P(Z=z) \geq 0$$

It remains to show that the condition given above is indeed a sharp representation, i.e.

$$\theta \in \Theta_I \qquad \Longleftrightarrow \qquad \mathbf{0} \in \underset{z\in\mathcal{Z}}{\oplus} \mathbb{E}[Q_{\theta,z}]P(Z=z)$$

The $\Rightarrow$ implication is trivial. Given a completion $(\tilde{Y},\tilde{X})$ satisfying A1 define $q_{\theta,z}(\omega) \equiv M_\theta(\tilde{Y}(\omega),\tilde{X}(\omega))$. Then, $q_{\theta,z}(\omega) \in Q_{\theta,z}(\omega)$, thus by definition of the Aumann expectation $\mathbf{E}[q_{\theta,z}] \in \mathbb{E}[Q_{\theta,z}]$ for each z, which fullfills the moment condition by Lemma 1 in the main text.

For $\Leftarrow$, note that the right hand side implies that there exists $\mu_z \in \mathbb{E}[Q_{\theta,z}]$ such that $\sum_z P(Z=z)\mu_z = 0$. The complication is that the Aumann expectation is defined as the closure of all selection expectations $\mathbb{E}[X] = cl(\{\mathbf{E}[x] : \quad x \in Sel(X), \mathbf{E}[x] < \infty\})$, a prior μ_z may only be the limit of expectations of selections. In general Banach spaces the set of selection integrals need not be closed (see Example 2.1.26 in Molchanov (2017)). However, by Lemma 2 we have $\mathbb{E}[Q_{\theta,z}] = \mathbb{E}[\bar{Q}_{\theta,z}]$ so we can work with the closed-valued random set $\bar{Q}_{\theta,z}$. Lemma 3 constructed a deterministic compact set K such that $\bar{Q}_{\theta,z}(\omega) \subset K$ for P_z-a.e. ω. Since $\bar{Q}_{\theta,z}$ is closed it is compact (hence weakly compact) and we can apply Theorem 2.1.37 of Molchanov (2017), implying that the Aumann integral is closed and the closure in the definition of $\mathbb{E}[\bar{Q}_{\theta,z}]$ is redundant. Thus, for each $z \in \mathcal{Z}$ there exists an integrable selection $q_{\theta,z} \in \text{Sel}(\bar{Q}_{\theta,z})$ such that $\mathbb{E}[q_{\theta,z}] = \mu_z$. □

B Proof of Theorem 2

We want to show $\rho_H(\hat{\Theta}_I^{K(n)}, \Theta_I) \to^p 0$ by applying Theorem 3.1 in Chernozhukov, Hong, and Tamer (2007). Their results rests on condition C.1 of the same paper, which we later verify for our setting.

For simplification, we note that the variable Z and the respectively constructed random sets are mainly introduced in the main text for expositional clarity and to compare our results with results in Aucejo, Bugni, and Hotz (2017) and the related literature. We can equivalently define one random set encapsulating the same information. Our $Q_{\theta,z}$ is obtained from the *law of total probability* applied to random sets:

$$Q_\theta \equiv \underset{z\in\mathcal{Z}}{\oplus} Q_{\theta,Z}\mathbb{I}\{Z=z\} \qquad Q_\theta^{K(n)} \equiv \underset{z\in\mathcal{Z}}{\oplus} Q_{\theta,Z}^{K(n)}\mathbb{I}\{Z=z\}$$

Let us also define the K dimensional unit ball using the supremum norm as $B_1^K \equiv \{u \in \mathbb{R}^K : \sup_{k=1,\cdots,K} |u_k| = 1\}$. Finally, as we follow the criterion function formulation of Chernozhukov, Hong, and Tamer (2007), let us define the following criterion functions:

$$T_n(\theta) \equiv \vec{\rho}_H\left(0, \mathbb{E}[Q_\theta]\right) \qquad \hat{T}_n(\theta) \equiv \vec{\rho}_H\left(0, \frac{1}{n}\overset{n}{\underset{i=1}{\oplus}} Q_{\theta,i}^{K(n)}\right)$$

The directed hausdorff distance $\vec{\rho}_H$, captures how close the origin 0 is to $\mathbb{E}[Q_\theta]$. Note $\vec{\rho}_H \geq 0$ and $\vec{\rho}_H$ is a hemi-metric: a non-symmetrical, distance function. $T_n(\theta)$ defines the identified set through the equivalence $0 \in \mathbb{E}[Q_\theta] \Leftrightarrow T_n(\theta) = 0 \Leftrightarrow \theta \in \Theta_I$. See $T_n(\theta)$ does not vary with n, but $\hat{T}_n(\theta)$ does. Using this notation we can define the key inferential quantity $\mathcal{C}_n \equiv \sup_{\theta\in\Theta_I} a_n T_n(\theta)$. We restate the estimator defined in the main text:

$$\begin{aligned}\hat{\Theta}_I^{K(n)} &= \{\theta \in \Theta : \min_{u\in B^{K(n)}} \sum_{z\in\mathcal{Z}} \frac{1}{n}\sum_{i=1}^n [h_i(Q_{\theta,z,i}, u)]P(Z=z_i) \geq -\tau_n\} \\ &= \{\theta \in \Theta : \ a_n\hat{T}_n(\theta) \leq \tau_n\}\end{aligned}$$

where $B_1^{K(n)} \equiv \{u \in \mathbb{R}^{K(n)} : \|u\|_1 \leq 1\}$. Under assumptions A1-A4 if $K(n)/n \to 0$, $K(n) \to \infty$, ($\sqrt{\frac{n}{K(n)}}\tau_n \to \infty$ and $\tau_n \to 0$.

Note that as Θ is compact by assumption 1, as Q_θ is (set) continuous in θ, it follows that rates of convergence for a fixed θ extend to the supremum over all $\theta \in \Theta$.

With these we are equipped to check conditions (a)-(e) in condition C.1 of Chernozhukov, Hong, and Tamer (2007):

a) Θ is a compact subset of $\mathbb{R}^d$

   Holds by assumption 2.

b) $T_n(\theta)$ is continuous and $\inf_{\theta \in \Theta} T_n(\theta) = 0$

   Follows directly from $\mathbb{E}[Q_{\theta,z}^{K(N)}]$ being continuous in θ for each $z \in \mathcal{Z}$ by assumption 3 and from continuity of the support function in u and θ.

c) $\hat{T}_n(\theta)$ is jointly measurable.

   Not strictly necessary according to Chernozhukov, Hong, and Tamer (2007). Nevertheless, holds from the measurability of $Q_\theta^{K(N)}$.

d) The sample criterion function is uniformly no smaller than the population function in

large samples:

$$
\begin{aligned}
\sup_{\theta\in\Theta}\big(T(\theta)-\widehat{T}_n(\theta)\big)_+ &= \sup_{\theta\in\Theta}\Big(\vec{\rho}_H\big(0,\mathbb{E}[Q_\theta]\big)-\vec{\rho}_H\Big(0,\frac{1}{n}\overset{n}{\underset{i=1}{\oplus}}Q_{\theta,i}^{K(n)}\Big)\Big)_+ \\
&\le \sup_{\theta\in\Theta}\Big|\vec{\rho}_H\big(0,\mathbb{E}[Q_\theta]\big)-\vec{\rho}_H\Big(0,\frac{1}{n}\overset{n}{\underset{i=1}{\oplus}}Q_{\theta,i}^{K(n)}\Big)\Big| \\
&= \sup_{\theta\in\Theta}\Big|\vec{\rho}_H\big(0,\mathbb{E}[Q_\theta]\big)-\vec{\rho}_H\big(0,\mathbb{E}[Q_\theta^{K(n)}]\big) \\
&\qquad +\vec{\rho}_H\big(0,\mathbb{E}[Q_\theta^{K(n)}]\big)-\vec{\rho}_H\Big(0,\frac{1}{n}\overset{n}{\underset{i=1}{\oplus}}Q_{\theta,i}^{K(n)}\Big)\Big| \\
&\le \sup_{\theta\in\Theta}\Big|\vec{\rho}_H\big(0,\mathbb{E}[Q_\theta]\big)-\vec{\rho}_H\big(0,\mathbb{E}[Q_\theta^{K(n)}]\big)\Big| \\
&\qquad +\sup_{\theta\in\Theta}\Big|\vec{\rho}_H\big(0,\mathbb{E}[Q_\theta^{K(n)}]\big)-\vec{\rho}_H\Big(0,\frac{1}{n}\overset{n}{\underset{i=1}{\oplus}}Q_{\theta,i}^{K(n)}\Big)\Big| \\
&\le \sup_{\theta\in\Theta}\Big|\vec{\rho}_H\big(0,\mathbb{E}[Q_\theta]\big)-\vec{\rho}_H\big(0,\mathbb{E}[Q_\theta^{K(n)}]\big)\Big| \\
&\qquad +\sup_{\theta\in\Theta}\rho_H\Big(\mathbb{E}[Q_\theta^{K(n)}],\frac{1}{n}\overset{n}{\underset{i=1}{\oplus}}Q_{\theta,i}^{K(n)}\Big) \\
&= O_p\left(K(n)^{-\beta}\right)+O_p\left(\sqrt{\frac{K(n)}{n}}\right) \\
&= \max\{O_p\left(K(n)^{-\beta}\right),O_p\left(\sqrt{\frac{K(n)}{n}}\right)\}=O_p(1/b_n)
\end{aligned}
$$

where the first line is definitional, the second uses $(x)_+ \equiv \max\{0,x\} \le |x|$. The third line adds and subtracts $\vec{\rho}_H\big(0,\mathbb{E}[Q_\theta^{K(n)}]\big)$ inside the absolute value. The fourth inequality uses the triangle inequality. The fifth line uses Lemma 3 which establishes $|\vec{\rho}(0,A)-\vec{\rho}(0,B)| \le \rho_H(A,B)$. Finally, by Lemma 4 we have $\Big|\vec{\rho}_H\big(0,\mathbb{E}[Q_\theta]\big)-\vec{\rho}_H\big(0,\mathbb{E}[Q_\theta^{K(n)}]\big)\Big| = O_p\left(K(n)^{-\beta}\right)$ and Lemma 5 establishes $\rho_H\Big(\mathbb{E}[Q_\theta^{K(n)}],\frac{1}{n}\overset{n}{\underset{i=1}{\oplus}}Q_{\theta,i}^{K(n)}\Big) = O_p\left(\sqrt{\frac{K(n)}{n}}\right)$. Thus

$$
b_n = \big(K(n)^\beta\big)\wedge\left(\sqrt{\frac{n}{K(n)}}\right) \quad \text{thus} \quad b_n\to\infty
$$

The first part of this rate captures and approximation error coming from using finitely many moments instead of infinitely many as the identification theorem would require.

The second rate gives the a rate for the sampling error of the estimator.

e) The sample criterion converges to the limit criterion function over the identified set Θ_I at the $1/a_n$ rate.

$$\begin{aligned}
\sup_{\theta\in\Theta_I}\big(T(\theta)-\widehat{T}_n(\theta)\big)_+ &= \sup_{\theta\in\Theta_I}\Big(\vec{\rho}_H\big(0,\mathbb{E}[Q_\theta]\big)-\vec{\rho}_H\Big(0,\frac{1}{n}\overset{n}{\underset{i=1}{\oplus}}Q_{\theta,i}^{K(n)}\Big)\Big)_+ \\
&\le \sup_{\theta\in\Theta_I}\Big|\,0-\vec{\rho}_H\Big(0,\frac{1}{n}\overset{n}{\underset{i=1}{\oplus}}Q_{\theta,i}^{K(n)}\Big)\Big| \\
&= \sup_{\theta\in\Theta_I}\vec{\rho}_H\Big(0,\frac{1}{n}\overset{n}{\underset{i=1}{\oplus}}Q_{\theta,i}^{K(n)}\Big) \\
&\le \sup_{\theta\in\Theta_I}\vec{\rho}_H\big(0,\mathbb{E}[Q_\theta^{K(n)}]\big)+\sup_{\theta\in\Theta_I}\vec{\rho}_H\Big(\mathbb{E}[Q_\theta^{K(n)}],\frac{1}{n}\overset{n}{\underset{i=1}{\oplus}}Q_{\theta,i}^{K(n)}\Big) \\
&= \sup_{\theta\in\Theta_I}\vec{\rho}_H\Big(\mathbb{E}[Q_\theta^{K(n)}],\frac{1}{n}\overset{N}{\underset{i=1}{\oplus}}Q_{\theta,i}^{K(n)}\Big) \\
&= \sup_{\theta\in\Theta_I}O_p\left(\sqrt{\frac{K(n)}{n}}\right)=O_p(1/a_n),\quad a_n=\sqrt{\frac{n}{K(n)}}.
\end{aligned}$$

The first line is definitional, the second uses $(x)_+ \equiv \max\{0,x\} \le |x|$ together with $T(\theta)=0$ for all $\theta\in\Theta_I$ and the third line uses the fact that $\vec{\rho}_H$ is a (hemi-) Hausdorff distance, thus non-negative. The fourth line follows by the triangle inequality for hemi-metrics. Next note that $\mathbb{E}[Q_\theta^{K(n)}]\subseteq\mathbb{E}[Q_\theta]$, thus $\vec{\rho}_H\big(0,\mathbb{E}[Q_\theta^{K(n)}]\big)=0$. Finally the last rate follows from Lemma 5.

To complete the proof, after verifying condition C.1 in Chernozhukov, Hong, and Tamer (2007), our theorem holds immediately from from Theorem 3.1(i) in Chernozhukov, Hong, and Tamer (2007) □

C Auxiliary Lemmata for Proof of Theorem 2

Here we prove some Lemmata used in the proof of Theorem 2 above. First, the following Lemma links the deviation of directed Hausdorff distances to the undirected Hausdorff distances between two sets.

Lemma 4. $\left|\vec{\rho}_H(0,A) - \vec{\rho}_H(0,B)\right| \leq \rho_H(A,B)$

Proof. By the triangle inequality of hemi-metrics $q \in B$, $\vec{\rho}_H(0,A) \leq \vec{\rho}_H(0,q) + \vec{\rho}_H(q,A)$ (see Lemma 6.2.16 in Goubault-Larrecq, Ch. 6). Therefore,

$$\begin{aligned}\vec{\rho}_H(0,A) &\leq \min_{q\in B}\Big(\vec{\rho}_H(0,q) + \vec{\rho}_H(q,A)\Big) \leq \min_{q\in B}\Big(\rho_H(0,q) + \rho_H(q,A)\Big)\\ &\leq \vec{\rho}_H(0,B) + \rho_H(A,B).\end{aligned}$$

Hence $\vec{\rho}_H(0,A) - \vec{\rho}_H(0,B) \leq \rho_H(A,B)$. Reversing the inequality, for all $q \in A$ we similarly obtain $\vec{\rho}_H(0,B) \leq \vec{\rho}_H(0,q) + \vec{\rho}_H(q,B)$ so that

$$\begin{aligned}\vec{\rho}_H(0,B) &\leq \min_{q\in A}\Big(\vec{\rho}_H(0,q) + \vec{\rho}_H(q,B)\Big) \leq \min_{q\in A}\Big(\rho_H(0,q) + \rho_H(q,B)\Big)\\ &\leq \vec{\rho}_H(0,A) + \rho_H(A,B).\end{aligned}$$

Hence $\vec{\rho}_H(0,B) - \vec{\rho}_H(0,A) \leq \rho_H(A,B)$ Combining both inequalities gives the desired result. □

Next let us bound the approximation error of using a finite number of moments $K(n)$ instead of infinitely many.

Lemma 5. *Let A1-A4 hold, and using $Q_\theta^{K(N)}$ defined as a truncation of moments using the instrument class $\mathcal{G}_{c-cube}$, and weighted by terms on the order of $O(K(N)^{-\beta})$ for $\beta > 0$. We suspect similar results that are faster than $O_p(1)$ will hold for a larger class of instruments and also unweighted instruments; note unweighted instrument functions greatly complicate proofs.*

$$\left|\vec{\rho}_H\big(0,\mathbb{E}[Q_\theta]\big) - \vec{\rho}_H\big(0,\mathbb{E}[Q_\theta^{K(n)}]\big)\right| = O_p\left(K(n)^{-\beta}\right)$$

Proof. Note that we can write the directional Hausdorff distance as $\vec{\rho}_H(0,A) = \inf_{q\in A}\|q\|_\infty$.

Hence

$$\left|\vec{\rho}_H\big(0,\mathbb{E}[Q_\theta^\infty]\big)-\vec{\rho}_H\big(0,\mathbb{E}[Q_\theta^{K(N)}]\big)\right| = \left|\inf_{q_1\in\mathbb{E}[Q_\theta^\infty]}\|q_1\|_\infty - \inf_{q_2\in\mathbb{E}[Q_\theta^{K(N)}]}\|q_2\|_\infty\right|$$

$$= \left|\inf_{(x,y)\in \mathrm{Sel}(X\times Y)}\max_{j\geq 1}\left|\mathbf{E}_{F_{x,y}}\big[m(x,y,\theta)g_j(x)\big]\right|\right.$$

$$\left. - \inf_{(x,y)\in \mathrm{Sel}(X\times Y)}\max_{1\leq j\leq K(n)}\left|\mathbf{E}_{F_{x,y}}\big[m(x,y,\theta)g_j(x)\big]\right|\right|$$

Connecting this to our choice of $\mathcal{G}_{c-cube}$: for each selection (x,y) and moment j, define $G_{x,y}(j)\equiv (j^\beta+100)^{-1}E_{F_{x,y}}[m(x,y,\theta)\mathbb{I}\{x\in B_j\}]$.

Then

$$\left|\vec{\rho}_H\big(0,\mathbb{E}[Q_\theta^\infty]\big)-\vec{\rho}_H\big(0,\mathbb{E}[Q_\theta^{K(N)}]\big)\right| = \left|\inf_{(x,y)}\sup_{j\geq 1}|G_{x,y}(j)| - \inf_{(x,y)}\max_{1\leq j\leq K(n)}|G_{x,y}(j)|\right|$$

$$\leq \sup_{(x,y)\in \mathrm{Sel}(X\times Y)}\left|\sup_{j\geq 1}|G_{x,y}(j)| - \max_{1\leq j\leq K(n)}|G_{x,y}(j)|\right|$$

by the inequality $|\inf f-\inf g|\leq \sup|f-g|$.

Further as $\sup_{1\leq j\leq K(n)}|G_{x,y}(j)| = \max_{1\leq j\leq K(n)}|G_{x,y}(j)|$

$$\left|\vec{\rho}_H\big(0,\mathbb{E}[Q_\theta^\infty]\big)-\vec{\rho}_H\big(0,\mathbb{E}[Q_\theta^{K(N)}]\big)\right| \leq \sup_{(x,y)\in \mathrm{Sel}(X\times Y)}\sup_{j>K(n)}|G_{x,y}(j)|$$

$$= \sup_{(x,y)\in \mathrm{Sel}(X\times Y)}\sup_{j>K(n)}(j^\beta+100)^{-1}|E_{F_{x,y}}[m(x,y,\theta)\mathbb{I}\{x\in B_j\}]| \leq C\cdot\sup_{j>K(n)}(j^\beta+100)^{-1}$$

Here we use assumption 1 to bound $m(x,y,\theta)$ over all selections (x,y) and all θ. Therefore we arrive at the conclusion

$$\left|\vec{\rho}_H\big(0,\mathbb{E}[Q_\theta]\big)-\vec{\rho}_H\big(0,\mathbb{E}[Q_\theta^{K(n)}]\big)\right| = O_p\left(K(n)^{-\beta}\right), \beta>0$$

□

Finally, the last auxiliary results establishes a law of large deviations for random sets with growing dimensions.

Lemma 6. *Let A1-A3 be true. then*

$$\rho_H\left(n^{-1}\overset{n}{\underset{i=1}{\oplus}} Q_{\theta,i}^{K(n)}, \mathbb{E}[Q_\theta^{K(n)}]\right) = O_p\left(\sqrt{\frac{K(n)}{n}}\right)$$

Proof. The proof mimics Molchanov (2017), Theorem 3.3.16, except that the necessary covering number of a unit sphere changes as its dimension $K(n)$ changes.
We first restrict attention to convex sets and define

$$co\left[Q_\theta^{K(n)}\right] = n^{-1}\overset{n}{\underset{i=1}{\oplus}} co\left[Q_{\theta,i}^{K(n)}\right]$$

The sets $\left\|Q_{\theta,i}^{K(n)}\right\| = \left\|\text{co}\left[Q_{\theta,i}^{K(n)}\right]\right\| \leq b$ are bounded as previously established. Define

$$\gamma \equiv \text{ess sup}\, \rho_H\Big(\text{co}\big[Q_{\theta,i}^{K(n)}\big], \mathbb{E}\big[Q_\theta^{K(n)}\big]\Big) \leq 2b,$$

which is bounded. Let $N\left(S^{K(n)-1}, \delta\right)$ represent the covering number of a $(K(n)-1)$-dimensional unit sphere by a net of radius δ. Then by Lemma 4.10 in Pisier (1989),

$$N\left(S^{K(n)-1}, \delta\right) \leq \left(1+\frac{2}{\delta}\right)^{K(n)}.$$

Repeating the steps of Molchanov (2017), Theorem 3.3.16, we represent the Hausdorff distance via support functions:

$$\mathbb{P}\left[\rho_H\left(\text{co}\big[Q_\theta^{K(n)}\big], \mathbb{E}\big[Q_\theta^{K(n)}\big]\right) > \varepsilon\right] = \mathbb{P}\left[\sup_{u\in S^{K(n)-1}}\left|h\left(\text{co}\big[Q_\theta^{K(n)}\big], u\right) - h\left(\mathbb{E}\big[Q_\theta^{K(n)}\big], u\right)\right| > \varepsilon\right]$$

Generate a net over $S^{K(n)-1}$,

$$\mathcal{U}(K(n))\left\{u_1,\ldots,u_M : u_j \in S^{K(n)-1},\ \forall u \in S^{K(n)-1},\ \exists j \text{ such that } \|u-u_j\|_1 \leq \delta\right\}$$

with cardinality $M \leq (1+2/\delta)^{K(n)}$ and holes of radius $\delta = \varepsilon/\left(4b(1-\rho^{1/2})\right)$. Therefore,

$$\mathbb{P}\left[\rho_H\left(\widehat{Q}_\theta^{K(n)}, Q_\theta^{K(n)}\right) > \varepsilon\right] \leq (1+2/\delta)^{K(n)} \exp\left(-\frac{\rho\varepsilon^2 n}{4\gamma^2}\right)$$

where the inequality follows from Equation (3.3.12) in Molchanov (2017), and β is bounded. Let $0<\rho<1$ be fixed arbitrarily.

Hence, we can write:

$$\mathbb{P}\left[\rho_H\left(\widehat{Q}_\theta^{K(n)}, Q_\theta^{K(n)}\right) > M\sqrt{\frac{K(n)}{n}}\right] \leq \left(1+\frac{2}{\delta}\right)^{K(n)} \exp\left(-\frac{\rho M^2 K(n)}{4\gamma^2}\right)$$

which simplifies to:

$$\exp\left(K(n)\log\left(1+\frac{2}{\delta}\right) - \frac{\rho M^2 K(n)}{4\beta^2}\right) = \exp\left(K(n)\left[\log\left(1+\frac{2}{\delta}\right) - \frac{\rho M^2}{4\gamma^2}\right]\right).$$

For sufficiently large M, the term in brackets becomes negative, and the probability bound converges to zero as $n \to \infty$. Therefore,

$$\rho_H\left(\widehat{Q}_\theta^{K(n)}, Q_\theta^{K(n)}\right) = O_p\left(\sqrt{\frac{K(n)}{n}}\right).$$

Finally let us consider non-convex sets $Q_\theta^{K(n)}$. From the Shapley-Folkman-Starr theorem (Molchanov (2017) Theorem D.13) with growing dimensionality, as $Q_{\theta,i}^{K(n)}$ is compact from previous results,

$$\rho_H\left(\mathrm{co}\left[\widehat{Q}_\theta^{K(n)}\right], \widehat{Q}_\theta^{K(n)}\right) \leq \frac{b\,K(n)}{n}.$$

It follows immediately that

$$\rho_H\Big(\mathrm{co}\big[\widehat{Q}_\theta^{K(n)}\big], \widehat{Q}_\theta^{K(n)}\Big) = O_p\left(\sqrt{\frac{K(n)}{n}}\right).$$

Thus, combining inequalities yields

$$\begin{aligned}\rho_H\Big(\widehat{Q}_\theta^{K(n)}, \mathbb{E}\big[Q_\theta^{K(n)}\big]\Big) &\leq \rho_H\Big(\widehat{Q}_\theta^{K(n)}, \mathrm{co}\big[\widehat{Q}_\theta^{K(n)}\big]\Big) + \rho_H\Big(\mathrm{co}\big[\widehat{Q}_\theta^{K(n)}\big], \mathbb{E}\big[Q_\theta^{K(n)}\big]\Big) \\ &= \max\left\{O_p\left(\sqrt{\frac{K(n)}{n}}\right), O_p\left(\sqrt{\frac{K(n)}{n}}\right)\right\} = O_p\left(\sqrt{\frac{K(n)}{n}}\right).\end{aligned}$$

□

D Proof of Proposition 1

This section verifies the assumptions in D. Andrews and Shi (2017) and thus proves the coverage results in Proposition 1.

In particular, we have to verify PS1 and PS2 in D. Andrews and Shi (2017), restated for convenience in their notation here:

Assumption PS1. For any $(\theta, F) \in \mathcal{F}^+$, (a) $\theta \in \Theta$, (b) $\{W_i : i \geq 1\}$ are i.i.d. under F (c) $\sigma_F(\theta) > 0$, (d) $\left|\frac{\tilde{m}(w,\theta,\tau)}{\sigma_F(\theta)}\right| \leq M(w)$, $\forall w \in \mathbb{R}^{d_w}$, $\forall \tau \in \mathcal{T}$ for some function $M : \mathbb{R}^{d_w} \to [0, \infty)$ and (e) $\mathbb{E}_F M^{2+\delta}(W_i) \leq C$.

Assumption PS2. For all sequences $\{(\theta_n, F_n) \in \mathcal{F}^+ : n \geq 1\}$, the triangular array of processes

$$\left\{\frac{\tilde{m}(W_{n,i}, \theta_n, \tau)}{\sigma_{F_n}(\theta_n)} : \tau \in \mathcal{T},\ i \leq n,\ n \geq 1\right\}$$

is manageable with respect to the envelopes $\{M(W_{n,i}) : i \leq n,\ n \geq 1\}$ where $\{W_{n,i} : i \leq n,\ n \geq 1\}$ is a rowwise i.i.d. triangular array with $W_{n,i} \sim F_n$ for all $i \leq n,\ n \geq 1$.

Following the choice of D. Andrews and Shi (2017) for the Beresteanu, I. Molchanov, and Molinari (2011) example, we choose $\sigma_F = 1$ for simplicity. Thus, assumptions PS1 (a)-(c) are fullfilled by assumption. For (d) we can bound the support function as follows

$$|h(Q_{\theta,z_i,i}(g), u)| \leq \max_{y_i \in [\,]_i, x_i \in [\,]_i} |\sum_{k \geq 1} m(w_i, \theta)\tau_k(x)| \leq C_1 \max_{y_i \in [\,]_i, x_i \in [\,]_i} \tau_k(x) \leq C_1$$

thus it follows that $\tilde{m}(W_{n,i}, \theta_n, \tau) \leq C_1$, so (d) is satisfied using $M(w) = C_1$ and (e) holds, completing the verification of PS1.

Given the choice $\sigma_F = 1$, it remains to show

$$\left\{ \tilde{m}(W_{n,i}, \theta_n, \tau) : \tau \in \mathcal{T}_n,\ i \leq n,\ n \geq 1 \right\}$$

is manageable in the above sense.

Remark. Note, this is incomplete but provides the structure for the proof. We have to check also Assumption PS3 in D. Andrews and Shi (2017).