Esteban M. Aucejo, Federico A. Bugni and V. Joseph Hotz

Identification and inference on regressions with missing covariate data

Article (Accepted version)
(Refereed)

Original citation:
Aucejo, Esteban M., Bugni, Federico A. and Hotz, V. Joseph (2017) *Identification and inference on regressions with missing covariate data.* Econometric Theory, 33 (1). pp. 196-241. ISSN 0266-4666

DOI: 10.1017/S0266466615000250

Identification and Inference on Regressions with Missing Covariate Data[*]

Esteban M. Aucejo
Department of Economics
London School of Economics
e.m.aucejo@lse.ac.uk

Federico A. Bugni
Department of Economics
Duke University
federico.bugni@duke.edu

V. Joseph Hotz
Department of Economics
Duke University, NBER, and IZA
hotz@econ.duke.edu

June 12, 2015

[*] We thank useful comments and suggestions from Arie Beresteanu, Ivan Canay, Shakeeb Khan, and the participants at the presentations in the 2010 World Congress in Shanghai, the 2010 Triangle Econometrics Conference, the 2011 Joint Statistical Meetings in Miami, and at the Statistics Seminar at Duke University. Bugni thanks the National Science Foundation for research support via grant SES-1123771. We also thank the co-editors and the two anonymous referees for several comments and suggestions that have significantly improved this paper. Takuya Ura provided excellent research assistance. Any and all errors are our own.

Abstract

This paper examines the problem of identification and inference on a conditional moment condition model with missing data, with special focus on the case when the conditioning covariates are missing. We impose no assumption on the distribution of the missing data and we confront the missing data problem by using a worst case scenario approach.

We characterize the sharp identified set and argue that this set is usually too complex to compute or to use for inference. Given this difficulty, we consider the construction of outer identified sets (i.e. supersets of the identified set) that are easier to compute and can still characterize the parameter of interest. Two different outer identification strategies are proposed. Both of these strategies are shown to have non-trivial identifying power and are relatively easy to use and combine for inferential purposes.

Keywords: Missing Data, Missing Covariate Data, Partial Identification, Outer Identified Sets, Inference, Confidence Sets.

JEL Classification Codes: C01, C10, C20, C25.

Running head: Regressions with Missing Covariate Data.

1 Introduction

The problem of missing data is ubiquitous in empirical social science research. When survey data is used to estimate an econometric model, researchers are often faced with a dataset that has missing observations. This paper examines the problem of identification and inference in a conditional moment equality model with missing data, with special focus on the case when the conditioning covariates are missing.

Our econometric model is as follows. We are interested in the true parameter value θ_0 that belongs to a parameter space $\Theta \subseteq \mathbb{R}^{d_\theta}$ and satisfies the following *(conditional) moment conditions*:

$$E_F[m(X, Y, \theta_0)|X = x] = \mathbf{0} \ \forall x \ F\text{−a.s.} \tag{1.1}$$

where $(\Omega, \mathcal{A}, F)$ denotes the underlying probability space, $Y : \Omega \to \mathbb{R}^{d_y}$ denotes the *outcome variables*, $X : \Omega \to \mathbb{R}^{d_x}$ denotes the *conditioning variables* or *covariates*, and $m : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \times \mathbb{R}^{d_\theta} \to \mathbb{R}^{d_m}$ is a known structural function. Throughout this paper, a variable is a covariate if it is part of the conditioning variables in Eq. (1.1) and is an outcome variable otherwise. Models characterized by Eq. (1.1) have been studied extensively in the econometrics literature, as we illustrate towards the end of this section.

We explore identification and inference of θ_0 characterized by Eq. (1.1) in the presence of missing data. In practice, the missing data problem affects both outcomes and covariates. From the perspective of identification analysis, missing outcome data and missing covariate data are very different problems, and the former one has been extensively studied in the literature. Therefore, the main text of this paper focuses on the case when only conditioning covariates are missing. In the appendix of the paper, we extend our results to allow for arbitrary missing data patterns on both outcome variables and covariates.

We confront the missing data problem by using a *worst case scenario approach*. This approach allows us to extract the information about θ_0 from the observed data without imposing (untestable) assumptions on the (unobserved) distribution of missing data. Under a worst case scenario approach to missing data, θ_0 is typically partially identified, i.e., the restrictions of the model do not necessarily restrict θ_0 to a unique value, but rather they constrain it to belong to an identified set.

According to our results, the identified set of θ_0 in the presence of missing covariate data is an extremely complex object to characterize and this naturally leads to an even more complicated inferential problem. To the best of our knowledge, the partial identification literature has not been able to address the problem of identification and inference of θ_0 characterized by Eq. (1.1) in the presence of missing covariate data. This paper is an attempt to fill this gap in the literature. Given the complications in dealing with the identified set, we consider several methods to construct supersets of this identified set, referred to as *outer identified sets*, which are relatively simple to compute. In particular, all outer identified sets proposed in this paper take the form of collection of moment inequalities and are thus amenable to inference using the current techniques in the partial identification literature.

The remainder of this paper is organized as follows. Section 1.1 collects several motivating examples and Section 1.2 reviews the related literature. Section 2 introduces our econometric model, characterizes the (sharp) identified set, and explains why it is extremely complex to compute or use for inference. This complexity justifies the construction of simple outer identified sets to characterize the parameter of interest, developed in Section 3. Section 4 proposes a methodology to construct confidence sets of these outer identified sets. Section 5 presents Monte Carlo simulations. Section 6 concludes the paper. The appendix of the paper collects most of the proofs and intermediate results. Finally, several proofs can be found in the online supplementary material to this paper (see Aucejo et al. (2015b)).

1.1 Examples

In order to illustrate the theoretical framework of this paper, we now relate it to econometric models that are routinely used in empirical applications.

Example 1.1 (Mean regression model)**.** *Consider the following econometric model:*

$$Y = f(X, \theta_0) + \varepsilon, \tag{1.2}$$

where $Y : \Omega \to \mathbb{R}$ denotes the outcome variable, $X : \Omega \to \mathbb{R}^{d_x}$ are the conditioning covariates, $\theta_0 \in \Theta \subseteq \mathbb{R}^{d_\theta}$ is the parameter of interest, $f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_\theta} \to \mathbb{R}$ is a known regression function for the conditional mean, i.e.,

$$f(X, \theta) = E_F[Y|X, \theta],$$

and $\varepsilon : \Omega \to \mathbb{R}$ is a mean independent error term with its mean normalized to zero, i.e.,

$$E_F[\varepsilon|X = x] = 0 \ \forall x \ F\text{−a.s.} \tag{1.3}$$

This model can be equivalently re-written as in Eq. (1.1) with $m(X, Y, \theta) \equiv Y - f(X, \theta)$.

For illustration purposes, we give special attention to the linear index regression model, in which $f(X, \theta) = G(X'\theta)$ for a known weakly increasing function $G : \mathbb{R} \to \mathbb{R}$. As special cases, this model includes the linear regression model (i.e. G is the identity function) and limited dependent binary choice models, such as probit or logit (i.e. G is the standard normal or the logistic CDF, respectively).

The mean regression model in Example 1.1 is arguably one of the most commonly used empirical frameworks. For example, it constitutes the basis of the related empirical application in Aucejo et al. (2015a), which we now briefly describe. Prior to the year 1998, the campuses in the University of California system were allowed to use affirmative action criteria in their admissions procedures. However, starting in 1998, a ban on affirmative action was mandated with the passage of Proposition 209, also known as the California Civil Rights Initiative. The objective of Aucejo et al. (2015a) is to estimate the effect of the ban on graduation rates for under-represented minorities. To achieve this goal, we use a random sample of students to estimate a probit version of Eq. (1.2), given by

$$Y = G(\theta_{0,0} + \theta_{0,1}R + \theta_{0,2}P209 + \theta_{0,3}(P209 \times R) + \theta_{0,4}Z) + \varepsilon, \tag{1.4}$$

where Y is a binary indicator of whether the student graduated (or not), R is an indicator of the student's minority status, $P209$ is a binary indicator of whether the student enrolled after the passage of Proposition 209 (or not), and Z is a vector of control variables considered in college admissions decisions, such as measures of the student's academic qualifications and family background characteristics (e.g. parental income, etc.).

The main identification problem in the estimation of θ_0 in Eq. (1.4) is that the covariate vector has a significant amount of missing data, both in its discrete components (e.g. race) and in its continuous components (e.g. parental income). Moreover, the conjectured reasons for the missing observations are varied and complex, making it implausible that these data are actually "missing at random".[1]

Example 1.2 (Quantile regression model)**.** *For some $\alpha \in (0, 1)$, consider the following econometric model:*

$$q_\alpha[Y|X] = f(X, \theta_0),$$

where $q_\alpha[Y|X]$ denotes the α-quantile of $\{Y|X\}$, $Y : \Omega \to \mathbb{R}$ denotes the outcome variable, $X : \Omega \to \mathbb{R}^{d_x}$ are the conditioning covariates, $f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_\theta} \to \mathbb{R}$ denotes a known regression function, $\theta_0 \in \Theta \subseteq \mathbb{R}^{d_\theta}$ is the parameter of interest. This model can be equivalently re-written as in condition (1.1) with $m(X, Y, \theta) \equiv 1[Y - f(X, \theta) \leq 0] - \alpha$.

Example 1.2 is the result of considering Example 1.1 but with the zero moment condition in Eq. (1.3) replaced by the zero quantile condition $q_\alpha(\varepsilon|X = x) = 0\ \forall x\ F-$a.s. In this sense, any empirical illustration that serves as motivation of Example 1.1 could also motivate Example 1.2 as long the modeling objective shifts from the conditional mean to conditional quantile. For the sake of illustration, consider the empirical application in Abrevaya (2001), who studies the effect of demographic characteristics (e.g. age, race, etc.) and maternal behavior during pregnancy (e.g. prenatal care, smoking, etc.) on the quantiles of the birthweight distribution (among other outcome variables). The paper uses U.S. data from the Natality Data Set and suffers from significant amounts of missing covariate data. In particular, data from California, Indiana, New York, and South Dakota were excluded from Abrevaya (2001) as they were missing key covariates such as smoking behavior of the mother (see Abrevaya (2001, Page 250)).

Example 1.3 (Simultaneous equations model)**.** *Consider an econometric model in which two or more outcome variables are simultaneously determined through a system of equations. For example, consider the following two equation case:*

$$\begin{aligned} Y_1 &= f(Y_2, X_1, X_2, \theta_0) + \varepsilon_1, \\ Y_2 &= f(Y_1, X_2, X_1, \theta_0) + \varepsilon_2, \end{aligned}$$

where $Y = (Y_1, Y_2) : \Omega \to \mathbb{R}^2$ denotes the outcomes variables, $X = (X_1, X_2) : \Omega \to \mathbb{R}^{d_x}$ denotes exogenous covariates, $f : \mathbb{R} \times \mathbb{R}^{2d_x} \times \mathbb{R}^{d_\theta} \to \mathbb{R}$ denotes a known regression function, $\theta_0 \in \Theta \subseteq \mathbb{R}^{d_\theta}$ is the parameter of interest, and $\varepsilon = (\varepsilon_1, \varepsilon_2) : \Omega \to \mathbb{R}^2$ denotes mean independent error terms with their means normalized to zero, i.e., $E_F[\varepsilon|X = x] = \mathbf{0}\ \forall x\ F-$a.s.

To illustrate Example 1.3, we can consider the empirical illustration in Lundberg (1988), who analyzes the labor supply decision of couples in a household using a simultaneous equations model. In this application, the outcome variables are hours worked by each family member and covariates include market wages and other family income. Lundberg (1988) estimates the model with data from the Denver Income Maintenance Experiment, which have a non-trivial amount of missing observations (see, e.g., Lundberg (1988, Table A.1)).

1.2 Literature review

There is a vast literature on identification and inference under missing data. However, the worst case scenario approach to missing information is relatively recent and intimately related with the development of partial identification. An excellent summary of this literature can be found in Manski (2003, Chapter 3).

Horowitz and Manski (1998) were the first to consider the identification problem of jointly missing outcome and covariate variables using worst case scenario analysis. They provide sharp bounds on $E_F[g(Y)|X \in A]$ for a known function g and a set A when either Y or X are missing. In general, $E_F[g(Y)|X \in A]$ is shown to be partially identified. By generalizing this result, one could use their analysis to provide sharp bounds for the parameter value θ_0 that satisfies $E_F[m(X, Y, \theta_0)|X \in A] = \mathbf{0}$ for any A when either Y or X are missing. While helpful for our analysis, this generalization does not characterize the set implied by our conditional moment restriction in Eq. (1.1). This is because Eq. (1.1) implies that the true parameter value satisfies the restriction $E_F[m(X,Y,\theta_0)|X \in A] = \mathbf{0}$ *simultaneously* for all A, rather than for a single A. In other words, if we were to select any set A and consider parameter values that satisfy $E_F[m(X,Y,\theta_0)|X \in A] = \mathbf{0}$, we might be losing a large amount of identifying power. In related work, Horowitz and Manski (2000) examine the case where the outcome variable Y is binary and consider partial identification of $P_F[Y=1|X=x]$ and $P_F[Y=1|X=x] - P_F[Y=1|X=\tilde{x}]$ for any x and $\tilde{x}$ when both Y and X are allowed to be missing. As in their earlier work, Horowitz and Manski (2000) consider identification conditional on a pair $(x,\tilde{x})$, while we are interested in a conditional moment restriction that *simultaneously* holds for all pairs $(x,\tilde{x})$.

Manski and Tamer (2002) consider the problem of inference on regressions with interval valued covariates. Since missing data can be considered a special case of interval data, one might hope that their methodology can be used to analyze our problem. Unfortunately, in our context, the assumptions imposed by Manski and Tamer (2002) imply that the data are missing at random, which we purposely want to avoid. We now explain this point using their notation. Let the covariates be (X,V), where V is subject to missing data and assumed to belong to $[V_0,V_1]$, let Z denote a variable that indicates if the covariate V is missing, and let v^L and v^H denote the logical lower and upper bounds of V. Missing covariate data implies that V can be either observed (i.e. $Z=0$ and, so, $V=V_0=V_1$) or unobserved (i.e. $Z=1$ and, so, $V_0=v^L<v^H=V_1$). According to this setup, $Z=0$ occurs if and only if $V_0=V=V_1$. On the other hand, their Mean Independence (MI) assumption implies that:

$$E_F[Y|X=x,V=v] = E_F[Y|X=x,V=v,V_0=v_0,V_1=v_1] \quad \forall(x,v,v_0,v_1),$$

By applying this assumption to any (x,v) such that $v_0=v_1=v$, it follows that:

$$\begin{aligned} E_F[Y|X=x,V=v] &= E_F[Y|X=x,V=v,V_0=v,V_1=v] \\ &= E_F[Y|X=x,V=v,\{V_0=V=V_1\}] \\ &= E_F[Y|X=x,V=v,Z=0] \end{aligned}$$

and, so, their MI assumption applied to the current setup implies that the data are missing at random.

In related work, Horowitz and Manski (2006) (HM06 hereafter), Horowitz et al. (2003), and Beresteanu et al. (2011, Section 4) (BMM11 hereafter) consider identification and inference of best linear predictions (BLPs) under squared loss in the presence of incomplete data, i.e., missing observations and/or interval-valued measures. We now briefly characterize these papers and discuss how they differ from our contribution. Under (unconditional) expected square loss, the BLP of $\{Y|X=x\}$ is equal to $x'\theta_0$, where θ_0 satisfies:

$$E_F[X\varepsilon] = \mathbf{0}, \quad \text{where } \varepsilon \equiv Y - X'\theta_0. \tag{1.5}$$

If $E_F[XX']$ is non-singular, Eq. (1.5) implies that θ_0 is uniquely identified and given by:

$$\theta_0 = E_F[XX']^{-1}E_F[XY]. \tag{1.6}$$

As HM06 and Horowitz et al. (2003) point out, the expression on the right hand side of Eq. (1.6) is not identified under missing covariate data because neither $E_F[XX']$ nor $E_F[XY]$ is identified. By discretizing the distribution of the covariates and imposing logical bounds, these papers develop worst case scenario bounds on θ_0. While these sharp bounds are conceptually easy to understand, they can be very computationally challenging to calculate or estimate. In particular, HM06 (Page 457) suggest that easier-to-compute outer bounds might be considered for this class of problems and that "further research is needed to assess the usefulness of these and other outer bounds in practice". In response to this challenge, BMM11 use the support function approach to conduct computationally feasible sharp inference in a broad class of incomplete data models, including the aforementioned BLP problem when data of $\{X, Y\}$ are interval valued (or, in our case, missing).

It is important to stress that the BLP problem with missing data considered by HM06 or BMM11 differs from the econometric problem of interest in this paper. To see this, consider the mean regression function model in Eq. (1.2) of Example 1.1 when $f(X, \theta_0) = X'\theta_0$. While HM06 and BMM11 assume that the residual in their BLP problem satisfies the *unconditional* moment restriction in Eq. (1.5), we assume instead that the residual satisfies the stronger *conditional* moment restriction in Eq. (1.3). The reason for considering conditional moment restrictions instead of the unconditional ones is two-fold. On the one hand, as Section 1.1 illustrates, there are numerous empirical applications which are modeled as conditional moment restrictions. In fact, this is precisely the case in the probit model with missing covariates in Aucejo et al. (2015a) that motivated this work. Second, unlike its unconditional counterpart, the conditional moment restriction model with missing conditioning covariate has received less attention in the literature. In particular, neither HM06 nor BMM11 can be used to study the identified set of interest in this paper.[2]

In the absence of missing data, θ_0 is typically (point) identified under either conditional or unconditional moment restrictions (in both cases, θ_0 is as in Eq. (1.6)). In the presence of missing covariate data, however, the identification of θ_0 under conditional or unconditional moment restrictions can produce different answers. The intuition behind this is simple. The unconditional moment restriction in Eq. (1.5) implies a finite number of unconditional moment restrictions, which typically lead to a strictly partially identified set for θ_0, which we denote by $\Theta_I^{unc}(F)$.[3] On the other hand, imposing the conditional moment restriction proposed in this paper, i.e., Eq. (1.3), implies simultaneously imposing a conditional moment condition of the form:

$$E_F[\varepsilon g(X)] = \mathbf{0}, \tag{1.7}$$

for every (measurable) function g, which include the unconditional moment conditions in Eq. (1.5) plus an infinite number of additional ones. We show in this paper that this can also lead to a strictly partially identified set for θ_0, which we denote by $\Theta_I^{cond}(F)$. By the law of iterated expectations, $\Theta_I^{cond}(F) \subseteq \Theta_I^{unc}(F)$ and, thus, $\Theta_I^{unc}(F)$ based on Eq. (1.5) can result in a superset of the identified set for the parameter θ_0 that is of interest to this paper. As we explain in Section 2, the computational complexity of $\Theta_I^{cond}(F)$ will force us to produce inference on a superset of the identified set, which we denote by $\Theta_S(F)$. Since the set $\Theta_I^{unc}(F)$ and our set $\Theta_S(F)$ are different supersets of the identified set $\Theta_I^{cond}(F)$, our contribution can be considered to be complementary to the existing literature.

In other work, BMM11 (Section 3) and Galichon and Henry (2011) consider the identification problem in economic games with possible multiplicity of equilibria. While this setup differs considerably from ours, their unobserved equilibrium selection rule plays an analogous role to the distribution of missing covariates in our framework. Galichon and Henry (2011) show that their identified set is characterized by the so-called core of the generalized likelihood predicted by the model and, consequently, the problem of identification reduces to checking whether a collection of conditional moment inequalities is satisfied or not for each parameter value. This collection is relatively easy to handle when the support of the outcome variable is finite and small, but can be computationally very challenging when the outcome variable takes a large number of values. In terms of our paper, their method would be hard to implement whenever the support of the missing covariates has numerous values. Since this last case is relevant in applications, we consider our contribution to be complementary to Galichon and Henry (2011).

There are several other papers that can also be related to our identification problem. Horowitz and Manski (1995) study the problem of corrupted and contaminated data. Lewbel (2002), Mahajan (2006), and Molinari (2008) study identification of the parameter of interest when there is misclassification error of a categorical covariate data. Finally, Chesher et al. (2013) and Chesher and Rosen (2014a,b) develop sharp identification results for instrumental variable models that can be related to our framework.

2 Identification analysis

We now present the identification analysis for models characterized by conditional moment restrictions that contain missing data. We begin with an assumption that formally characterizes our econometric framework.

Assumption A.1. Let the following conditions hold.

(i) Let $(\Omega, \mathcal{A}, F)$ be the probability space of (X, Y, W), let $Y : \Omega \to S_Y \subseteq \mathbb{R}^{d_y}$ be the outcome variables, let $X = (X_1, X_2) : \Omega \to S_{X_1} \times S_{X_2} \equiv S_X \subseteq \mathbb{R}^{d_x} = \mathbb{R}^{d_1+d_2}$ be the covariates, where X_1 is always observed and X_2 is subject to missing data, and let $W : \Omega \to \{0, 1\}$ denote the binary random variable that takes value 1 if X_2 is unobserved, and 0 otherwise.

(ii) There is a *known* function $m : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \times \mathbb{R}^{d_\theta} \to \mathbb{R}^{d_m}$ such that the true parameter value $\theta_0 \in \Theta \subseteq \mathbb{R}^{d_\theta}$ satisfies Eq. (1.1), i.e., $E_F[m(X, Y, \theta_0)|X = x] = \mathbf{0}\ \forall x\ F-$a.s.

Assumption A.1 characterizes the structure of the data and its "missingness". According to Assumption A.1(i), the covariate vector X has two parts, X_1 and X_2, and only X_2 is subject to missing data. As mentioned earlier, the appendix extends all of our results to allow for arbitrary missing data patterns on both X and Y. Assumption A.1(ii) restates the conditional moment restriction in Eq. (1.1).

By definition, the *sharp identified set* of θ_0 is the smallest subset of the parameter space Θ that is consistent with our assumptions. For a given distribution of the data F, this identified set is denoted by $\Theta_I(F)$ and is characterized by the following result.

Lemma 2.1 (Identified set)**.** *Assume Assumption A.1. Then,*

$$\Theta_I(F) \equiv \left\{ \begin{array}{l} \theta \in \Theta \text{ s.t. } \exists g_1 : \mathbb{R}^{d_x} \to \mathbb{R}_+ \text{ and } g_2 : \mathbb{R}^{d_y} \times \mathbb{R}^{d_x} \to \mathbb{R}_+ \text{ that satisfy:} \\ \text{(i) } g_1(x) = 0\ \forall x \notin S_X \text{ and } g_2(y, x) = 0\ \forall (y, x) \notin S_Y \times S_X \\ \text{(ii) } \int_{x_2} g_1(x) dx_2 = 1\ \forall x_1 \in \mathbb{R}^{d_1}\ F-\text{a.s.} \\ \text{(iii) } \int_y g_2(y, x) dy = 1\ \forall x \in \mathbb{R}^{d_x}\ (F, g_1)-\text{a.s.} \\ \text{(iv) } \int_{x_2} g_2(y, x) g_1(x) dx_2 = dP_F\ \forall (x_1, y) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_y}\ F-\text{a.s.} \\ \text{(v) } \left\{ \begin{array}{l} \left(\int_y m(x, y, \theta) g_2(y, x) dy\right) g_1(x) dP_F[X_1 = x_1|W = 1] P_F[W = 1] + \\ E_F[m(X, Y, \theta)|X = x, W = 0] dP_F[X = x|W = 0] P_F[W = 0] = \mathbf{0} \\ \forall x \in \mathbb{R}^{d_x}\ (F, g_1)-\text{a.s.} \end{array} \right\} \end{array} \right\}, \quad (2.1)$$

where dP_F denotes the probability distribution function that induces P_F.

The presence of missing covariate data implies that the following two distributions are unobserved: $dP_F[X_2 = x_2|X_1 = x_1, W = 1]$ and $dP_F[Y = y|X = x, W = 1]$. Nevertheless, the conditional moment restriction in Eq. (1.1) imposes restrictions on these two unknown distributions that are specified in Lemma 2.1, where $g_1(x)$ represents $dP_F[X_2 = x_2|X_1 = x_1, W = 1]$ and $g_2(x, y)$ represents $dP_F[Y = y|X = x, W = 1]$.

The identified set described by Lemma 2.1 is typically extremely complicated to compute in practice. In particular, in order to determine whether a specific parameter value belongs to the identified set (or not) we need to prove (or disprove) the existence of a pair of functions (g_1 and g_2) that satisfies certain properties. These functions need to satisfy a possibly large (even uncountable) number of integral restrictions (i.e. conditions (ii)-(iv) in Eq. (2.1)), which is a challenging mathematical problem. This identified set appears to be more complex than the traditional moment inequalities and equalities considered by the standard literature of partially identified econometric models.

To illustrate the complexity in computing the identified set described by Lemma 2.1, consider a special case of the mean regression model in Example 1.1 with a binary outcome, i.e., $Y \in \{0, 1\}$, and a univariate covariate affected by missing data and with finite support, i.e., $X = X_2$ and $S_X = S_{X_2} = \{x_j\}_{j=1}^N$ and $N > 1$. In this simplified setting, Lemma 2.1 implies that:

$$\Theta_I(F) = \left\{ \begin{array}{l} \theta \in \Theta \text{ s.t. } \exists \gamma_1, \gamma_2 \in \mathbb{R}_+^N \text{ with } \gamma_2 \leq \gamma_1 \text{ that satisfy:} \\ \sum_{j=1}^N \gamma_{1,j} = 1, \\ \sum_{j=1}^N \gamma_{2,j} = P_F[Y = 1|W = 1], \\ \quad (P_F[Y = 1|X = x_j, W = 0] - f(x_j, \theta))P_F[X = x_j, W = 0] \\ \quad +(\gamma_{2,j} - \gamma_{1,j} f(x_j, \theta))P_F[W = 1] = 0 \;\; \forall j = 1, \ldots, N. \end{array} \right\}, \tag{2.2}$$

where γ_1 and γ_2 can be related to functions g_1 and g_2 in Lemma 2.1.[4] As a consequence, in order to check whether a parameter value belongs to $\Theta_I(F)$ (or not), we need to solve a linear system of $N + 2$ equations with $2N$ unknowns subject to non-negativity constraints. This is easy to solve when $N = 2$. On the other hand, when $N > 2$, the system of linear equations becomes under-determined, with a degree of indeterminacy that grows as N increases.

The preceding example is arguably the simplest econometric model based on conditional moment conditions with missing covariate data. Nevertheless, the computation of the identified set can be complicated when the support of the missing covariate has a large number of values. Furthermore, this complexity can be shown to increase substantially as the model structure and/or the missing data patterns become richer.

The complexity of the identification problem motivates us to propose simpler ways of characterizing the identified set in the class of models characterized by Assumption A.1. To this end, we propose the use supersets of the identified set or, as they are referred to in the literature, *outer identified sets*.

Definition 2.1 (Outer Identified set). *An outer identified set is a superset of the identified set.*

An outer identified set provides a (possibly non-sharp) characterization of the parameter of interest. By definition, any parameter value that lies outside of the outer identified set also lies outside of the identified set and, thus, can be eliminated as a candidate for the true parameter value. Of course, if an outer identified set is a *strict* superset of the identified set, it must imply some loss of information about the parameter of interest. Nevertheless, given the challenges described earlier, outer identified sets that are easier to compute and use for inference can be an attractive option for applied researchers.

3 Outer Identification analysis

In the next two subsections, we propose outer identification strategies to produce outer identified sets.

3.1 Outer identification analysis using boxes

Our first approach to constructing outer identified sets is to consider the implication of the conditional moment condition in Eq. (1.1) over a specific class of sets that we refer to as *boxes*. In particular, let $B(x,\nu)$ denote a d_x-dimensional box with "center" at $x \in \mathbb{R}^{d_x}$ and "length" $\nu \in \mathbb{R}^{d_x}_{++}$, formally defined as follows:

$$B(x,\nu) \equiv \{\tilde{x} \in \mathbb{R}^{d_x} : \{x_j - \nu_j < \tilde{x}_j \leq x_j + \nu_j\}_{j=1}^{d_x}\}. \tag{3.1}$$

For any arbitrary $\overline{r} \in (0,\infty]$, the conditional moment restriction in Eq. (1.1) implies the following collection of unconditional moment restrictions:

$$E_F[\ m(X,Y,\theta)\ 1[X \in B(x,\nu)]\ ] = \mathbf{0} \quad \forall (x,\nu) \in \mathbb{R}^{d_x} \times (0,\overline{r})^{d_x}. \tag{3.2}$$

In fact, the results in Domínguez and Lobato (2004) and Andrews and Shi (2013) imply that the informational content in the conditional moment restriction in Eq. (1.1) is equivalent to that in the collection of unconditional moment restrictions in Eq. (3.2). The objective of the subsection is to develop an (outer) identification region based on Eq. (3.2).

In the presence of missing covariate data, the unconditional moment restrictions in Eq. (3.2) are not identified for the same reason as in Eq. (1.1), i.e., they depend on the two unknown distributions: $P_F[X_2|X_1, W = 1]$ and $P_F[Y|X, W = 1]$. The outer identification strategy of this section imposes logical bounds for the unobserved distribution for each individual member of this collection.

Before we describe the result, it is necessary to introduce additional notation. According to Assumption A.1, the covariate X has two components; $X_1 \in \mathbb{R}^{d_1}$ which is always observed, and $X_2 \in \mathbb{R}^{d_2}$ which is subject to missing data. Let $B_1(x_1,\nu_1)$ and $B_2(x_2,\nu_2)$ be the d_1 and d_2-dimensional sub-boxes that result from projecting $B(x,\nu)$ along the dimensions of these two types of covariates, formally defined as follows:

$$\begin{aligned} B_1(x_1,\nu_1) &\equiv \{\tilde{x} \in \mathbb{R}^{d_1} : \{x_j - \nu_j < \tilde{x}_j \leq x_j + \nu_j\}_{j=1}^{d_1}\}, \\ B_2(x_2,\nu_2) &\equiv \{\tilde{x} \in \mathbb{R}^{d_2} : \{x_j - \nu_j < \tilde{x}_j \leq x_j + \nu_j\}_{j=d_1+1}^{d_x}\}, \end{aligned} \tag{3.3}$$

where $x_1 \equiv \{x_j\}_{j=1}^{d_1}$, $x_2 \equiv \{x_j\}_{j=d_1+1}^{d_x}$, $\nu_1 \equiv \{\nu_j\}_{j=1}^{d_1}$, and $\nu_2 \equiv \{\nu_j\}_{j=d_1+1}^{d_x}$, for any $(x,\nu) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_x}_{++}$. With this notation in place, we are ready to state our first outer identified set for θ_0.

Theorem 3.1. *Assume Assumption A.1 and choose $\overline{r} \in (0,\infty]$ arbitrarily. Let $Z \equiv (Y, X_1, (1-W)X_2, W)$*

and let $M_1(Z,\theta,x,\nu) \equiv \{M_{1,j}(Z,\theta,x,\nu)\}_{j=1}^{d_m}$ *with*

$$M_{1,j}(Z,\theta,x,\nu) \equiv$$
$$\left[\begin{array}{c} -\left(\begin{array}{l} \inf\limits_{(\xi_2,y)\in\{S_{X_2}\cap B_2(x_2,\nu_2)\}\times S_Y} m_j((X_1,\xi_2),y,\theta)\ 1[S_{X_2}\cap B_2(x_2,\nu_2)\neq\emptyset, X_1\in B_1(x_1,\nu_1), W=1] \\ +m_j(X,Y,\theta)\ 1[X\in B(x,\nu),W=0] \end{array}\right), \\ \left(\begin{array}{l} \sup\limits_{(\xi_2,y)\in\{S_{X_2}\cap B_2(x_2,\nu_2)\}\times S_Y} m_j((X_1,\xi_2),y,\theta)\ 1[S_{X_2}\cap B_2(x_2,\nu_2)\neq\emptyset, X_1\in B_1(x_1,\nu_1), W=1] \\ +m_j(X,Y,\theta)\ 1[X\in B(x,\nu),W=0] \end{array}\right) \end{array}\right] \tag{3.4}$$

for all $(\theta,x,\nu,j)\in\Theta\times\mathbb{R}^{d_x}\times\mathbb{R}_{++}^{d_x}\times\{1,\ldots,d_m\}$, *and where* B, B_1, *and* B_2 *are defined as in Eqs.* (3.1) *and* (3.3). *Consider the following set:*

$$\Theta_{S_1}(F)\equiv\left\{\theta\in\Theta:\ E_F[M_1(Z,\theta,x,\nu)]\geq\mathbf{0}\ \ \forall(x,\nu)\in\mathbb{R}^{d_x}\times(0,\bar{r})^{d_x}\right\}. \tag{3.5}$$

Then, $\Theta_{S_1}(F)$ *is an outer identified set, i.e.,* $\Theta_I(F)\subseteq\Theta_{S_1}(F)$.

The outer identified set $\Theta_{S_1}(F)$ in Theorem 3.1 is the result of imposing logical bounds on unobserved terms of each member of the collection of unconditional moment restrictions in Eq. (3.2). These bounds are logically possible from the point of view of each element of the collection, but may not be from the point of view of the collection as a whole. In fact, the connection between elements of the collection of unconditional moment restrictions is the main contributing factor to the complexity of the sharp identified set $\Theta_I(F)$. In contrast, the outer identified set $\Theta_{S_1}(F)$ takes the form of a collection of unconditional moment inequalities, which makes it amenable to computation and inference.

The computation of $\Theta_{S_1}(F)$ requires minimizing and maximizing $\{m_j((X_1,\xi_2),y,\theta)\}_{j=1}^{d_m}$ with respect to $(\xi_2,y)\in\{S_{X_2}\cap B_2(x_2,\nu_2)\}\times S_Y$ for all values of $(X_1,\theta)\in S_{X_1}\times\Theta$.[5] The difficulty of these operations will depend on the structure of the model under consideration. For example, in the case of a linear index regression version of Example 1.1, i.e., $m(x,y,\theta)=y-G(x'\theta)$ with weakly increasing function $G(\cdot)$ and $S_{X_2}=\mathbb{R}^{d_2}$, these optimization problems have a closed form solution. In particular, if we set $y_L\equiv\inf_{y\in S_Y}y$ and $y_H\equiv\sup_{y\in S_Y}y$,

$$\inf_{(\xi_2,y)\in\{S_{X_2}\cap B_2(x_2,\nu_2)\}\times S_Y} m((X_1,\xi_2),y,\theta)=$$
$$y_L-G\left(X_1'\theta_1+\textstyle\sum_{j=1}^{d_2}\left((x_{2,j}+\nu_{2,j})1[\theta_{2,j}>0]+(x_{2,j}-\nu_{2,j})1[\theta_{2,j}<0]\right)\theta_{2,j}\right),$$

$$\sup_{(\xi_2,y)\in\{S_{X_2}\cap B_2(x_2,\nu_2)\}\times S_Y} m((X_1,\xi_2),y,\theta)=$$
$$y_H-G\left(X_1'\theta_1+\textstyle\sum_{j=1}^{d_2}\left((x_{2,j}-\nu_{2,j})1[\theta_{2,j}>0]+(x_{2,j}+\nu_{2,j})1[\theta_{2,j}<0]\right)\theta_{2,j}\right).$$

3.2 Outer identification analysis by integrating out

Our second approach to constructing outer identified sets is to integrate out the covariates suffering from missing data, thus removing them from the conditioning set. In particular, the conditional moment restriction in Eq. (1.1) implies the following equation:

$$E_F[m(X,Y,\theta_0)|X_1=x_1]=\mathbf{0}\ \ \forall x\ F\text{−a.s.} \tag{3.6}$$

The difference between Eq. (1.1) and Eq. (3.6) lies in the set of covariates each is conditioned on. While Eq. (1.1) conditions on all the covariates, Eq. (3.6) only conditions on the fully observed covariates. Since Eq. (3.6) does not suffer from a missing covariate data problem, we can characterize its informational content by applying a more traditional worst case scenario bounds analysis. As a result, we obtain our second outer identified set for θ_0.

Theorem 3.2. *Assume Assumption A.1 and choose $\bar{r} \in (0, \infty]$ arbitrarily. Let $Z \equiv (Y, X_1, (1-W)X_2, W)$ and let $M_2(Z, \theta, x, \nu) \equiv \{M_{2,j}(Z, \theta, x, \nu)\}_{j=1}^{d_m}$ with*

$$M_{2,j}(Z,\theta,x,\nu) \equiv$$
$$\left[\begin{array}{c} -\inf_{\xi_2 \in S_{X_2}} m_j((X_1,\xi_2),Y,\theta)\ 1[X_1 \in B_1(x_1,\nu_1), W=1] - m_j(X,Y,\theta)\ 1[X_1 \in B_1(x_1,\nu_1), W=0], \\ \\ \sup_{\xi_2 \in S_{X_2}} m_j((X_1,\xi_2),Y,\theta)\ 1[X_1 \in B_1(x_1,\nu_1), W=1] + m_j(X,Y,\theta)\ 1[X_1 \in B_1(x_1,\nu_1), W=0] \end{array}\right] \tag{3.7}$$

for all $(\theta, x, \nu, j) \in \Theta \times \mathbb{R}^{d_x} \times \mathbb{R}_{++}^{d_x} \times \{1, \ldots, d_m\}$ and where B_1 is defined as in Eq. (3.3). Consider the following set:

$$\Theta_{S_2}(F) \equiv \left\{\theta \in \Theta : \ E_F[M_2(Z,\theta,x,\nu)] \geq \mathbf{0}, \ \forall (x,\nu) \in \mathbb{R}^{d_x} \times (0,\bar{r})^{d_x}\right\}. \tag{3.8}$$

Then, $\Theta_{S_2}(F)$ is an outer identified set, i.e., $\Theta_I(F) \subseteq \Theta_{S_2}(F)$.

As explained earlier, the outer identified set $\Theta_{S_2}(F)$ is entirely based on Eq. (3.6). The reason why $\Theta_{S_2}(F)$ might not be a sharp identified set for θ_0 is that, in general, there will be a loss of information in the process of integrating out covariates with missing data that takes us from Eq. (1.1) to Eq. (3.6). As with our first outer identified set, the outer identified set $\Theta_{S_2}(F)$ takes the form of a collection of unconditional moment inequalities, which makes it amenable to computation and inference.

The computation of $\Theta_{S_2}(F)$ requires minimizing and maximizing $\{m_j((X_1,\xi_2),Y,\theta)\}_{j=1}^{d_m}$ with respect to $\xi_2 \in S_{X_2}$ for all values of $(X_1, Y, \theta) \in S_{X_1} \times S_Y \times \Theta$. Once again, the difficulty of these operations will depend on the structure of the model under consideration. For example, in the case of a linear index regression version of Example 1.1, i.e., $m(x,y,\theta) = y - G(x'\theta)$ with weakly increasing function $G(\cdot)$, these optimization problems have a closed form solution. In particular, if we set $x_{2,j}^L \equiv \inf_{x_2 \in S_{X_2}} x_{2,j}$ and $x_{2,j}^H \equiv \sup_{x_2 \in S_{X_2}} x_{2,j}$ for all $j = 1, \ldots, d_2$,

$$\inf_{\xi_2 \in S_{X_2}} m((X_1,\xi_2),Y,\theta) = Y - G\left(X_1'\theta_1 + \textstyle\sum_{j=1}^{d_2} \left(x_{2,j}^L 1[\theta_{2,j} > 0] + x_{2,j}^H 1[\theta_{2,j} < 0]\right)\theta_{2,j}\right),$$

$$\sup_{\xi_2 \in S_{X_2}} m((X_1,\xi_2),Y,\theta) = Y - G\left(X_1'\theta_1 + \textstyle\sum_{j=1}^{d_2} \left(x_{2,j}^H 1[\theta_{2,j} > 0] + x_{2,j}^L 1[\theta_{2,j} < 0]\right)\theta_{2,j}\right).$$

3.3 Summary of outer identification strategies

Sections 3.1 and 3.2 characterize two outer identification strategies for θ_0 under Assumption A.1. It is easy to verify that these outer identification strategies are different and thus provide different restrictions to the parameter of interest. To verify this, consider the linear index regression problem, i.e., $m(x,y,\theta) = y - G(x'\theta)$ with weakly increasing function $G(\cdot)$. In this case, if the outcome variable Y has bounded support and the missing covariate X_2 has unbounded support, then the first outer identified set is informative while the second one is not (i.e. $\Theta_{S_1}(F) \subset \Theta = \Theta_{S_2}(F)$). On the other hand, if the outcome variable Y has unbounded support and the missing covariate X_2 has bounded support, then the previous result is reversed, with the second outer identified set being informative and the first one being uninformative (i.e. $\Theta_{S_2}(F) \subset \Theta = \Theta_{S_1}(F)$).

Both of these outer identified sets take the form of collection of unconditional moment inequalities. As a result, one can easily combine both collections to generate a sharper (i.e. more informative) outer identified set, also defined by a collection of unconditional moment inequalities. This is the content of the next result.

Theorem 3.3. *Assume Assumption A.1 and choose $\overline{r} \in (0,\infty]$ arbitrarily. Let $Z \equiv (Y, X_1, (1-W)X_2, W)$ and*

$$M(Z,\theta,x,\nu) \;=\; [M_1(Z,\theta,x,\nu)', M_2(Z,\theta,x,\nu)']' \tag{3.9}$$

for all $(\theta,x,\nu) \in \Theta \times \mathbb{R}^{d_x} \times (0,\overline{r}]^{d_x}$, where M_1 and M_2 are defined as in Eqs. (3.4) and (3.7), respectively. Consider the following set:

$$\Theta_S(F) \equiv \left\{\theta \in \Theta : \; E_F[M(Z,\theta,x,\nu)] \geq \mathbf{0} \; \forall (x,\nu) \in \mathbb{R}^{d_x} \times (0,\overline{r})^{d_x}\right\}. \tag{3.10}$$

Then, $\Theta_S(F)$ is an outer identified set, i.e., $\Theta_I(F) \subseteq \Theta_S(F)$.

The outer identified set $\Theta_S(F)$ is given by a collection of unconditional moment restrictions that represents both identification strategies. In the remainder of the paper, we use this outer identified set to conduct econometric inference.[6]

4 Inference

The objective of this section is to construct a confidence set, denoted CS_n, that covers the true parameter value θ_0 with an asymptotic confidence size of $(1-\alpha)$ (or more). Given our results in previous sections, it is important to choose an inferential method that allows the parameter of interest to be partially identified.

Following Theorem 3.3, our outer identified set is characterized by an uncountable collection of p-dimensional unconditional moment inequalities with $p \equiv 4d_m$. To the best of our knowledge, our framework does not exactly coincide with the typical econometric model used in the partially identified literature. On the one hand, we have an uncountable number of unconditional moment inequalities, which is not allowed in the standard framework for unconditional moment inequalities.[7] On the other hand, our framework with unconditional moment conditions is not directly captured by any of the existing inference methods for conditional moment inequalities. In any case, our outer identification analysis closely resembles the ideas of Andrews and Shi (2013) (hereafter, referred to as AS13), who also translate conditional moment inequalities into unconditional ones. For this reason, we find it natural to implement inference by adapting the Generalized Moment Selection (GMS, henceforth) method developed by AS13. Although less natural in our setting, one could also have implemented inference by adapting the results from other references in the conditional moment inequality literature, which include Kim (2008), Ponomareva (2010), Armstrong (2012, 2014), Armstrong and Chan (2012), Chetverikov (2012), and Chernozhukov et al. (2013), among others. It is relevant to point out that, to the best of our knowledge, there are no inferential procedures that can be applied to the complex structure of the sharp identified set in Lemma 2.1. In other words, the possibility of conducting inference along the lines of any of these references is a consequence of the simplification introduced by our outer identification strategies.

In order to conduct inference, we assume to observe an i.i.d. sample of observations of $\{Z_i\}_{i=1}^n$ with $Z \equiv (Y, X_1, (1-W)X_2, W)$ distributed according to F. Following the GMS procedure in AS13, we propose to construct a confidence set for θ_0 by hypothesis test inversion, that is, by collecting all parameter values that are not rejected in a hypothesis test with $H_0 : \theta_0 = \theta$ vs. $H_1 : \theta_0 \neq \theta$. In particular, we propose:

$$CS_n \;\equiv\; \{\theta \in \Theta \;:\; T_n(\theta) \leq \hat{c}_n(\theta, 1-\alpha)\},$$

where $T_n(\theta)$ is the Cramér-von Mises test statistic and $\hat{c}_n(\theta, 1-\alpha)$ is the GMS critical value for aforementioned hypothesis test. In the remainder of this section, we specify the components of CS_n and we discuss its main asymptotic properties. For reasons of brevity, several details of this section are deferred to Appendix A.2.

4.1 Definition of the test statistic $T_n(\theta)$

Given the i.i.d. sample of data $\{Z_i\}_{i=1}^n$ and for any parameter $\theta \in \Theta$, the Cramér-von Mises test statistic is defined as follows:

$$T_n(\theta) \equiv \int S\left(\, \sqrt{n}\, \overline{M}_n(\theta, x, \nu),\; \overline{\Sigma}_n(\theta, x, \nu) \,\right) d\mu(x, \nu), \tag{4.1}$$

where $(x, \nu) \in \mathbb{R}^{d_x} \times \mathbb{R}_+^{d_x}$, $\overline{M}_n(\theta, x, \nu)$ denotes the sample mean of $\{M(Z_i, \theta, x, \nu)\}_{i=1}^n$, $\overline{\Sigma}_n(\theta, x, \nu)$ denote a slight modification of the sample variance of $\{M(Z_i, \theta, x, \nu)\}_{i=1}^n$ (see Eq. (4.2)), and S and μ are a function and a probability measure chosen by the researcher according to assumptions in Appendix A.2.2.

According to Theorem 3.3, $\Theta_S(F)$ is composed of parameter values $\theta \in \Theta$ that satisfy a collection of p moment inequalities. Our test statistic replaces these population moment inequalities with their properly scaled sample analogue, $\sqrt{n}\, \overline{M}_n(\theta, x, \nu)$, weights them according to their sample variance, evaluates their value according to the criterion function S, and aggregates them across values of (x, ν) according to the probability measure μ. In the language of the criterion function approach developed by Chernozhukov et al. (2007), $T_n(\theta)$ is the sample analogue of the criterion function, which indicates whether θ belongs to the outer identified set $\Theta_S(F)$ or not. This statement is formalized in Theorem A.3 in Appendix A.2.3.

We proceed by specifying each of the components of the test function $T_n(\theta)$. For any $\theta \in \Theta$ and $(x, \nu) \in \mathbb{R}^{d_x} \times (0, \overline{r})^{d_x}$, the sample mean, the sample variance, and its modified version are defined as follows:

$$\begin{aligned}
\overline{M}_n(\theta, x, \nu) \quad &\equiv \quad n^{-1} \sum_{i=1}^{n} M(Z_i, \theta, x, \nu), \\
\hat{\Sigma}_n(\theta, x, \nu) \quad &\equiv \quad n^{-1} \sum\nolimits_{i=1}^{n} \left[M(Z_i, \theta, x, \nu) - \overline{M}_n(\theta, x, \nu)\right] \left[M(Z_i, \theta, x, \nu) - \overline{M}_n(\theta, x, \nu)\right]', \\
\overline{\Sigma}_n(\theta, x, \nu) \quad &\equiv \quad \hat{\Sigma}_n(\theta, x, \nu) \;+\; \lambda\, D_n(\theta),
\end{aligned} \tag{4.2}$$

where λ is an arbitrarily small positive constant[8] and $D_n(\theta)$ a positive definite diagonal matrix defined in Eq. (A.5) in Appendix A.2.1. The role of the modification is to ensure that we use a measure of sample variance that is positive definite in a scale invariant fashion.

4.2 Definition of the GMS critical value $\hat{c}_n(\theta, 1-\alpha)$

The GMS critical value $\hat{c}_n(\theta, 1-\alpha)$ is an approximation to the $(1-\alpha)$-quantile of the asymptotic distribution of $T_n(\theta)$ under $H_0 : \theta_0 = \theta$. According to AS13 (Section 4.1), this asymptotic distribution is:

$$T(h) \equiv \int S\left(\, v_{h_2}(x, \nu) + h_1(x, \nu)\;,\; h_2(x, \nu) + \lambda\, \mathbf{I}_{p \times p} \,\right) d\mu(x, \nu), \tag{4.3}$$

where $h \equiv (h_1, h_2)$, h_1 indicates the amount of slackness of the moment inequalities, h_2 is the limiting variance-covariance kernel, and v_{h_2} is a mean zero $\mathbb{R}^p$-valued Gaussian process with covariance kernel $h_2(\cdot, \cdot)$.

To define the GMS approximation to the distribution in Eq. (4.3), it is first necessary to define certain auxiliary expressions. For every $\theta \in \Theta$ and $(x,\nu),(\tilde{x},\tilde{\nu}) \in \mathbb{R}^{d_x} \times (0,\overline{r})^{d_x}$, define:

$$\begin{aligned}
\hat{h}_{2,n}(\theta,(x,\nu),(\tilde{x},\tilde{\nu})) &\equiv D_n^{-1/2}(\theta)\ \hat{\Sigma}_n(\theta,(x,\nu),(\tilde{x},\tilde{\nu}))\ D_n^{-1/2}(\theta), \\
\hat{h}_{2,n}(\theta,x,\nu) &\equiv \hat{h}_{2,n}(\theta,(x,\nu),(x,\nu)), \\
x_n(\theta,x,\nu) &\equiv \kappa_n^{-1}\sqrt{n}\ \hat{D}_n^{-1/2}(\theta)\ \overline{M}_n(\theta,x,\nu), \\
\varphi_n(\theta,x,\nu) &\equiv \{B_n \times 1\left[x_{n,j}(\theta,x,\nu) > 1\right]\}_{j=1}^{p}\,, \qquad (4.4)
\end{aligned}$$

where $\{\kappa_n\}_{n\geq 1}$ and $\{B_n\}_{n\geq 1}$ are sequences chosen by the researcher according to the restrictions in Appendix A.2.2. We briefly describe each of these expressions. First, $\hat{h}_{2,n}(\theta,(x,\nu),(\tilde{x},\tilde{\nu}))$ and $\hat{h}_{2,n}(\theta,x,\nu)$ are the standardized versions of the sample variance-covariance kernel and sample variance kernel, respectively. Second, $x_n(\theta,x,\nu)$ is a sample measure of the slackness of the moment inequalities and $\varphi_n(\theta,x,\nu)$ is an increasing function of this measure that is used in the construction of GMS quantiles. With these definitions in place, the GMS critical value is defined as follows:

$$\hat{c}_n(\theta,1-\alpha) \equiv \eta + c(\varphi_n(\theta,\cdot),\hat{h}_{2,n}(\theta,\cdot,\cdot),1-\alpha+\eta),$$

where η is an arbitrarily small positive constant[9] and $c(\varphi_n(\theta,\cdot),\hat{h}_{2,n}(\theta,\cdot,\cdot),1-\alpha+\eta)$ is the (conditional) $(1-\alpha+\eta)$-quantile of the following random variable:

$$\int S\ (\ v_{\hat{h}_{2,n}}(x,\nu) + \varphi_n(\theta,x,\nu)\ ,\ \hat{h}_{2,n}(\theta,x,\nu)\ +\ \lambda\,\mathbf{I}_{p\times p}\ )\ d\mu(x,\nu), \qquad (4.5)$$

and $v_{\hat{h}_{2,n}}$ is a mean zero $\mathbb{R}^p$-valued Gaussian process with covariance kernel $\hat{h}_{2,n}(\theta,\cdot,\cdot)$. The intuition behind the GMS approximation can be understood by comparing Eqs. (4.3) and (4.5). First, the sample analogue variance-covariance kernel $\hat{h}_{2,n}(\theta,\cdot,\cdot)$ estimates the limiting covariance kernel h_2. Second, the empirical slackness measure $\varphi_n(\theta,\cdot)$ approximates the limiting slackness in the moment inequalities h_1.

There are several details regarding the computation of the GMS quantiles from Eq. (4.5). First, the Gaussian process $v_{\hat{h}_{2,n}}$ requires simulation and there are various methods that can be used to implement this. Second, Eqs. (4.1) and (4.5) require integration with respect to the measure μ. All of these approximations can be conducted with arbitrary accuracy by methods described in detail in AS13 (Section 3.5). For the sake of convenience, we include a brief description of these approximation methods in Appendix A.2.4.

4.3 Properties of the GMS confidence sets

The formal results in AS13 suggest that GMS confidence sets provide excellent asymptotic properties. While these results do not immediately apply to our outer identified framework, it is not hard to adapt their arguments in order to establish analogous results. For the sake of completeness, this section announces some of these results and their proofs can be found in the online supplementary material (see Aucejo et al. (2015b)).

In order to discuss formal coverage properties, it is necessary to introduce some basic notation regarding the parameter space. As it is customary in the literature of moment inequality models, one can consider the parameters of the model to be (θ, F), where θ is the finite dimensional parameter of interest and F is the distribution of the data. In order to produce asymptotic results, we restrict these parameters to a *baseline* parameter space, denoted by $\mathcal{F}$ and formally defined in Definition A.1 in Appendix A.2.1. It is worthwhile to point out that the baseline parameter space $\mathcal{F}$ includes both parameter values (θ, F) for which θ satisfies the moment inequalities of our outer identified set (i.e. $\theta \in \Theta_S(F)$) and parameter values $(\tilde{\theta}, \tilde{F})$ for which $\tilde{\theta}$ does not satisfy the moment inequalities of our outer identified set (i.e. $\tilde{\theta} \notin \Theta_S(\tilde{F})$). In order to establish coverage results, we further restrict the baseline parameter space $\mathcal{F}$ to a relevant *null* parameter space, denoted by $\bar{\mathcal{F}}_0$, which imposes the moment inequalities of our outer identified set (among other technical conditions). In other words, $\bar{\mathcal{F}}_0 \subset \mathcal{F}$ and, by definition, $\bar{\mathcal{F}}_0$ is only composed of parameter values (θ, F) such that $\theta \in \Theta_S(F)$. The formal definition of the parameter space $\bar{\mathcal{F}}_0$ is deferred to Definition A.3 in Appendix A.2.1.

We are now ready to introduce our main asymptotic coverage result, which establishes that the GMS confidence set covers each parameter θ in $\Theta_S(F)$ with a limiting probability of $(1-\alpha)$ or more.

Theorem 4.1. *Assume Assumptions A.2, A.5-A.6 and let $\bar{\mathcal{F}}_0$ be as in Definition A.3. Then,*

$$\liminf_{n\to\infty} \inf_{(\theta,F)\in\bar{\mathcal{F}}_0} P_F[\theta \in CS_n] \;\geq\; (1-\alpha). \tag{4.6}$$

There are a couple of relevant aspects in this result that are worth pointing out. First, recall that $(\theta, F) \in \bar{\mathcal{F}}_0$ implies $\theta \in \Theta_S(F)$ and so the coverage of all $(\theta, F) \in \bar{\mathcal{F}}_0$ implies the coverage of all $\theta \in \Theta_S(F)$ for the relevant collection of distributions F. Second, notice that Eq. (4.6) computes limits as $n \to \infty$ *after* considering the infimum of $(\theta, F) \in \bar{\mathcal{F}}_0$. In this sense, the asymptotic coverage result holds *uniformly* over a relevant subset of the parameters $(\theta, F) \in \bar{\mathcal{F}}_0$. According to the literature on partially identified moment inequality models, obtaining uniformly valid asymptotic results is the only way to guarantee that the asymptotic analysis provides an accurate approximation to finite sample results. The reason for this is that the limiting distribution of the test statistic is discontinuous in the slackness of the moment inequalities, while the finite sample distribution of this statistic does not exhibit these discontinuities. In consequence, asymptotic results obtained for any fixed distribution (i.e. pointwise asymptotics) can be grossly misleading, and possibly produce confidence sets that undercover (even asymptotically).[10]

Our next result describes the asymptotic power properties of the GMS confidence set, which shows that the GMS confidence set excludes any fixed θ outside of $\Theta_S(F)$ with probability approaching one.[11]

Theorem 4.2. *Assume Assumptions A.2-A.6, and let $(\theta, F) \in \mathcal{F}$ such that $\theta \notin \Theta_S(F)$. Then,*

$$\lim_{n\to\infty} P_F[\theta \notin CS_n] = 1.$$

By repeating arguments in AS13, it is also possible to show that inference based on GMS confidence sets have non-trivial power against a set of relevant $n^{-1/2}$-local alternatives, and strictly higher power than inference based on alternative methods, such as plug-in asymptotics or subsampling. These results are omitted from the paper for reasons of brevity.

Given the structure of our outer identified set, adapting the GMS inference method developed by AS13 to our problem was the most natural choice. However, as we explained earlier, the recent literature offers several other methods to implement inference for conditional moment inequality models and it is important to understand how the GMS inference method compares to the rest of the literature. In this respect, the literature offers many interesting discussions and we now briefly summarize some of the main ideas. Chernozhukov et al. (2013, page 672) explain that their method and that of AS13 differ in the detection rate of $n^{-1/2}$-local alternatives depending on whether these take the form of a constant deviation from the null hypothesis on a set of positive Lebesgue measure (the so-called flat alternative), or not. On the one hand, the GMS inference method will have non-trivial power against these flat alternatives, while the inference method in Chernozhukov et al. (2013) will not. On the other hand, the inference method in Chernozhukov et al. (2013) presents near optimal detection rates of non-flat alternatives, while the GMS inference method presents sub-optimal rates. Both flat and non-flat alternatives are relevant in applications and so both contributions should be considered complementary. More recently, Armstrong (2012, 2014), Chetverikov (2012), and Armstrong and Chan (2012) propose new inference methods based on optimally weighted Kolmogorov-Smirnov (KS) type statistics. While AS13 also considers KS statistics, their method implicitly impose restrictions on the choice of weights due to technical reasons (See Armstrong (2012) for a comprehensive analysis of the power properties of the AS13). By using novel arguments, these new references show that using an optimal weighted KS statistics can lead to significant power improvements.

We conclude the section by considering the effect of misspecification on inference. By definition, any outer identified set is non-empty if the model is correctly specified, while it may or may not be empty if the model is incorrectly specified. By applying Theorem 4.1, we conclude that a correctly specified model will produce a non-empty confidence set with a limiting probability of $(1-\alpha)$ or more. However, a misspecified model with an empty outer identified set can generate an empty confidence set. These ideas can be used as a basis of model specification in partially identified moment inequality models as in Andrews and Soares (2010, section 5) and Bugni et al. (2015).

5 Monte Carlo simulations

In this section, we illustrate our results using Monte Carlo simulations based on the probit linear regression model in Example 1.1. In this setup, the researcher correctly assumes that the true value of the parameter $\theta_0 = (\theta_{0,1}, \theta_{0,2}) \in \Theta \equiv [-2,2]^2$ satisfies the conditional moment condition:

$$E_F[Y|X=(x_1,x_2)] = \Phi(x_1\theta_1 + x_2\theta_2), \tag{5.1}$$

where $Y \in S_Y = \{0,1\}$ is a binary outcome random variable, $X = (X_1, X_2)$ are the covariates with $X_1 \in S_{X_1} = [0,1]$ that is always observed, and $X_2 \in S_{X_2} = \{0,1\}$ that is subject to missing data. As in the main text, W is a binary variable that indicates whether X_2 is missing. In order to conduct inference, the researcher observes an i.i.d. sample of $\{(Y, X_1, (1-W)X_2, W)\}_{i=1}^n$.

We next discuss aspects of the simulations that are unknown to the researcher. The covariates X_1 and X_2 independently distributed with $X_1 \sim U[0,1]$ and $X_2 \sim Be(0.5)$. The data are missing according to:

$$P_F[W=1|X=(x_1,x_2)] = \pi(x_2) \;\text{ for } x_2 \in \{0,1\}. \tag{5.2}$$

Notice that Eq. (5.2) allows the conditional probability of missing data to be constant (i.e. $\pi(0)=\pi(1)$) or not (i.e. $\pi(0) \neq \pi(1)$). Finally, in all of our simulations, the data that are not missing are also distributed according to a probit regression model with parameter value $\tilde{\theta}_0 = (\tilde{\theta}_{0,1}, \tilde{\theta}_{0,2})$, i.e.,

$$E_F[Y|X=(x_1,x_2), W=0] = \Phi(x_1\tilde{\theta}_1 + x_2\tilde{\theta}_2), \tag{5.3}$$

Notice that Eq. (5.3) allows data to be missing at random (i.e. $\theta_0 = \tilde{\theta}_0$) or not (i.e. $\theta_0 \neq \tilde{\theta}_0$). Finally, we notice that while the researcher correctly assumes Eq. (5.1), he is unaware that Eqs. (5.2) and (5.3) hold.

We consider five Monte Carlo designs that differ in the value of the population parameters. These parameter values are specified in Table 1 and are chosen to illustrate cases with and without missing at random and with and without a constant probability of missing data.

Design	$\theta_{0,1}$	$\theta_{0,2}$	$\tilde{\theta}_{0,1}$	$\tilde{\theta}_{0,2}$	$\pi(0)$	$\pi(1)$	Missing at random	Missing probability
Design 1	0.5	1	0.5	1	0.15	0.15	yes	constant
Design 2	0.5	0.5	0.5	0.5	0.15	0.15	yes	constant
Design 3	0.5	1	0.5	1	0.25	0.15	yes	not constant
Design 4	0.5	1	0.75	1	0.15	0.15	no	constant
Design 5	0.5	1	0.75	1	0.25	0.15	no	not constant

Table 1: Parameter values for Eqs. (5.1), (5.2), and (5.3) used in our simulations. In this framework, data are missing at random if and only if $\theta_0 = \tilde{\theta}_0$, and the missing probability is constant if and only if $\pi(0) = \pi(1)$.

The Monte Carlo setup we consider in our simulations is admittedly simple but is very useful to illustrate the value of our outer identification strategies. Since the outcome is binary and the missing data can only take two possible values, we are in a situation where the identified set is simple enough to be computed analytically (i.e. $N = 2$ in Eq. (2.2)). For each design, we compute the sharp identified set $\Theta_I(F)$, the outer identified set 1, $\Theta_{S,1}(F)$ (Theorem 3.1), the outer identified set 2, $\Theta_{S,2}(F)$ (Theorem 3.2), and the proposed outer identified set, $\Theta_S(F) = \Theta_{S,1}(F) \cap \Theta_{S,2}(F)$ (Theorem 3.3). We arbitrarily choose $\bar{r} = 0.5$ to compute our outer identified sets. By looking individually at outer identified sets 1 and 2, we can understand the identifying power of each outer identification strategy. By comparing the proposed outer identified sets and the sharp identified set, we can quantify the loss of information involved in our outer identification strategies.

We also use the Monte Carlo simulations to implement our inference method. For this purpose, we simulate $2,000$ datasets, each with a sample size of $n = 500$ and, for each of these samples, we construct confidence sets with confidence size $(1 - \alpha) = 90\%$ using two inference methods. First, we construct the GMS confidence set for the outer identified set $\Theta_S(F)$ proposed in Section 4. According to our theoretical results, this confidence set should cover the true parameter value with a limiting probability of $(1 - \alpha)$ or more (Theorem 4.1) and should cover any fixed parameter value outside of the outer identified set with a probability that converges to zero (Theorem 4.2). Second, we construct a confidence set using only the fully observed data under the implicit assumption that the data are missing at random. By using standard arguments, this second inference method can be shown to cover the parameter value $\tilde{\theta}_0$ with a limiting probability of $(1 - \alpha)$ and should cover any other fixed parameter value with a probability that converges to zero. In cases in which the data are missing at random (i.e. $\theta_0 = \tilde{\theta}_0$ as in designs 1-3) the second confidence set is expected to be size correct and have optimal power properties. On the other hand, in cases in which data are not missing at random (i.e. $\theta_0 \neq \tilde{\theta}_0$ as in designs 4-5) the second confidence set is expected to suffer from undercoverage problems.

In order to illustrate our coverage properties, we choose 12 specific parameter values that are purposely chosen in the interior, boundary, and exterior of the outer identified set $\Theta_S(F)$. First, we consider the true parameter value $\theta_{interior} \equiv \theta_0$, which is always located in the interior of $\Theta_S(F)$. Second, we consider the parameter value $\theta_{boundary}$ that is on the boundary of the outer identified set located directly to the east of the true parameter value, i.e.,

$$\theta_{boundary} \equiv (\theta_{0,1} + C, \theta_{0,2}), \tag{5.4}$$

where the constant $C > 0$ is chosen so that $\theta_{boundary}$ lies *exactly* in the boundary of the $\Theta_S(F)$. Next, we consider a list of 10 additional parameter values $\{\theta_{exterior,v}\}_{v=1}^{10}$ chosen according to the following rule:

$$\theta_{exterior,v} \equiv (\theta_{0,1} + C + v/\sqrt{n}, \theta_{0,2}) \quad \text{for} \quad v \in \{1, \ldots, 10\},$$

where $C > 0$ is as in Eq. (5.4). Since $\theta_{boundary}$ lies exactly in the boundary of $\Theta_S(F)$, $\{\theta_{exterior,v}\}_{v=1}^{10}$ lie in the exterior of $\Theta_S(F)$ and at a distance to this set that increases with the index $v = 1, \ldots, 10$.

We conclude by describing the parameters used to implement the GMS method, which is explained in full detail in Appendix A.2.4. We construct GMS confidence sets with a function S given by the modified method of moments (see Eq. (A.6) in Appendix A.2.2) and, following AS13, we specify the measure μ to be uniform distribution using the information regarding the support of the covariates, i.e., $\mu(x, \nu) = \prod_{j=1}^{2} \mu_{1,j}(x_{1,j}) \times \mu_2(\nu_{L,j}) \times \mu_2(\nu_{H,j})$, where $\mu_{1,1}$ is equal to $U(0,1)$, $\mu_{1,2}$ is equal to $Be(0.5)$, and $\{\mu_2(\nu_{L,j}), \mu_2(\nu_{H,j})\}_{j=1,2}$ are all equal to $U(0, \bar{r})$ with $\bar{r} = 0.5$. Every integral is approximated by Monte Carlo integration with $s_n = 1,000$. Following AS13, we choose $\kappa_n = (0.3 \ln(n))^{1/2}$, $B_n = (0.4 \ln(n) / \ln \ln(n))^{1/2}$, and $\eta = 10^{-6}$. Finally, GMS quantiles are computed by simulation using $\tau_{reps} = 1,000$ repetitions.

5.1 Design 1

Figure 5.1 describes the identification analysis in Design 1. It shows the true parameter value θ_0, the identified set, the two outer identified sets, and their intersection. The outer identified set 1 is a relatively large region of the parameter space while the outer identified set 2 is relatively smaller. Neither of these sets is a subset of the other and, consequently, there is an informative gain in considering their intersection. In fact, the size of the intersection of the outer identified sets is slightly larger that the size of the identified set. In other words, in the current setup, the combination of our outer identification strategies captures most of the information that is available in the data.

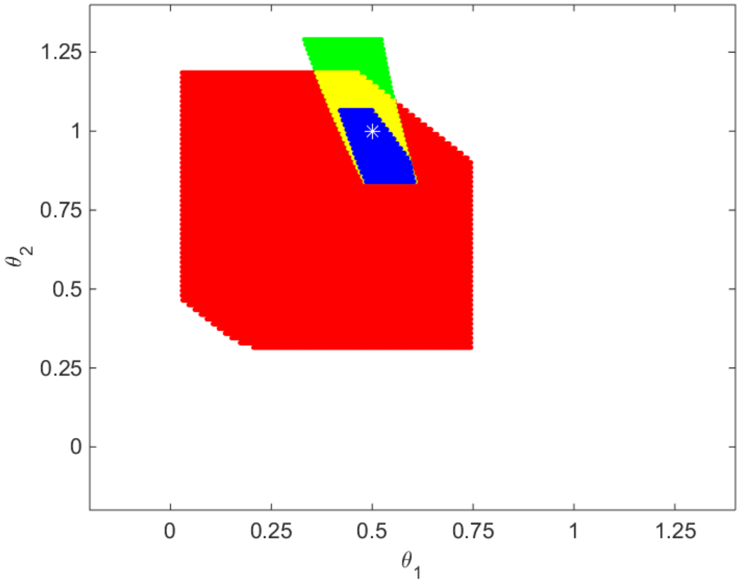

Figure 5.1: Identification analysis in Design 1. The white asterisk is θ_0, the blue region is $\Theta_I(F)$, the yellow region is $\Theta_S(F) \cap \Theta_I(F)^c$, the red region is $\Theta_{S_1}(F) \cap \Theta_S(F)^c$, and the green region is $\Theta_{S_2}(F) \cap \Theta_S(F)^c$.

Figure 5.2 shows coverage probabilities that result from our GMS confidence set and from inference using a probit model based only on fully observed data. As expected, the GMS inference on our outer identified set covers θ_0 and θ_1 (labeled -1 and 0) with probability exceeding $(1 - \alpha)$, and all remaining points are covered less frequently, with coverage frequency decreasing monotonically as we move further away from the outer identified set. Inference based on fully observed data covers θ_0 (labeled -1) with (limiting) probability $(1 - \alpha)$, and all other points are covered less frequently, with coverage dropping monotonically as we move further to the right. Since data are missing at random in Design 1, it is no surprise that inference based on fully observed data is significantly more powerful than inference based on our outer identified set.

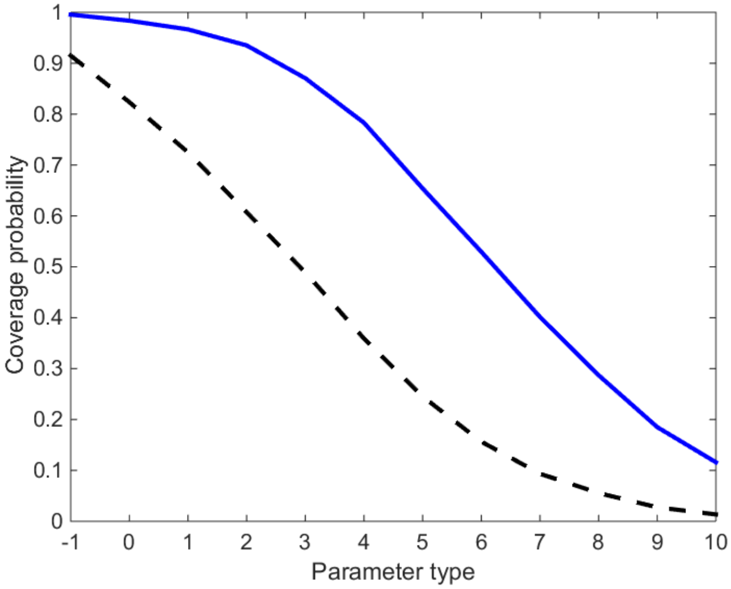

Figure 5.2: Empirical coverage frequency with $(1-\alpha) = 90\%$ for several parameter types in Design 1. Solid line represents coverage with our GMS confidence set and dashed line represents coverage using a probit model based only on fully observed data. Parameter types are as follows: "-1" refers to $\theta_{interior} = \theta_0$, "0" refers to $\theta_{boundary}$, and "1-10" refer to $\{\theta_{exterior,v}\}_{v=1}^{10}$.

5.2 Design 2

Figure 5.3 presents the identification analysis using the parameters in Monte Carlo Design 2. Notice that the first two designs share the fact that the data are missing at random and the probability of missing data is constant. Nevertheless, the results of the identification analysis in these two simulations are very different. In Design 2, the outer identified set 2 is a strict subset of the outer identified set 1 and, as a consequence, the intersection of the outer identified sets coincides with the outer identified set 2. As in Design 1, the intersection of outer identified sets is slightly larger than the sharp identified set and, therefore, captures most of the information that is available in the data. Figure 5.4 presents the inferential results for Design 2. The coverage probabilities are qualitatively and quantitatively similar to Design 1.

5.3 Design 3

Figures 5.5 and 5.6 repeat the analysis using the parameters in Design 3. The purpose of this simulation is to explore the effect of having a larger and non-constant probability of missing data (i.e. $\pi(0) \neq \pi(1)$). Increasing the percentage of missing data enlarges the outer identified sets, leading to a larger intersection of outer identified sets. Nevertheless, the combination of our outer identification strategies is still reasonably close to the sharp identified set. Inferential results are similar to those in previous designs, both qualitatively and quantitatively.

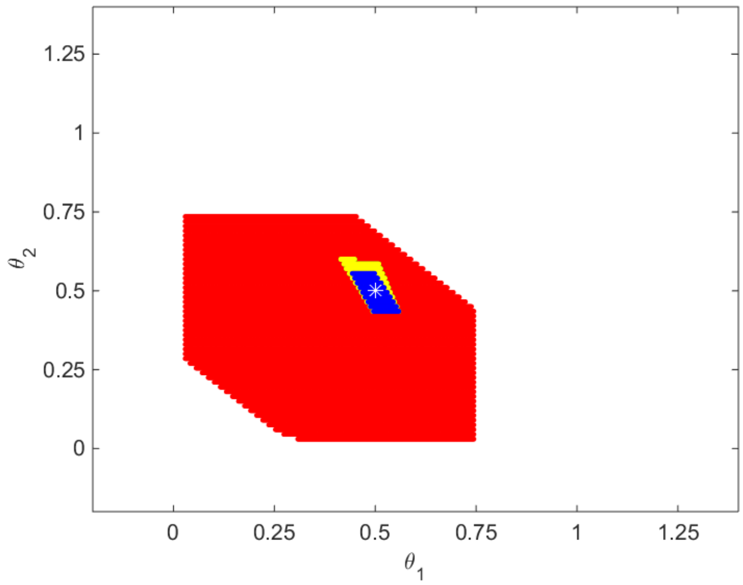

Figure 5.3: Identification analysis in Design 2. The white asterisk is θ_0, the blue region is $\Theta_I(F)$, the yellow region is $\Theta_S(F) \cap \Theta_I(F)^c$, the red region is $\Theta_{S_1}(F) \cap \Theta_S(F)^c$, and the green region is $\Theta_{S_2}(F) \cap \Theta_S(F)^c$.

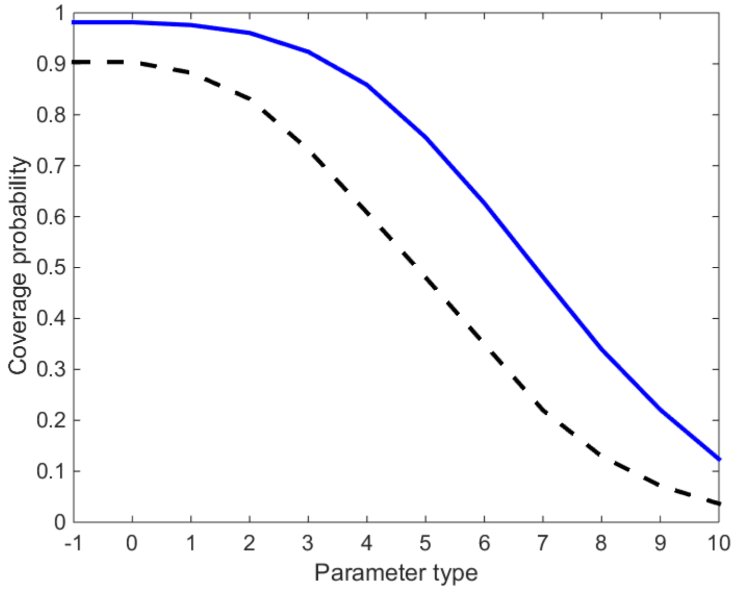

Figure 5.4: Empirical coverage frequency with $(1-\alpha) = 90\%$ for several parameter types in Design 2. Solid line represents coverage with our GMS confidence set and dashed line represents coverage using a probit model based only on fully observed data. Parameter types are as follows: "-1" refers to $\theta_{interior} = \theta_0$, "0" refers to $\theta_{boundary}$, and "1-10" refer to $\{\theta_{exterior,v}\}_{v=1}^{10}$.

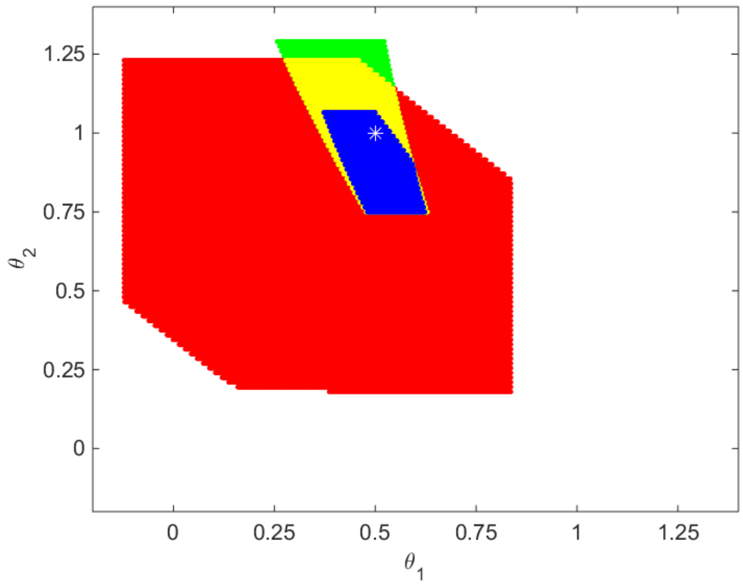

Figure 5.5: Identification analysis in Design 3. The white asterisk is θ_0, the blue region is $\Theta_I(F)$, the yellow region is $\Theta_S(F) \cap \Theta_I(F)^c$, the red region is $\Theta_{S_1}(F) \cap \Theta_S(F)^c$, and the green region is $\Theta_{S_2}(F) \cap \Theta_S(F)^c$.

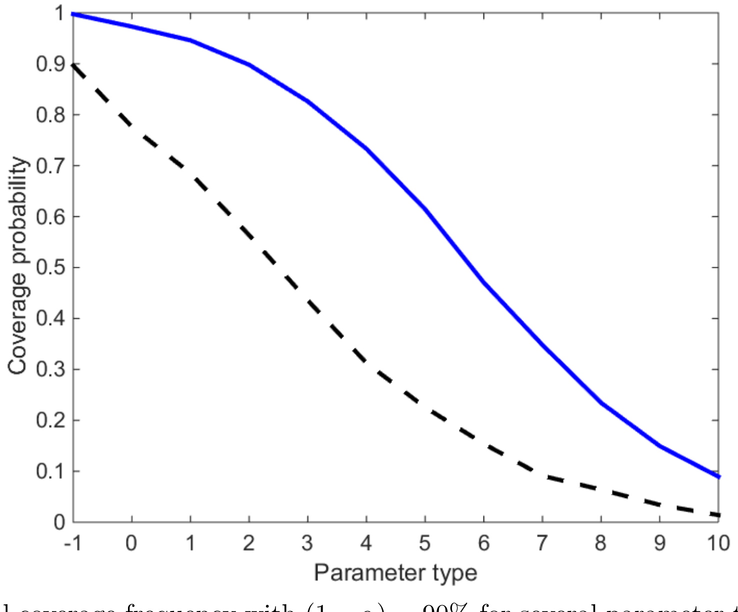

Figure 5.6: Empirical coverage frequency with $(1 - \alpha) = 90\%$ for several parameter types in Design 3. Solid line represents coverage with our GMS confidence set and dashed line represents coverage using a probit model based only on fully observed data. Parameter types are as follows: "-1" refers to $\theta_{interior} = \theta_0$, "0" refers to $\theta_{boundary}$, and "1-10" refer to $\{\theta_{exterior,v}\}_{v=1}^{10}$.

5.4 Design 4

Figures 5.7 and 5.8 repeat the analysis with the parameters in Design 4. As opposed to previous simulations, the data in this design are not missing at random (i.e. $\theta_0 \neq \tilde{\theta}_0$). As with all previous designs, the GMS inference on our outer identified set covers θ_0 and θ_1 (i.e. parameters types -1 and 0, respectively) with probability exceeding $(1-\alpha)$, and all remaining points are covered less frequently, with coverage decreasing monotonically as we move further away from the outer identified set. Unlike previous designs, inference based only on fully observed data suffers from an undercoverage problem. In particular, the empirical coverage of the true parameter value θ_0 is 50%, which is significantly below the desired coverage level of $(1-\alpha) = 90\%$. As explained earlier, this undercoverage is an expected consequence of the fact that data are not missing at random. Inference based only on fully observed data can be shown to cover $\tilde{\theta}_0 = (0.75, 1)$ instead of the true parameter value $\theta_0 = (0.5, 1)$, which also explains why the coverage increases as we consider parameter values located to the east of θ_0 and located in the exterior of the outer identified set.

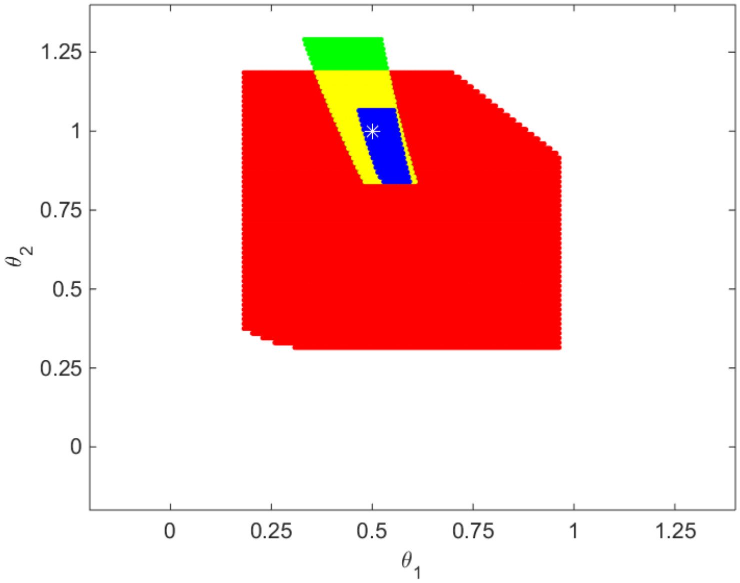

Figure 5.7: Identification analysis in Design 4. The white asterisk is θ_0, the blue region is $\Theta_I(F)$, the yellow region is $\Theta_S(F) \cap \Theta_I(F)^c$, the red region is $\Theta_{S_1}(F) \cap \Theta_S(F)^c$, and the green region is $\Theta_{S_2}(F) \cap \Theta_S(F)^c$.

5.5 Design 5

Figures 5.9 and 5.10 repeat the analysis using the parameters in Design 5. The purpose of this simulation is to explore the combined effect of having probability of missing data that is not constant and data that are not missing at random. The identification analysis produces qualitative results that are similar to a combination of Designs 3 and 4. The data in this design are not missing at random, which causes an expected undercoverage problem for inference based only on fully observed data.

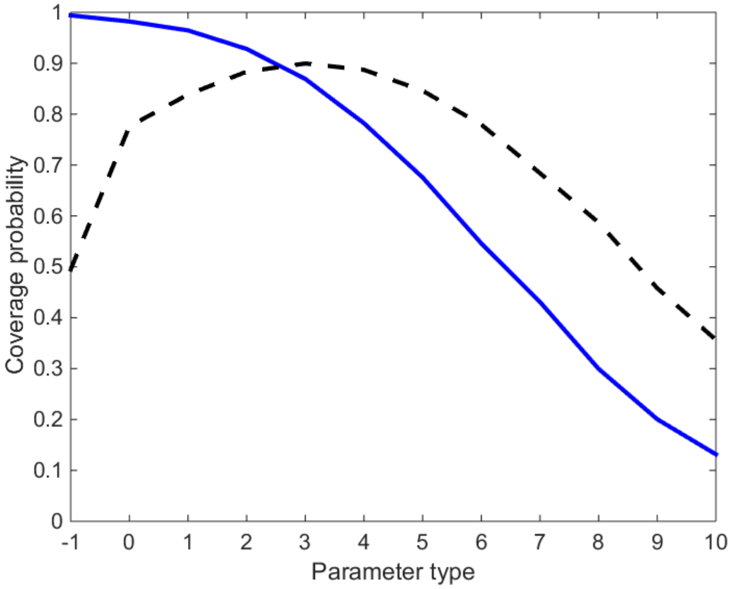

Figure 5.8: Empirical coverage frequency with $(1-\alpha) = 90\%$ for several parameter types in Design 4. Solid line represents coverage with our GMS confidence set and dashed line represents coverage using a probit model based only on fully observed data. Parameter types are as follows: "-1" refers to $\theta_{interior} = \theta_0$, "0" refers to $\theta_{boundary}$, and "1-10" refer to $\{\theta_{exterior,v}\}_{v=1}^{10}$.

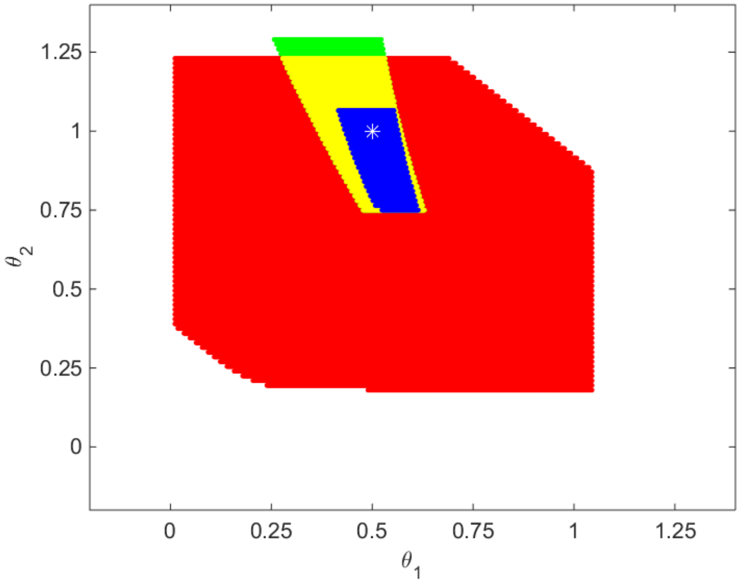

Figure 5.9: Identification analysis in Design 5. The white asterisk is θ_0, the blue region is $\Theta_I(F)$, the yellow region is $\Theta_S(F) \cap \Theta_I(F)^c$, the red region is $\Theta_{S_1}(F) \cap \Theta_S(F)^c$, and the green region is $\Theta_{S_2}(F) \cap \Theta_S(F)^c$.

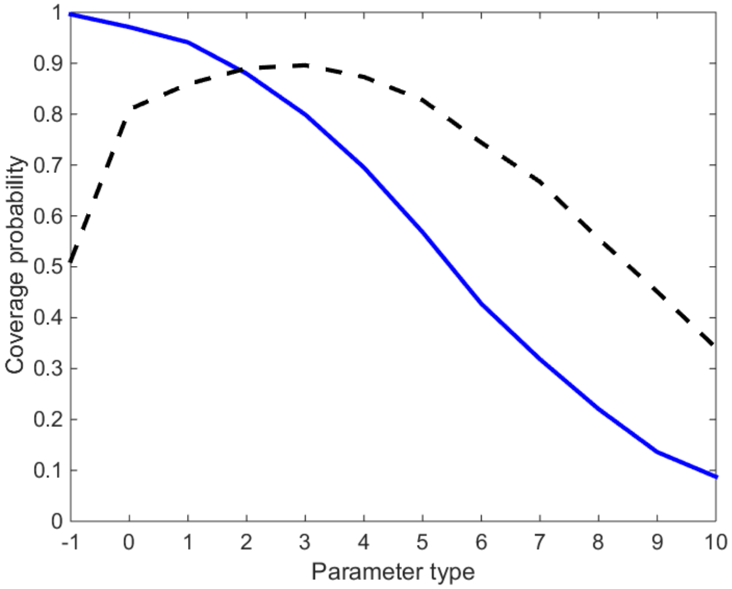

Figure 5.10: Empirical coverage frequency with $(1-\alpha) = 90\%$ for several parameter types in Design 5. Solid line represents coverage with our GMS confidence set and dashed line represents coverage using a probit model based only on fully observed data. Parameter types are as follows: "-1" refers to $\theta_{interior} = \theta_0$, "0" refers to $\theta_{boundary}$, and "1-10" refer to $\{\theta_{exterior,v}\}_{v=1}^{10}$.

6 Conclusions

This paper examines the problem of identification and inference on an econometric model with missing data, with special focus on the case when covariates are missing. Our econometric model is characterized by conditional moment conditions, which are routinely used in econometric applications. In order to address the missing data problem, we adopt a worst case scenario approach, which extracts the information from the observed data without imposing (untestable) assumptions on the (unobserved) distribution of missing data.

We show that having unobserved covariate observations implies that, in general, the parameter of interest is partially identified. We characterize the sharp identified set and show that it is usually prohibitively complex to compute or use for inference (at least with currently available methods). For this reason, we consider the construction of outer identified sets (i.e. supersets of the identified set) that are relatively easier to compute and use for inference.

We provide two different strategies to construct outer identified sets. The first strategy is based on using the conditional moment condition to derive a collection of unconditional moment conditions within boxes. The second strategy is based on integrating out the missing covariates in the conditional moment condition. We argue that these two outer identified sets contain non-trivial identifying power. Furthermore, we show that the two strategies provide different identifying power which can be easily combined to create a sharper outer identified set. The resulting outer identified set is relatively easy to compute and, most importantly, amenable to inference using recent developments in the literature on inference in partially identified models.

A Appendix

This appendix uses the following abbreviations. We use "RHS" and "LHS" to denote "right hand side" and "left hand side", respectively. We also use "s.t." to abbreviate "such that". Furthermore, for any population parameter A, we let $\mathcal{I}(A)$ denote the (sharp) identified set of A. Finally, we use $\mathcal{G} \equiv \mathbb{R}^{d_x} \times (0, \bar{r})^{d_x}$.

A.1 Appendix to Sections 2 and 3

Results in this section are developed under the following generalization of Assumption A.1.

Assumption B.1. Let the following conditions hold.

(i) Let $(\Omega, \mathcal{A}, F)$ be the probability space of (X, Y, W_X, W_Y), let $Y : \Omega \to S_Y \subseteq \mathbb{R}^{d_y}$ be the outcome variables, let $X : \Omega \to S_X \subseteq \mathbb{R}^{d_x}$ be the covariates, and let $W_X : \Omega \to \{0, \ldots, 2^{d_x} - 1\}$ and $W_Y : \Omega \to \{0, \ldots, 2^{d_y} - 1\}$ denote the missing data patterns of X and Y, respectively.[12] Any of the coordinates of X or Y may be subject to missing data.

(ii) There is a *known* function $m : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \times \mathbb{R}^{d_\theta} \to \mathbb{R}^{d_m}$ such that the true parameter value $\theta_0 \in \Theta \subseteq \mathbb{R}^{d_\theta}$ satisfies Eq. (1.1), i.e., $E_F[m(X, Y, \theta_0)|X = x] = \mathbf{0}$ $\forall x$ $F-$a.s.

Assumption B.2. The outcome random variable Y has no missing data.

We briefly comment on these assumptions. Assumption B.1 generalizes Assumption A.1 by allowing any arbitrary missing data pattern for outcome variables and covariates. Assumption B.2 is used only in order to simplify the statement and the proof of Lemma A.1.

According to Assumption B.1(i), $W_X : \Omega \to \{0, \ldots, 2^{d_x} - 1\}$ and $W_Y : \Omega \to \{0, \ldots, 2^{d_y} - 1\}$ denote the missing data patterns of X and Y, respectively. We now explain these variables further. Since X has d_x dimensions and each of them are allowed to be individually missing or not, there are 2^{d_x} possible missing covariate data patterns. The variable $W_X : \Omega \to \{0, \ldots, 2^{d_x} - 1\}$ indicates which one of these patterns occur, where $W_X = 0$ indicates that all of the covariates are observed and $W_X = 2^{d_x} - 1$ indicates that all of the covariates are unobserved. Notice that this is a special case of the main text in which there are only two missing data patterns, which gives rise to $W_X = W \in \{0, 1\}$. For every $w = 0, \ldots, 2^{d_x} - 1$, let $X_{1,w} : \Omega \to S_{X_{1,w}}$ be the sub-vector of X that is observed and let $X_{2,w} : \Omega \to S_{X_{2,w}}$ be the sub-vector of X that is unobserved. In a similar fashion, Y has d_y dimensions and each of them are allowed to be individually missing or not, there are 2^{d_y} possible missing outcome data patterns. The variable $W_Y : \Omega \to \{0, \ldots, 2^{d_y} - 1\}$ indicates which one of these patterns occur, where $W_Y = 0$ indicates that all of the outcome variables are observed and $W_Y = 2^{d_y} - 1$ indicates that all of the outcome variables are unobserved. For every $w = 0, \ldots, 2^{d_y} - 1$, let $Y_{1,w} : \Omega \to S_{Y_{1,w}}$ be the sub-vector of Y that is observed and let $Y_{2,w} : \Omega \to S_{Y_{2,w}}$ be the sub-vector of Y that is unobserved.

A.1.1 Proofs for results in Section 2

For the sake of simplicity, we characterize the identified set with arbitrary missing covariate data patterns but fully observed outcomes.[13]

Lemma A.1. *Assume Assumptions B.1-B.2. Then,* $\Theta_I(F)$ *is given by:*

$$\left\{ \begin{array}{l} \theta \in \Theta \ s.t. \ \forall w = 1, \dots, 2^{d_x} - 1, \\ \exists g_{1,w} : \mathbb{R}^{d_x} \to \mathbb{R}_+ \ \textit{and} \ g_{2,w} : \mathbb{R}^{d_y} \times \mathbb{R}^{d_x} \to \mathbb{R}_+ \ \textit{that satisfy:} \\ \text{(i)} \ g_{1,w}(x) = 0 \ \forall x \notin S_{X,w} \ \textit{and} \ g_{2,w}(y,x) = 0 \ \forall (y,x) \notin S_Y \times S_{X,w} \\ \text{(ii)} \ \int g_{1,w}(x) dx_{2,w} = 1 \ \forall x_{1,w} \in \mathbb{R}^{d_{x1,w}} \ P_F\text{–a.s.} \\ \text{(iii)} \ \int g_{2,w}(y,x) dy = 1 \ \forall x \in \mathbb{R}^{d_x} \ (P_F, g_{1,w})\text{–a.s.} \\ \text{(iv)} \ \left\{ \begin{array}{l} \int g_{2,w}(y,x) g_{1,w}(x) dx_{2,w} dy P_F[W_X = w] = dP_F[Y = y | X_{1,w} = x_{1,w}, W_X = w] P_F[W_X = w] \\ \forall (x_{1,w}, y) \in \mathbb{R}^{d_{x1,w}} \times \mathbb{R}^{d_y} \ P_F\text{–a.s.} \end{array} \right\} \\ \text{(v)} \ \left\{ \begin{array}{l} E_F[m(x,Y,\theta) | X = x, W_X = 0] dP_F[X = x | W_X = 0] P_F[W_X = 0] \ + \\ \sum_{w=1}^{2^{d_y}-2} (\int m(x,y,\theta) g_{2,w}(y,x) dy) g_{1,w}(x) dP_F[X_{1,w} = x_{1,w} | W_X = w] P_F[W_X = w] \ + \\ (\int m(x,y,\theta) g_{2,2^{d_x}-1}(y,x) dy) g_{1,2^{d_x}-1}(x) P_F[W_X = 2^{d_x} - 1] = \mathbf{0} \ \forall x \in \mathbb{R}^{d_x} \ (F, g_1)\text{–a.s.} \end{array} \right\} \end{array} \right\},$$

where dP_F *denotes the probability distribution function that induces* P_F*.*

Proof. By definition, $\Theta_I(F)$ is composed of $\theta \in \Theta$ for which the observed distributions and the restrictions on the parameter space do not contradict $E_F[m(Y,x,\theta)|X = x] = \mathbf{0} \ \forall x \ F$–a.s.

Step 1. For every $w = 1, \dots, 2^{d_x} - 1$, we derive:

$$\mathcal{I} \left(\begin{array}{l} dP_F[X_{2,w} = x_{2,w} | W_X = w, X_{1,w} = x_{1,w}] : x \in \mathbb{R}^{d_x}, \\ dP_F[Y = y | W_X = w, X = x] : (x,y) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \end{array} \right).$$

Fix $w \in \{1, \dots, 2^{d_x} - 2\}$ arbitrarily. Conditional on $W_X = w$, the object of interest is not identified because the distribution of $\{X_{2,w} | X_{1,w} = x_{1,w}, W_X = w\}$ is not observed. In order to obtain any expression that is identified, the dependence on the unobserved variable needs to be integrated out.

Define the set of functions $\Psi(w)$ as follows:

$$\left\{ \begin{array}{l} g_{1,w} : \mathbb{R}^{d_x} \to \mathbb{R}_+ \text{ and } g_{2,w} : \mathbb{R}^{d_y} \times \mathbb{R}^{d_x} \to \mathbb{R}_+ \text{ that satisfy:} \\ \text{(i)} \ g_{1,w}(x) = 0 \ \forall x \notin S_{X,w} \text{ and } g_{2,w}(y,x) = 0 \ \forall (y,x) \notin S_Y \times S_{X,w} \\ \text{(ii)} \ \int g_{1,w}(x) dx_{2,w} = 1 \ \forall x_{1,w} \in \mathbb{R}^{d_{x1,w}} \ P_F\text{–a.s.} \\ \text{(iii)} \ \int g_{2,w}(y,x) dy = 1 \ \forall x \in \mathbb{R}^{d_x} \ (P_F, g_{1,w})\text{–a.s.} \\ \text{(iv)} \ \left\{ \begin{array}{l} (\int g_{2,w}(y,x) g_{1,w}(x) dx_{2,w} dy) P_F[W_X = w] = dP_F[Y = y | X_{1,w} = x_{1,w}, W_X = w] P_F[W_X = w] \\ \forall (x_{1,w}, y) \in \mathbb{R}^{d_{x1,w}} \times \mathbb{R}^{d_y} \ P_F\text{–a.s.} \end{array} \right\} \end{array} \right\}$$

We now show that:

$$\mathcal{I} \left(\begin{array}{l} dP_F[X_{2,w} = x_{2,w} | W_X = w, X_{1,w} = x_{1,w}] : x \in \mathbb{R}^{d_x} \\ dP_F[Y = y | W_X = w, X = x] : (x,y) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \end{array} \right) = \Psi(w). \tag{A.1}$$

We first show that the identified set in the identified set on LHS of Eq. (A.1) is included in $\Psi(w)$. Consider a vector $(\bar{g}_{1,w}, \bar{g}_{2,w})$ that belongs to the identified set. Since these are distributions, they need to be non-negative functions and integrate to one. Furthermore, they also need to have zero density outside the support. Moreover, when we combine these distributions and integrate out $X_{2,w}$ they must be able to generate $\{dP_F[Y = y|X_{1,w} = x_{1,w}, W_X = w] : (x,y) \in \mathbb{R}^{d_x} \times \mathbb{R}\}$, whenever $P_F[W_X = w] > 0$. Hence, $(\bar{g}_{1,w}, \bar{g}_{2,w}) \in \Psi(w)$.

We now show the reverse inclusion. Consider $(\bar{g}_{1,w}, \bar{g}_{2,w}) \in \Psi(w)$. In order to show that $(\bar{g}_{1,w}, \bar{g}_{2,w})$ belongs to the identified set in the LHS of Eq. (A.1), we need to argue that the properties in $\Psi(w)$ exhaust all the necessary properties for the vector of distributions.

First, since $\bar{g}_{1,w}$ and $\bar{g}_{2,w}$ play the role of $\{dP_F[X_{2,w} = x_{2,w}|W_X = w, X_{1,w} = x_{1,w}] : x \in \mathbb{R}^{d_x}\}$ and $\{dP_F[Y = y|W_X = w, X = x] : (x,y) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}\}$, respectively, they need to satisfy all of the known restrictions regarding the support of (X, Y). This is guaranteed by condition (i).

Second, $\bar{g}_{1,w}$ needs to be a non-negative function that integrates to one with respect to $x_{2,w}$ to satisfy the (individual) necessary restrictions to be $\{dP_F[X_{2,w} = x_{2,w}|W_X = w, X_{1,w} = x_{1,w}] : x \in \mathbb{R}^{d_x}\}$. Similarly, $\bar{g}_{2,w}$ needs to be non-negative function that integrates to one with respect to y to satisfies all the (individual) necessary restrictions to be $\{dP_F[Y = y|W_X = w, X = x] : (x,y) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}\}$. These are guaranteed by conditions (ii) and (iii), respectively.

Third, if $P_F[W_X = w] > 0$, then there are restrictions that need to be satisfied by combination of these functions. First, if $\bar{g}_{1,w}$ plays the role of $\{dP_F[X_{2,w} = x_{2,w}|W_X = w, X_{1,w} = x_{1,w}] : x \in \mathbb{R}^{d_x}\}$, then the restrictions on $\bar{g}_{2,w}$ that need to be satisfied for $X_{2,w} = x_{2,w}$ may be allowed to be violated on a negligible set, which explains that the restrictions on $\bar{g}_2$ need to be satisfied $\bar{g}_{1,w} - a.s.$ Second, if $\bar{g}_1$ plays the role of $\{dP_F[X_{2,w} = x_{2,w}|W_X = w, X_{1,w} = x_{1,w}] : x \in \mathbb{R}^{d_x}\}$ and $\bar{g}_2$ plays the role of $\{dP_F[Y = y|W_X = w, X = x] : (x,y) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}\}$, then the combination of these two can be used to integrate out the unobserved vector $X_{2,w}$ and generate objects identified in the data. In particular, for any $(x_{1,w}, y) \in \mathbb{R}^{d_{x1,w}} \times \mathbb{R}^{d_y}$, the integral of $\{\bar{g}_2(y,x)\bar{g}_{1,w}(x) : x_{2,w} \in \mathbb{R}^{d_{x2,w}}\}$ produces $dP_F[Y = y|X_{1,w} = x_{1,w}, W_X = w]$. If $P_F[W_X = w] = 0$, $dP_F[Y = y|X_{1,w} = x_{1,w}, W_X = w]$ is not properly defined and the condition becomes vacuous. This is guaranteed by condition (iv).

Finally, since we are constructing probability distributions of all identified objects and these completely characterize the behavior of the random variables, this implies we have exhausted all available information.

Step 2. Derive an expression for $E_F[m(x, Y, \theta)|X = x]$ in terms of primitive probability distributions.

This step follows the structure of Manski (2003, Section 3.4). Fix $x \in \mathbb{R}^{d_x}$ arbitrarily and consider the following argument. By the law of iterated expectations:

$$E_F[m(x, Y, \theta)|X = x] = \sum_{w=0}^{2^{d_x}-1} E[m(x, Y, \theta)|X = x, W_X = w] P_F[W_X = w|X = x].$$

For every $w = 0, \ldots, 2^{d_x} - 1$, Bayes' theorem implies that:

$$P_F[W_X = w|X = x] = \frac{dP_F[X = x|W_X = w]P_F[W_X = w]}{\sum_{\zeta=0}^{2^{d_x}-1} dP_F[X = x|W_X = \zeta]P_F[W_X = \zeta]}.$$

By replacing this on the previous equation and expanding the expressions:

$$E_F[m(x, Y, \theta)|X = x] = \frac{\sum_{w=0}^{2^{d_x}-1} E[m(x, Y, \theta)|X = x, W_X = w] dP_F[X = x|W_X = w] P_F[W_X = w]}{\sum_{\zeta=0}^{2^{d_x}-1} dP_F[X = x|W_X = \zeta]P_F[W_X = \zeta]} = \frac{N(x)}{D(x)},$$

where $N(x)$ and $D(x)$ are given by:

$$N(x) \equiv \left\{ \begin{array}{l} E_F[m(x,Y,\theta)|X=x,W_X=0]dP_F[X=x|W_X=0]P_F[W_X=0]+ \\ \left\{ \begin{array}{l} (\int_{y\in\mathbb{R}} m(x,y,\theta)dP_F[Y=y|W_X=2^{d_x}-1,X=x]dy)\times \\ dP_F[X=x|W_X=2^{d_x}-1]P_F[W_X=2^{d_x}-1] \end{array} \right\}+ \\ \sum_{w=1}^{2^{d_x}-2}[\int_{y\in\mathbb{R}^{d_y}} m(x,y,\theta)dP[Y=y|W_X=w,X=x]dy\times \\ dP_F[X_{2,w}=x_{2,w}|W_X=w,X_{1,w}=x_{1,w}]dP_F[X_{1,w}=x_{1,w}|W_X=w]P_F[W_X=w]] \end{array} \right\},$$

$$D(x) \equiv \left\{ \begin{array}{l} dP_F[X=x|W_X=0]P_F[W_X=0]+ \\ \sum_{w=1}^{2^{d_x}-2} dP_F(X_{2,w}=x_{2,w}|W_X=w,X_{1,w}=x_{1,w})dP_F[X_{1,w}=x_{1,w}|W_X=w]P_F[W_X=w] \\ +dP_F(X_{2,w}=x_{2,w}|W_X=2^{d_x}-1)P_F[W_X=2^{d_x}-1] \end{array} \right\}.$$

Notice that the expressions for $N(x)$ and $D(x)$ are identified except for $dP_F[X=x|W_X=2^{d_x}-1]$, $dP_F[Y=y|W_X=w,X=x]$, and $dP_F[X_{2,w}=x_{2,w}|W_X=w,X_{1,w}=x_{1,w}]$, with $w=1,\ldots,2^{d_x}-2$.

Step 3. Fix $\theta\in\Theta$ arbitrarily and derive $\mathcal{I}(\{E_F[m(x,Y,\theta)|X=x]:x\in\mathbb{R}^{d_x}\})$.

Step 1 derives the identified set for a vector of distribution functions conditional on $W_X=w$ for $w=1,\ldots,2^{d_x}-1$. Given that the events $\{W_X=w\}$ and $\{W_X=\tilde{w}\}$ are disjoint for $w\neq\tilde{w}$, the identified set for the joint vector of functions for $w=1,\ldots,2^{d_x}-1$ is the product of the sets derived in step 1, i.e.,

$$\begin{array}{l} \mathcal{I}\left(\left\{ \begin{array}{l} dP_F[X_{2,w}=x_{2,w}|W_X=w,X_{1,w}=x_{1,w}]:x\in\mathbb{R}^{d_x} \\ dP_F[Y=y|W_X=w,X=x]:(x,y)\in\mathbb{R}^{d_x}\times\mathbb{R}^{d_y} \end{array} \right\}_{w=1,\ldots,2^{d_x}-1}\right) \\ =\prod_{w=1,\ldots,2^{d_x}-1}\mathcal{I}\left(\begin{array}{l} dP_F[X_{2,w}=x_{2,w}|W_X=w,X_{1,w}=x_{1,w}]:x\in\mathbb{R}^{d_x} \\ dP_F[Y=y|W_X=w,X=x]:(x,y)\in\mathbb{R}^{d_x}\times\mathbb{R}^{d_y} \end{array} \right). \end{array}$$

When we combine this with step 2, it follows that:

$$\mathcal{I}(\{E_F(m(x,Y,\theta)|X=x):x\in\mathbb{R}^{d_x}\})=$$

$$\left\{ \begin{array}{l} f:\mathbb{R}^{d_x}\to\mathbb{R}^{d_m} \text{ s.t. } \forall w=1,\ldots,2^{d_x}-1,\ \exists g_{1,w}:\mathbb{R}^{d_x}\to\mathbb{R}_+ \text{ and } g_{2,w}:\mathbb{R}^{d_y}\times\mathbb{R}^{d_x}\to\mathbb{R}_+ \text{ that satisfy:} \\ \text{(i) } g_{1,w}(x)=0\ \forall x\notin S_{X,w}, g_{2,w}(y,x)=0\ \forall(y,x)\notin S_Y\times S_{X,w} \\ \text{(ii) } \int g_{1,w}(x)dx_{2,w}=1\ \forall x_{1,w}\in\mathbb{R}^{d_{x1,w}}\ P_F\text{−a.s.} \\ \text{(iii) } \int g_{2,w}(y,x)dy=1\ \forall x\in\mathbb{R}^{d_x}\ (P_F,g_{1,w})\text{−a.s.} \\ \text{(iv) } \left\{ \begin{array}{l} \int g_{2,w}(y,x)g_{1,w}(x)dx_{2,w}dyP_F[W_X=w]=dP_F[Y=y|X_{1,w}=x_{1,w},W_X=w]P_F[W_X=w] \\ \forall(x_{1,w},y)\in\mathbb{R}^{d_{x1,w}}\times\mathbb{R}^{d_y}\ P_F\text{−a.s.} \end{array} \right\} \\ \text{(v) } f(x)=N(x,g_1,g_2)/D(x,g_1,g_2)\ (F,g_1)\text{−a.s.} \end{array} \right\},$$

where $N(x,g_1,g_2)$ and $D(x,g_1,g_2)$ are similar to $N(x)$ and $D(x)$ in step 2, except that the unidentified expressions are replaced by the functions $\{g_{1,w}\}_{w=1}^{2^{d_x}-1}$ and $\{g_{2,w}\}_{w=1}^{2^{d_x}-1}$, i.e.,

$$N(x,g_1,g_2) \equiv \left\{ \begin{array}{l} E_F[m(x,Y,\theta)|X=x,W_X=0]dP_F[X=x|W_X=0]P_F[W_X=0]\ + \\ (\int_{y\in\mathbb{R}} m(x,y,\theta)g_{2,2^{d_x}-1}(y,x)dy)g_{1,2^{d_x}-1}(x)P_F[W_X=2^{d_x}-1]\ + \\ \sum_{w=1}^{2^{d_x}-2}(\int_{y\in\mathbb{R}^{d_y}} m(x,y,\theta)g_{2,w}(y,x)dy)\ g_{1,w}(x)dP_F[X_{1,w}=x_{1,w}|W_X=w]P_F[W_X=w] \end{array} \right\},$$

$$D(x,g_1,g_2) \equiv \left\{ \begin{array}{l} dP_F[X=x|W_X=0]P_F[W_X=0]\ + \\ \sum_{w=1}^{2^{d_x}-2} g_{1,w}(x)dP_F[X_{1,w}=x_{1,w}|W_X=w]P_F[W_X=w] \\ +g_{1,2^{d_x}-1}(x)P_F[W_X=2^{d_x}-1] \end{array} \right\}.$$

Step 4. Conclude the proof.

By definition, $\theta\in\Theta_I(F)$ if and only if the zero function belongs to $\mathcal{I}(\{E_F(m(x,Y,\theta)|X=x): x\in\mathbb{R}^{d_x}\})$. The characterization in the statement follows from imposing the existence of the zero function in the definition of $\mathcal{I}(\{E_F(m(x,Y,\theta)|X=x): x\in\mathbb{R}^{d_x}\})$. □

Proof of Lemma 2.1. This result is a special case of Lemma A.1. □

A.1.2 Proofs for results in Section 3

Recall that $W_X = w$ for $w=0,\ldots,2^{d_x}-1$ determines the missing data pattern of the covariates. For any $(x,\nu)\in\mathcal{G}$ and any $w=0,\ldots,2^{d_x}-1$, let $B_1(x_{1,w},\nu_{1,w})$ and $B_2(x_{2,w},\nu_{2,w})$ denote the projection of the d_x-dimensional set $B(x,\nu)$ onto the space of the observed covariates $X_{1,w}$ and unobserved covariates $X_{2,w}$, respectively. With some abuse of notation, we define:

$$\begin{aligned} B_{1,w}(x_{1,w},\nu_{1,w}) &\equiv \{x_{1,w}\in\mathbb{R}^{d_{1,w}} \text{ s.t. } \exists x_{2,w}\in\mathbb{R}^{d_{2,w}} \text{ with } (x_{1,w},x_{2,w})\in B(x,\nu)\}, \\ B_{2,w}(x_{2,w},\nu_{2,w}) &\equiv \{x_{2,w}\in\mathbb{R}^{d_{2,w}} \text{ s.t. } \exists x_{1,w}\in\mathbb{R}^{d_{1,w}} \text{ with } (x_{1,w},x_{2,w})\in B(x,\nu)\}, \end{aligned} \tag{A.2}$$

where the abuse of notation occurs in the reshuffling of coordinates in the expression "$(x_{1,w},x_{2,w})$". We note that these definitions imply that $B(x,\nu)\equiv B_{1,w}(x_{1,w},\nu_{1,w})\times B_{2,w}(x_{2,w},\nu_{2,w})$.

Theorem A.1. *Assume Assumption B.1 and choose $\bar{r}\in(0,\infty]$ arbitrarily. Let $Z\equiv(\sum_{\tilde{w}=0}^{2^{d_y}-1}1[W_Y=\tilde{w}]Y_{1,\tilde{w}},\sum_{w=0}^{2^{d_y}-1}1[W_X=w]X_{1,w},W_X,W_Y)$ and let $M_1(Z,\theta,x,\nu)=\{M_{1,j}(Z,\theta,x,\nu)\}_{j=1}^{d_m}$ with*

$$M_{1,j}(Z,\theta,x,\nu)\equiv$$

$$\left[\begin{array}{c} -\left(\begin{array}{l} \sum_{\tilde{w}=0}^{2^{d_y}-1}\sum_{w=1}^{2^{d_x}-1} \inf\limits_{(\xi_{2,w},y)\in\{S_{X_{2,w}}\cap B_{2,w}(x_{2,w},\nu_{2,w})\}\times S_{Y,\tilde{w}}} m_j((X_{1,w},\xi_{2,w}),y,\theta) \\ \times 1[S_{X_{2,w}}\cap B_{2,w}(x_{2,w},\nu_{2,w})\neq\emptyset, X_{1,w}\in B_{1,w}(x_{1,w},\nu_{1,w}), W_X=w, W_Y=\tilde{w}] \\ +\sum_{\tilde{w}=1}^{2^{d_y}-1} \inf\limits_{y_{2,\tilde{w}}\in S_{Y_{2,\tilde{w}}}} m_j(X,(Y_{1,\tilde{w}},y_{2,\tilde{w}}),\theta)\times 1[X\in B(x,\nu),W_X=0,W_Y=\tilde{w}] \\ +m_j(X,Y,\theta)\times 1[X\in B(x,\nu),W_X=0,W_Y=0] \end{array} \right), \\ \left(\begin{array}{l} \sum_{\tilde{w}=0}^{2^{d_y}-1}\sum_{w=1}^{2^{d_x}-1} \sup\limits_{(\xi_{2,w},y)\in\{S_{X_{2,w}}\cap B_{2,w}(x_{2,w},\nu_{2,w})\}\times S_{Y,\tilde{w}}} m_j((X_{1,w},\xi_{2,w}),y,\theta) \\ \times 1[S_{X_{2,w}}\cap B_{2,w}(x_{2,w},\nu_{2,w})\neq\emptyset, X_{1,w}\in B_{1,w}(x_{1,w},\nu_{1,w}), W_X=w, W_Y=\tilde{w}] \\ +\sum_{\tilde{w}=1}^{2^{d_y}-1} \sup\limits_{y_{2,\tilde{w}}\in S_{Y_{2,\tilde{w}}}} m_j(X,(Y_{1,\tilde{w}},y_{2,\tilde{w}}),\theta)\times 1[X\in B(x,\nu),W_X=0,W_Y=\tilde{w}] \\ +m_j(X,Y,\theta)\times 1[X\in B(x,\nu),W_X=0,W_Y=0] \end{array} \right) \end{array} \right],$$

for all $(\theta,(x,\nu)) \in \Theta \times \mathcal{G}$ and where $B(\cdot)$ is defined as in Eq. (3.1) and $B_{1,w}(\cdot)$, and $B_{2,w}(\cdot)$ are defined as in Eq. (A.2). Consider the following set:

$$\Theta_{S_1}(F) \equiv \left\{\theta \in \Theta : \; E_F[M_1(Z,\theta,x,\nu)] \geq \mathbf{0} \; \forall (x,\nu) \in \mathbb{R}^{d_x} \times (0,\bar{r})^{d_x} \right\}.$$

Then, $\Theta_I(F) \subseteq \Theta_{S_1}(F)$, i.e., $\Theta_{S_1}(F)$ is an outer identified set.

Proof. Consider any arbitrary $(\theta,(x,\nu)) \in \Theta_I(F) \times \mathcal{G}$. By definition, this implies that $E_F[m(X,Y,\theta)|X = x] = \mathbf{0}$ P_F-a.s. and, thus, by multiplying this expression by $1(X \in B(x,\nu))$ and integrating with respect to the density of X, we deduce that $E_F[m(X,Y,\theta)1(X \in B(x,\nu))] = \mathbf{0}$, or, equivalently,

$$E_F\left[\begin{array}{l} \sum_{\tilde{w}=0}^{2^{d_y}-1}\sum_{w=1}^{2^{d_x}-1} m((X_{1,w},X_{2,\tilde{w}}),Y,\theta) \\ \times 1[X_{2,w} \in B_{2,w}(x_{2,w},\nu_{2,w}), X_{1,w} \in B_{1,w}(x_{1,w},\nu_{1,w}), W_X = w, W_Y = \tilde{w}] \\ +\sum_{\tilde{w}=1}^{2^{d_y}-1} m(X,Y,\theta) \times 1[X \in B(x,\nu), W_X = 0, W_Y = \tilde{w}] \\ +m(X,Y,\theta) \times 1[X \in B(x,\nu), W_X = 0, W_Y = 0] \end{array}\right] = \mathbf{0}.$$

For each $w = 1,\ldots,2^{d_x}-1$, the value of $X_{2,w}$ is unobserved. In addition, even if the value of Y is observed, the value of Y conditional on the event of $\{X_{2,w} \in B_{2,w}(x_{2,w},\nu_{2,w})\}$ is also unobserved. Finally, for each $\tilde{w}-1$, the value of $Y_{2,\tilde{w}}$ is unobserved. By imposing logical lower and upper bounds on the unknown variables for each of the d_m coordinates, the desired result follows. □

Proof of Theorem 3.1. This result is a special case of Theorem A.1. The only difference occurs in the definition of Z, which we now explain. Let Z be as defined in Theorem A.1. In this case, $W_Y = 0$ and $W = W_X \in \{0,1\}$, leading to $Y_{1,0} = Y$, $Y_{2,0} = \emptyset$, $X_{1,0} = X$, $X_{2,0} = \emptyset$, $X_{1,1} = X_1$, and $X_{2,1} = X_2$, and so $Z \equiv ((Y,(X_1,X_2),W = 0, W_Y = 1),(Y,X_1,W = 1,W_Y = 1))$. To complete the proof, notice that the information in Z can be equivalently re-expressed by $(Y,X_1,(1-W)X_2,W)$, leading to the definition of Z in the statement of Theorem 3.1. □

Theorem A.2. *Assume Assumption B.1 and choose $\bar{r} \in (0,\infty]$ arbitrarily. Let $Z \equiv (\sum_{\tilde{w}=0}^{2^{d_y}-1} 1[W_Y = \tilde{w}]Y_{1,\tilde{w}}, \sum_{w=0}^{2^{d_y}-1} 1[W_X = w]X_{1,w}, W_X, W_Y)$. There are two possible cases.*

1. *No covariates that are always observed. Then set $M_2(Z,\theta) = \{M_{2,j}(Z,\theta)\}_{j=1}^{d_m}$ with*

$$M_{2,j}(Z,\theta) \equiv$$
$$\left[\begin{array}{c} \left(\begin{array}{l} \sum_{\tilde{w}=1}^{2^{d_y}-1}\sum_{w=1}^{2^{d_x}-1} \sup\limits_{(x_{2,w},y_{2,\tilde{w}}) \in S_{X_{2,w}} \times S_{Y_{2,\tilde{w}}}} m_j((X_{1,w},x_{2,w}),(Y_{1,\tilde{w}},y_{2,\tilde{w}}),\theta)1[W_X = w, W_Y = \tilde{w}] \\ +\sum_{w=1}^{2^{d_x}-1} \sup\limits_{x_{2,w} \in S_{X_{2,w}}} m_j((X_{1,w},x_{2,w}),Y,\theta)1[W_X = w, W_Y = 0] \end{array}\right), \\ -\left(\begin{array}{l} \sum_{\tilde{w}=1}^{2^{d_y}-1}\sum_{w=1}^{2^{d_x}-1} \inf\limits_{(x_{2,w},y_{2,\tilde{w}}) \in S_{X_{2,w}} \times S_{Y_{2,\tilde{w}}}} m_j((X_{1,w},x_{2,w}),(Y_{1,\tilde{w}},y_{2,\tilde{w}}),\theta)1[W_X = w, W_Y = \tilde{w}] \\ +\sum_{w=1}^{2^{d_x}-1} \inf\limits_{x_{2,w} \in S_{X_{2,w}}} m_j((X_{1,w},x_{2,w}),Y,\theta)1[W_X = w, W_Y = 0] \end{array}\right) \end{array}\right].$$

Consider the following set:

$$\Theta_{S_2}(F) \equiv \{\theta \in \Theta : \; E_F[M_2(Z,\theta)] \geq \mathbf{0}\}.$$

Then, $\Theta_I(F) \subseteq \Theta_{S_2}(F)$, i.e., $\Theta_{S_2}(F)$ is an outer identified set.

2. *Some covariates that are always observed. Denote the sub-vector of the covariates that are always observed by X^{AO}, denote its support by $S_{X^{AO}} \in \mathbb{R}^{d_{AO}}$. The remaining covariates that are not always observed are denoted by $X^{NAO} \in \mathbb{R}^{d_{NAO}}$ and, with a slight abuse of notation, these can take the role of X in the previous case, i.e., set $M_2(Z,\theta,x^{AO},\nu^{AO}) = \{M_{2,j}(Z,\theta,x^{AO},\nu^{AO})\}_{j=1}^{d_m}$ with*

$$M_{2,j}(Z,\theta,x^{AO},\nu^{AO}) \equiv$$

$$\left[\begin{array}{c} \left(\begin{array}{l} \sum_{\tilde{w}=1}^{2^{d_y}-1}\sum_{w=1}^{2^{d_{NAO}}-1}\left(\begin{array}{l} \sup\limits_{(x_{2,w}^{NAO},y_{2,\tilde{w}})\in S_{X_{2,w}^{NAO}}\times S_{Y_{2,\tilde{w}}}} m_j((X^{AO},X_{1,w}^{NAO},x_{2,w}^{NAO}),(Y_{1,\tilde{w}},y_{2,\tilde{w}}),\theta) \\ \times 1[W_{X^{NAO}}=w,W_Y=\tilde{w}] \end{array}\right) \\ +\sum_{w=1}^{2^{d_{NAO}}-1} \sup\limits_{x_{2,w}^{NAO}\in S_{X_{2,w}^{NAO}}} m_j((X^{AO},X_{1,w}^{NAO},x_{2,w}^{NAO}),Y,\theta)1[W_{X^{NAO}}=w,W_Y=0] \end{array}\right), \\ -\left(\begin{array}{l} \sum_{\tilde{w}=1}^{2^{d_y}-1}\sum_{w=1}^{2^{d_{NAO}}-1}\left(\begin{array}{l} \inf\limits_{(x_{2,w}^{NAO},y_{2,\tilde{w}})\in S_{X_{2,w}^{NAO}}\times S_{Y_{2,\tilde{w}}}} m_j((X^{AO},X_{1,w}^{NAO},x_{2,w}^{NAO}),(Y_{1,\tilde{w}},y_{2,\tilde{w}}),\theta) \\ \times 1[W_{X^{NAO}}=w,W_Y=\tilde{w}] \end{array}\right) \\ +\sum_{w=1}^{2^{d_{NAO}}-1} \inf\limits_{x_{2,w}^{NAO}\in S_{X_{2,w}^{NAO}}} m_j((X^{AO},X_{1,w}^{NAO},x_{2,w}^{NAO}),Y,\theta)1[W_{X^{NAO}}=w,W_Y=0] \end{array}\right) \end{array}\right]$$

$$\times\, 1(X^{AO} \in B(x^{AO},\nu^{AO})).$$

Consider the following set:

$$\Theta_{S_2}(F) \equiv \left\{\theta \in \Theta : \ E_F[M_2(Z,\theta,x^{AO},\nu^{AO})] \geq \mathbf{0} \ \forall (x^{AO},\nu^{AO}) \in \mathbb{R}^{d_{AO}} \times (0,\bar{r})^{d_{AO}}\right\}.$$

Then, $\Theta_I(F) \subseteq \Theta_{S_2}(F)$, i.e., $\Theta_{S_2}(F)$ is an outer identified set.

Proof. We only cover the proof of part 1. The proof for part 2 follows exactly from the same arguments as part 1, except that (a) inside the expectations, there is an extra $1[X^{AO} \in B(x^{AO},\nu^{AO})]$ term, and (b) the proof should be repeated for every $(x^{AO},\nu^{AO}) \in \mathbb{R}^{d_{AO}} \times (0,\bar{r})^{d_{AO}}$.

Fix $\theta \in \Theta_I(F)$ arbitrarily. By definition, this implies that $E_F[m(X,Y,\theta)|X=x] = \mathbf{0}$ P_F-a.s. and, thus, $E_F[m(X,Y,\theta)] = \mathbf{0}$. Next, consider the following argument. The law of iterated expectations implies that:

$$E_F[m(X,Y,\theta)] = \sum_{\tilde{w}=0}^{2^{d_y}-1}\sum_{w=0}^{2^{d_x}-1}\left\{\begin{array}{l} \int_{x\in S_X} E_F[m(x,(Y_{1,\tilde{w}},Y_{2,\tilde{w}}),\theta)|X=x,W_X=w,W_Y=\tilde{w}]\times \\ dP_F[X=x|W_X=w,W_Y=\tilde{w}]P_F[W_X=w,W_Y=\tilde{w}] \end{array}\right\}.$$

The RHS is the sum of several terms. The expression is not identified because $\{dP_F[X=x|W_X=w,W_y=\tilde{w}] : x \in \mathbb{R}^{d_x}\}$ and $\{dP_F[Y_{2,\tilde{w}}|X=x,W_X=w,W_y=\tilde{w}] : (y_{2,\tilde{w}},x) \in \mathbb{R}^{d_{y_{2,\tilde{w}}}} \times \mathbb{R}^{d_x}\}$ are not identified for $w>0$ and $\tilde{w}>0$, respectively. By imposing logical lower and upper bounds on the unknown variables for each of the d_m coordinates, the desired result follows. □

Proof of Theorem 3.2. This result is a special case of Theorem A.2. Notice that X_1 in Theorem 3.2 takes the role of X^{AO} in Theorem A.2 as they are always observed. The only other difference occurs in the definition of Z, which can be explained by repeating the argument used in the proof of Theorem 3.1. □

A.2 Appendix to Section 4

This section provides the details regarding the properties of our confidence sets. We first introduce relevant definitions, follow with our assumptions, and conclude by establishing formal results.

A.2.1 Definitions

Our moment inequality model described in Theorem 3.3 has parameters (θ, F), where $\theta \in \Theta$ denotes a generic value for the parameter of interest and F denotes the distribution of the data. We now define the random variable $M(Z,\theta)$ that is an envelope for the collection of random variables $\{M(Z,\theta,x,\nu) \ : \ (x,\nu) \in \mathcal{G}\}$ (see, e.g., Pollard (1990, Page 19)). By definition, for any (θ, F) with $Z \sim F$, the envelope $M(Z,\theta)$ satisfies:

$$|M(Z,\theta,x,\nu)| \leq M(Z,\theta) \ \ \forall (x,\nu) \in \mathcal{G}. \tag{A.3}$$

In the context of Assumption A.1, the natural envelope is as follows:

$$M(Z,\theta) \equiv \left[\begin{array}{c} \left\{ \begin{array}{c} \sup\limits_{(\xi_2,y)\in S_{X_2}\times S_Y} |m_j((X_1,\xi_2),y,\theta)| \ 1[W=1] + |m_j(X,Y,\theta)| \ 1[W=0], \\ \sup\limits_{(\xi_2,y)\in S_{X_2}\times S_Y} |m_j((X_1,\xi_2),y,\theta)| \ 1[W=1] + |m_j(X,Y,\theta)| \ 1[W=0] \end{array} \right\}_{j=1}^{d_m}, \\ \left\{ \begin{array}{c} \sup\limits_{\xi_2 \in S_{X_2}} |m_j((X_1,\xi_2),Y,\theta)| \ 1[W=1] + |m_j(X,Y,\theta)| \ 1[W=0], \\ \sup\limits_{\xi_2 \in S_{X_2}} |m_j((X_1,\xi_2),Y,\theta)| \ 1[W=1] + |m_j(X,Y,\theta)| \ 1[W=0], \end{array} \right\}_{j=1}^{d_m} \end{array} \right].$$

Under the more involved setup described in Assumption B.1, one could define an analogous envelope function. This is available from the authors upon request.

For any $(x,\nu), (\tilde{x},\tilde{\nu}) \in \mathcal{G}$, we define the following population objects:

$$\begin{aligned} D_F(\theta) &\equiv Diag(Var_F(M(Z,\theta))), \\ \Sigma_F(\theta,(x,\nu),(\tilde{x},\tilde{\nu})) &\equiv Cov_F[\ M(Z,\theta,x,\nu) \ , \ M(Z,\theta,\tilde{x},\tilde{\nu}) \ ], \\ h_{1,n,F}(\theta,x,\nu) &\equiv \sqrt{n} \ D_F^{-1/2}(\theta) \ E_F[M(Z,\theta,x,\nu)], \\ h_{2,F}(\theta,(x,\nu),(\tilde{x},\tilde{\nu})) &\equiv D_F^{-1/2}(\theta) \times \Sigma_F(\theta,(x,\nu),(\tilde{x},\tilde{\nu})) \times D_F^{-1/2}(\theta), \\ \mathcal{H}_2 &\equiv \{h_{2,F}(\theta,\cdot,\cdot) : (\theta,F) \in \mathcal{F}\}. \end{aligned} \tag{A.4}$$

The diagonal matrix $D_F(\theta)$ is used to standardize the random variable $M(Z,\theta,x,\nu)$ in a scale-invariant and uniform (in (x,ν)) way at the population level. $h_{1,n,F}(\theta,x,\nu)$ and $h_{2,F}(\theta,(x,\nu),(\tilde{x},\tilde{\nu}))$ are standardized version of the slackness in the moment inequalities $\sqrt{n}E_F[M(Z,\theta,x,\nu)]$ and the variance-covariance kernel $\Sigma_F(\theta,(x,\nu),(\tilde{x},\tilde{\nu}))$, respectively. Finally, $\mathcal{H}_2$ is the parameter space for the standardized variance-covariance kernels. This is a space of $p \times p$-matrix-valued covariance kernels on $\mathcal{G} \times \mathcal{G}$, which we metrize with the sup-norm, i.e., for $h_{2,F}(\theta,\cdot,\cdot), \check{h}_{2,\check{F}}(\theta,\cdot,\cdot) \in \mathcal{H}_2$,

$$d(h_{2,F}(\theta,\cdot,\cdot), h_{2,\tilde{F}}(\tilde{\theta},\cdot,\cdot)) \equiv \sup_{(x,\nu),(\tilde{x},\tilde{\nu})\in\mathcal{G}} ||h_{2,F}(\theta,(x,\nu),(\tilde{x},\tilde{\nu})) - h_{2,\check{F}}(\check{\theta},(x,\nu),(\tilde{x},\tilde{\nu}))||.$$

Furthermore, for an i.i.d. sample $\{Z_i\}_{i=1}^n$ distributed according to F, we define the following sample objects associated to $\{M(Z_i,\theta)\}_{i=1}^n$:

$$\begin{aligned}\overline{M}_n(\theta) &\equiv n^{-1}\sum\nolimits_{i=1}^n M(Z_i,\theta),\\ \hat{\Sigma}_n(\theta) &\equiv n^{-1}\sum\nolimits_{i=1}^n \left[M(Z_i,\theta)-\overline{M}_n(\theta)\right]\left[M(Z_i,\theta)-\overline{M}_n(\theta)\right]',\\ D_n(\theta) &\equiv Diag(\hat{\Sigma}_n(\theta)). \end{aligned} \tag{A.5}$$

By definition, $\overline{M}_n(\theta)$ and $\hat{\Sigma}_n(\theta)$ are the sample mean and sample covariance of $\{M(Z_i,\theta)\}_{i=1}^n$. The diagonal matrix $D_n(\theta)$ is the sample analogue of $D_F(\theta)$.

Finally, we define the following "mixed" (i.e. part sample and part population) objects:

$$\begin{aligned} v_{n,F}(\theta,x,\nu) &\equiv n^{-1/2}\sum\nolimits_{i=1}^n D_F^{-1/2}(\theta)\ (M(Z_i,\theta,x,\nu)-E_F[M(Z,\theta,x,\nu)]),\\ \hat{h}_{2,n,F}(\theta,(x,\nu),(\tilde{x},\tilde{\nu})) &\equiv D_F^{-1/2}(\theta)\times\hat{\Sigma}_n(\theta,(x,\nu),(\tilde{x},\tilde{\nu}))\times D_F^{-1/2}(\theta).\end{aligned}$$

Notice that $v_{n,F}(\theta,x,\nu)$ and $\hat{h}_{2,n,F}(\theta,(x,\nu),(\tilde{x},\tilde{\nu}))$ are the standardized empirical process and variance covariance kernel, where the standardization is conducted using the population variance $D_F(\theta)$ in Eq. (A.4).

We now define several relevant parameter spaces for (θ,F). The first parameter space is the *baseline* parameter space. The second parameter space is the *null* parameter space and is the subset of the baseline parameter space in which the moment inequalities of our outer identified set are satisfied. The third parameter space is a subset of the null parameter space where the variance-covariance kernel is restricted to an arbitrary compact set. This last parameter space is related to the parameter space in AS13, Theorems 1 and 2, and is used to establish the uniform coverage result in Theorem 4.1.[14]

Definition A.1 (Baseline parameter space)**.** *The baseline parameter space, denoted by $\mathcal{F}$, is the collection of parameter values (θ,F) that satisfy the following conditions:*

(i) $\theta\in\Theta$,

(ii) $\{Z_i\}_{i=1}^n$ *are i.i.d. distributed according to* F,

(iii) $\sigma^2_{F,j}(Z,\theta)\equiv Var_F[M_j(Z,\theta)]\in(0,\infty)$, *for* $j=1,\dots,p$,

(iv) $E_F|M_j(Z,\theta)/\sigma_{F,j}(Z,\theta)|^{2+\delta}\leq K$ *for* $j=1,\dots,p$,

for some constants $\delta,K\in(0,\infty)$, *where* $M(Z,\theta)$ *satisfies Eq. (A.3).*

Definition A.2 (Null parameter space)**.** *The null parameter space, denoted by $\mathcal{F}_0$, is the collection of parameter values (θ,F) that satisfy conditions (i)-(iv) in Definition A.1 plus the following one:*

(v) $E_F[M(Z,\theta,x,\nu)]\geq\mathbf{0}\ \ \forall(x,\nu)\in\mathcal{G}$ *or, equivalently,* $\theta\in\Theta_S(F)$.

In words, the null parameter space $\mathcal{F}_0$ is the subset of parameters in the baseline parameter space $\mathcal{F}$ that satisfies the moment inequalities of our outer identified set.

Definition A.3 (Restricted null parameter space)**.** *Let $\bar{\mathcal{H}}_2$ denote an arbitrary compact subset of $\mathcal{H}_2$ (metrized with the sup-norm). The restricted null parameter space, denoted by $\bar{\mathcal{F}}_0$, is defined as follows:*

$$\bar{\mathcal{F}}_0\ \equiv\ \{(\theta,F)\in\mathcal{F}_0\ :\ h_{2,F}(\theta,\cdot,\cdot)\in\bar{\mathcal{H}}_2\}.$$

We conclude this section by defining sequences of parameters that are relevant for our asymptotic analysis.

Definition A.4. *For any $h_2 \in \bar{\mathcal{H}}_2$, $SubSeq(h_2)$ is the set of sequences $\{(\theta_n, F_n)\}_{n\geq 1}$ for which:*

$$\sup_{(x,\nu),(\tilde{x},\tilde{\nu})\in\mathcal{G}} ||h_{2,F_n}(\theta_n,(x,\nu),(\tilde{x},\tilde{\nu})) - h_2((x,\nu),(\tilde{x},\tilde{\nu}))|| \to 0.$$

A.2.2 Assumptions

Our results require the following assumptions which are directly related to those in AS13.

Assumption A.2. Let $\mathcal{W}$ denote the set of $p \times p$ positive definite matrices and let $\mathbb{R}^p_{[+\infty]}$ denote p copies of $\mathbb{R}_{[+\infty]} \equiv \mathbb{R} \cup \{+\infty\}$. For every $(y, \Sigma) \in \mathbb{R}^p_{[+\infty]} \times \mathcal{W}$, the function S used in Eq. (4.1) satisfies:

(a) $S(Dy, D\Sigma D) = S(y,\Sigma)$ $\forall D \in \Delta$, where Δ denotes the space of positive definite diagonal $p \times p$ matrices,

(b) $S(y,\Sigma)$ is non-increasing in each element of y,

(c) $S(y,\Sigma) \geq 0$,

(d) S is uniformly continuous in the sense that $\sup_{\mu\in\mathbb{R}^p_+} |S(\tilde{y}+\mu,\tilde{\Sigma}) - S(y+\mu,\Sigma)| \to 0$ as $(\tilde{y},\tilde{\Sigma}) \to (y,\Sigma)$,

(e) $S(y,\Sigma) \leq S(y,\Sigma+\Sigma_1)$ for any $p \times p$ positive semi-definite matrix Σ_1,

(f) $S(y,\Sigma) > 0$ if and only if $y_j < 0$ for some $j = 1,\ldots,p$,

(g) For some $\chi > 0$, $S(ay,\Sigma) = a^\chi S(y,\Sigma)$ for any scalar $a > 0$.

Assumption A.3. The probability measure μ used in Eq. (4.1) has full support on $\mathcal{G}$.

Assumption A.4. For every $\theta \in \Theta$ and $(\bar{x},\bar{\nu}) \in \mathcal{G}$, $\lim_{B(x,\nu)\downarrow B(\bar{x},\bar{\nu})} E_F[M(Z,\theta,x,\nu)] = E_F[M(Z,\theta,\bar{x},\bar{\nu})]$, where the convergence $B(x,\nu) \downarrow B(\bar{x},\bar{\nu})$ occurs in the Hausdorff distance, i.e., $\sup_{a\in B(x,\nu)} \inf_{b\in B(\bar{x},\bar{\nu})} ||a - b|| \to 0$.

Assumption A.5. For any $s = 1,\ldots,p$, the triangular array of processes $\{\{M(Z_i,\theta,x,\nu) : (x,\nu) \in \mathcal{G}\}_{i=1}^n\}_{n\geq 1}$ is manageable with respect to the envelopes $\{\{M(Z_i,\theta)\}_{i=1}^n\}_{n\geq 1}$ in the sense of Pollard (1990, Definition 7.9).

Assumption A.6. $\{\kappa_n\}_{n\geq 1}$ and $\{B_n\}_{n\geq 1}$ are non-decreasing sequences of positive constants such that $n \to \infty$ implies that: (a) $\kappa_n \to \infty$, (b) $B_n/\kappa_n \to 0$, (c) $B_n \to \infty$, and (d) $\sqrt{n}/\kappa_n \to \infty$.

We now briefly explain each of these assumptions. Assumption A.2 combines Assumptions S1-S4 in AS13, who propose several candidates for S that satisfy all of these necessary conditions. For convenience, we describe two of these candidates that are already tailored to the setup of this paper. The first example is the modified method of moments (MMM) test function:

$$S_1(y,\Sigma) = \sum_{j=1}^{p} [y_j/\Sigma_{[j,j]}]^2_-, \tag{A.6}$$

where $[z]_- \equiv |z| \times 1(z < 0)$. The second example is the quasi-likelihood ratio (QLR) test function:

$$S_2(y,\Sigma) = \inf_{t\in\mathbb{R}^p_{+,\infty}} (y-t)'\Sigma^{-1}(y-t).$$

The measure μ is analogous to the weight function Q in AS13 and so Assumption A.3 corresponds to their Assumption Q.[15] By this assumption, any subset of $\mathcal{G}$ with positive Lebesgue measure will be assigned a positive probability. There are many possible candidates for this measure. For example, we could consider the following product measure:

$$\mu(x,\nu) = \prod_{j=1}^{d_x} \mu_1(x_j) \times \mu_2(\nu_j),$$

where μ_1 is any continuous distribution with full support on $\mathbb{R}$ (e.g. standard normal $N(0,1)$) and μ_2 is any continuous distribution with support on $(0,\overline{r})$.

Assumption A.4 is a smoothness assumption on the moment conditions that define our partially identified model. Recall from Eqs. (3.4) and (3.7) that $E_F[M(Z,\theta,x,\nu)]$ is the result of integrating a function on a box $B(x,\nu)$, whose center is x and whose width is determined by ν. Assumption A.4 requires that this the expectation changes continuously as we infinitesimally increase the size of the box $B(x,\nu)$. This assumption can be considered mild because it applies to an expectation, which is a smoothing operator. For example, it is satisfied in Example 1.1 provided that G is continuous.

Assumption A.5 is analogous to Assumption M(c) in AS13. This assumption provides a sufficient condition to obtain a functional version of the law of large numbers and the central limit theorem, which are the key to our inferential results.

Finally, Assumption A.6 specifies thresholding sequences that need to be chosen by the researcher in order to implement the GMS approximation. These sequences are typical in GMS type of inference (see, e.g., Andrews and Soares (2010) and Bugni (2010), among others). While Assumption A.6 restricts these sequences in terms of rates of convergence, they provide little guidance on how to choose them in practice for a given sample size. Based on experience drawn from their Monte Carlo simulation, AS13 (Page 643) recommend using $\kappa_n \equiv (0.3\ln(n))^{1/2}$ and $B_n \equiv (0.4\ln(n)/\ln\ln(n))^{1/2}$, which we use in our own simulations.

A.2.3 Results on identification

Our next result has the objective of providing a formal justification for our definition of the test function T_n in Eq. (4.1). Our confidence set is an example of the criterion function approach to inference in partially identified models developed by Chernozhukov et al. (2007).

A central population object in this approach is the so-called criterion function, denoted by $T_F(\theta) : \Theta \to \mathbb{R}_+$ with the defining property that it takes value of zero if and only if $\theta \in \Theta_S(F)$. The following result proposes a particular function and verifies that it is a criterion function for the current problem.

Theorem A.3. *Assume Assumptions A.2-A.4. For any $(\theta,F) \in \mathcal{F}$, define the following function:*

$$T_F(\theta) \equiv \int S(E_F[M(Z,\theta,x,\nu)], Var_F[M(Z,\theta,x,\nu)] + \lambda D_F(\theta)) d\mu(x,\nu),$$

where $D_F(\theta)$ is as in Eq. (A.4). Then, $T_F(\theta)$ is a population criterion function for $\Theta_S(F)$, i.e., $T_F(\theta) \geq 0$ and $T_F(\theta) = 0$ if and only if $\theta \in \Theta_S(F)$.

Proof. $T_F(\theta) \geq 0$ for $\theta \in \Theta$ follows directly from Assumptions A.2(c) and A.3. First, consider $\theta \in \Theta_S(F)$. By definition in Theorem 3.3, $\theta \in \Theta_S(F)$ implies $E_F[M(Z,\theta,x,\nu)] \geq \mathbf{0}$ for all $(x,\nu) \in \mathcal{G}$. Then, Assumptions A.2(c,f) and A.3 imply that $T_F(\theta) = 0$.

For the remainder of the proof, consider $\theta \notin \Theta_S(F)$. By definition in Theorem 3.3, $\theta \notin \Theta_S(F)$ implies $E_F[M_j(Z,\theta,\bar{x},\bar{\nu})] < 0$ for some $(j,(\bar{x},\bar{\nu})) \in \{1,\ldots,p\} \times \mathcal{G}$. Let $\epsilon \equiv |E_F[M_j(Z,\theta,\bar{x},\bar{\nu})]|/2$.

For any $\delta > 0$, define the set $A(\delta) \equiv \{(x,\nu) \in [\bar{x} - \bar{\nu}\delta, \bar{x} + \bar{\nu}\delta] \times [\bar{\nu}(1+2\delta), \bar{\nu}(1+3\delta)]\}$. We now verify that $A(\delta)$ has several properties. First, the fact that $\min\{\bar{\nu}_s\}_{s=1}^{d_x}\delta > 0$ implies that $A(\delta)$ has a positive Lebesgue measure. Second, $(x,\nu) \in A(\delta)$ implies that $x - \nu \in [\bar{x} - \bar{\nu} - 4\bar{\nu}\delta, \bar{x} - \bar{\nu} - \bar{\nu}\delta]$ and $x + \nu \in [\bar{x} + \bar{\nu} + \bar{\nu}\delta, \bar{x} + \bar{\nu} + 4\bar{\nu}\delta]$ and these, combined with $\min\{\bar{\nu}_s\}_{s=1}^{d_x}\delta > 0$, imply that $B(\bar{x},\bar{\nu}) \subseteq B(x,\nu)$. Third, $||(x,\nu) - (\bar{x},\bar{\nu})|| \leq 3\max\{\bar{\nu}_s\}_{s=1}^{d_x}\delta$ and this implies that $B(x,\nu) \downarrow B(\bar{x},\bar{\nu})$ (in Hausdorff distance) as $\delta \downarrow 0$. By $\epsilon \equiv |E_F[M_j(Z,\theta,\bar{x},\bar{\nu})]|/2$ and Assumption A.4, it follows that $\exists\bar{\delta}_1 > 0$ s.t. $E_F[M_j(Z,\theta,x,\nu)] \leq -\epsilon$ for all $\delta \in (0,\bar{\delta}_1)$. Finally, the fact that $\max\{\bar{\nu}_s\}_{s=1}^{d_x} < \bar{r}$ implies that $\exists\bar{\delta}_2 > 0$ s.t. $\max\{\bar{\nu}_s\}_{s=1}^{d_x}(1+3\delta) < \bar{r}$ and so $A(\delta) \subseteq \mathcal{G}$ for all $\delta \in (0,\bar{\delta}_2)$. For the rest of the proof, define $A \equiv A(\bar{\delta})$ for $\bar{\delta} \equiv \min\{\bar{\delta}_1,\bar{\delta}_2\} > 0$.

We now show that $\exists\bar{\eta} > 0$ s.t.

$$S(\ E_F[M(Z,\theta,x,\nu)]\ ,\ Var_F[M(Z,\theta,x,\nu)]\ +\ \lambda\ D_F(\theta)\ ) \geq \bar{\eta} \quad \forall(x,\nu) \in A. \tag{A.7}$$

By Assumption A.2(e) and the fact that $Var_F[M(Z,\theta,x,\nu)]$ is positive semi-definite, it suffices to show that $\exists\bar{\eta} > 0$ s.t. $S(E_F[M(Z,\theta,x,\nu)], \lambda D_F(\theta)) \geq \bar{\eta}\ \forall(x,\nu) \in A$. Suppose that this is not true, i.e., suppose that $\exists\{(x_s,\nu_s)\}_{s\geq1}$ with $(x_s,\nu_s) \in A\ \forall s \in \mathbb{N}$ s.t. $\lim_{s\to\infty} S(E_F[M(Z,\theta,x_s,\nu_s)], \lambda D_F(\theta)) = 0$. By the compactness of A, $\{(x_s,\nu_s)\}_{s\geq1}$ has a convergent subsequence in A with a limit point $(x^*,\nu^*) \in A$ s.t. $S(E_F[M(Z,\theta,x^*,\nu^*)], \lambda D_F(\theta)) = 0$, which is a contradiction to Assumption A.2(f) and $E_F[M_j(Z,\theta,x^*,\nu^*)] \leq -\epsilon$.

To conclude the proof, consider the following argument:

$$T_F(\theta) \geq \int_A S(\ E_F[M(Z,\theta,x,\nu)], Var_F[M(Z,\theta,x,\nu)] + \lambda D_F(\theta)\ ) d\mu(x,\nu) \geq \eta\mu(A) > 0,$$

where the first inequality holds by Assumptions A.2(c) and A.3, the second inequality holds by Eq. (A.7), and the strict inequality holds by Assumption A.3. □

A.2.4 Computation of GMS confidence sets

This paper considers confidence sets of the form:

$$CS_n = \{\theta \in \Theta : T_n(\theta) \leq \hat{c}_n(\theta, 1-\alpha)\}.$$

In practice, both the test statistic $T_n(\theta)$ and the GMS critical value $\hat{c}_n(\theta, 1-\alpha)$ require integration with respect to the probability measure μ. Furthermore, $c_n(\theta, 1-\alpha)$ also requires computation of quantiles of a certain Gaussian process. The objective of this section is to describe how to implement these approximations.

First of all, integrals with respect to probability measure μ can be approximated with arbitrary accuracy by Monte Carlo simulation, i.e., we draw an arbitrarily large sample:

$$\{(x_u,\nu_u)\}_{u=1}^{s_n} \text{ is i.i.d. and distributed according to } \mu(x,\nu), \tag{A.8}$$

and approximate the integral with a sample average. The quality of the approximation to these integrals is controlled by the number of random draws used, denoted by s_n and assumed to satisfy $s_n \to \infty$ as $n \to \infty$. Following AS13 (Sections 3.5 and 4.2), we only draw the sample according to Eq. (A.8) once and use it to approximate integrals in both $T_n(\theta)$ and $c_n(\theta, 1-\alpha)$ for all $\theta \in \Theta$.

Approximating the test function $T_n(\theta)$ in Eq. (4.1) is a matter of replacing the integral with a sample average. In particular, we use:

$$\overline{T}_{n,s_n}(\theta) \equiv \frac{1}{s_n}\sum_{u=1}^{s_n} S(n^{1/2}\overline{M}_n(\theta,x_u,\nu_u),\overline{\Sigma}_n(\theta,x_u,\nu_u)),$$

where $\{(x_u,\nu_u)\}_{u=1}^{s_n}$ is the i.i.d. sample in Eq. (A.8), and $\overline{M}_n(\theta,x,\nu)$ and $\overline{\Sigma}_n(\theta,x,\nu)$ are as in Eq. (4.2).

Approximating the GMS critical value is slightly more involved. We provide two algorithms that can be used to approximate $\hat{c}_n(\theta,1-\alpha)$, referred to as asymptotic approximation and bootstrap. Both algorithms approximate integrals by Monte Carlo integration but differ in the method used to approximate the Gaussian process. In both of these algorithms, the quality of the approximation is controlled by the number of repetitions involved, denoted by τ_{reps} and assumed to satisfy $\tau_{reps}\to\infty$ as $n\to\infty$.

Approximation of $\hat{c}_n(\theta,1-\alpha)$ by simulation.

1. Draw an i.i.d. sample $\{\{\zeta_{\tau,i}\}_{i=1}^{n}\}_{\tau=1}^{\tau_{reps}}$ where $\zeta_{\tau,i}\sim N(0,1)$.

2. For each $\tau=1,\ldots,\tau_{reps}$ and $u=1,\ldots,s_n$, define

$$v_\tau(\theta,x_u,\nu_u)\equiv n^{-1/2}\sum_{i=1}^{n}\zeta_{\tau,i}\times D_n^{-1/2}(\theta)(M(Z_i,\theta,x_u,\nu_u)-\overline{M}_n(\theta,x_u,\nu_u)).$$

   where $\hat{\Sigma}_n(\theta)$ and $D_n(\theta)$ are as in Eq. (A.5), and $\{(x_u,\nu_u)\}_{u=1}^{s_n}$ is the i.i.d. sample in Eq. (A.8).

3. For each $\tau=1,\ldots,\tau_{reps}$, compute the sample $\overline{T}_{s_n,\tau}(\theta)$ as follows:

$$\overline{T}_{s_n,\tau}(\theta)=s_n^{-1}\sum_{u=1}^{s_n}S(v_\tau(\theta,x_u,\nu_u)+\varphi_n(\theta,x_u,\nu_u),\hat{h}_{2,n}(\theta,x_u,\nu_u)+\lambda\mathbf{I}_{p\times p}),$$

   where $\hat{h}_{2,n}$ and φ_n are as in Eq. (4.4).

4. $\hat{c}_n(\theta,1-\alpha)$ is approximated by η plus the empirical $(1-\alpha+\eta)$-quantile of $\{\overline{T}_{s_n,\tau}(\theta)\}_{\tau=1}^{\tau_{reps}}$.

Approximation of $\hat{c}_n(\theta,1-\alpha)$ by the bootstrap.

1. Draw an i.i.d. sample $\{\{Z_{\tau,i}^*\}_{i=1}^{n}\}_{\tau=1}^{\tau_{reps}}$ where $Z_{\tau,i}^*$ is a bootstrap draw from the empirical distribution of $\{Z_i\}_{i=1}^{n}$.

2. For each $\tau=1,\ldots,\tau_{reps}$ and $u=1,\ldots,s_n$, define

$$v_\tau^*(\theta,x_u,\nu_u)\equiv n^{-1/2}\sum_{i=1}^{n}\hat{D}_n^{-1/2}(\theta)(M(Z_{\tau,i}^*,\theta,x_u,\nu_u)-\overline{M}_n(\theta,x_u,\nu_u)).$$

   where $\hat{\Sigma}_n(\theta)$ and $D_n(\theta)$ are as in Eq. (A.5), and $\{(x_u,\nu_u)\}_{u=1}^{s_n}$ is the i.i.d. sample in Eq. (A.8).

3. For each $\tau=1,\ldots,\tau_{reps}$, compute the sample $\overline{T}_{s_n,\tau}(\theta)$ as follows:

$$\overline{T}_{s_n,\tau}(\theta)=s_n^{-1}\sum_{u=1}^{s_n}S(v_\tau(\theta,x_u,\nu_u)+\varphi_n(\theta,x_u,\nu_u),\hat{h}_{2,n}(\theta,x_u,\nu_u)+\lambda\mathbf{I}_{p\times p}),$$

   where $\hat{h}_{2,n}$ and φ_n are as in Eq. (4.4).

4. $\hat{c}_n(\theta, 1-\alpha)$ is approximated by η plus the empirical $(1-\alpha+\eta)$-quantile of $\{\overline{T}_{s_n,\tau}(\theta)\}_{\tau=1}^{\tau_{reps}}$.

Notes

[1] See Arcidiacono et al. (2012) for more evidence and claims about the non-reporting in these data.

[2] By the arguments in Domínguez and Lobato (2004), the methods described in HM06 or BMM11 would address our identification problem if one could apply their methodology to a model with an infinite number of unconditional moment inequalities. Unfortunately, neither of these methods are computationally feasible in this situation. In the case of HM06, see the discussion in Horowitz et al. (2003). In the case of BMM11, an infinite number of moment inequalities implies that the number of terms in the objective function of their optimization problem becomes computationally unmanageable.

[3] As previously explained, neither HM06 nor BMM11 allow for an infinite number of unconditional moment inequalities.

[4] To be specific, Eq. (2.2) is equivalent to Eq. (2.1) with g_1 and g_2 defined as follows. If $x = x_j$ for some $j = 1, \ldots, N$, $g_1(x) \equiv \gamma_{1,j}$, $g_2(1,x) \equiv \gamma_{2,j}/\gamma_{1,j}$, and $g_2(0,x) \equiv 1 - \gamma_{2,j}$. For any other x or any $y \notin \{0,1\}$, $g_1(x) = g_2(y,x) = 0$.

[5] If the set $S_{X_2} \cap B_2(x_2, \nu_2)$ is empty, then two things occur. First, the associated inf and sup are equal to ∞ and $-\infty$, respectively. Second, the indicator function multiplying these expressions equals zero. Here and throughout the paper, we define $\infty \times 0 \equiv 0$. Consequently, $S_{X_2} \cap B_2(x_2, \nu_2)$ being empty implies that the associated expression in Eq. (3.4) equals zero.

[6] As explained in the introduction, our results can be generalized to allow for arbitrary missing data patters on both outcomes and covariates. The corresponding outer identified set can be deduced directly from Theorem 3.3 if we replace M_1 and M_2 defined as in Eqs. (3.4) and (3.7) with the corresponding functions defined in Theorems A.1 and A.2, respectively.

[7] These are developed and discussed in Andrews et al. (2004), Imbens and Manski (2004), Galichon and Henry (2006, 2013), Chernozhukov et al. (2007), Beresteanu and Molinari (2008), Romano and Shaikh (2008, 2010), Rosen (2008), Andrews and Guggenberger (2009), Stoye (2009), Andrews and Soares (2010), Bugni (2010, 2015), Canay (2010), Andrews and Jia-Barwick (2012), Bontemps et al. (2012), and Pakes et al. (2014). In fact, these references could be applied to our problem without loss of information if the conditioning covariate had finite support.

[8] This positive constant controls the amount of modification introduced in the computation of the sample variance of $\{M(Z_i, \theta, x, \nu)\}_{i=1}^{n}$. Following AS13 (Page 644), we implement our results with $\lambda = 5\%$.

[9] This is a universal uniformity factor used to circumvent problems that arise due to the presence of the infinite-dimensional nuisance parameter associated to the slackness of the moment conditions. Following AS13 (Page 644), we implement our results with $\eta = 10^{-6}$.

[10] See Imbens and Manski (2004), Andrews and Guggenberger (2009), Andrews and Soares (2010), and AS13 (Section 5.1).

[11] Since $\Theta_S(F)$ is a superset of the sharp identified set $\Theta_I(F)$, there could be parameter values that belong to $\Theta_S(F)$ and lie outside of $\Theta_I(F)$. Even though these parameter values cannot (logically) correspond to the true parameter value, our inference method (based on outer identification) will not have any (non-trivial) power against them, even asymptotically.

[12] See explanation below on how the missing data patterns of X and Y translate into the value of W_X and W_Y, respectively.

[13] The analogous result for missing outcome data is not conceptually hard but it is cumbersome to express. It is available from the authors upon request.

[14] As explained in AS13, this restriction is not particularly problematic in practice, as the potential uniformity problems arise because the limiting distribution of the test statistic is discontinuous in the slackness of the moment inequalities and not its variance-covariance kernel.

[15] To be precise, our measure μ corresponds exactly to their measure Q^*.

References

Abrevaya, J. (2001): "The effects of demographics and maternal behavior on the distribution of birth outcomes," *Empirical Economics*, 26, 247–257.

Andrews, D. W. K., S. Berry, and P. Jia-Barwick (2004): "Confidence Regions for Parameters in Discrete Games with Multiple Equilibria with an Application to Discount Chain Store Location," Mimeo: Yale University and M.I.T.

Andrews, D. W. K. and P. Guggenberger (2009): "Validity of Subsampling and "Plug-in Asymptotic" Inference for Parameters Defined by Moment Inequalities," *Econometric Theory*, 25, 669–709.

Andrews, D. W. K. and P. Jia-Barwick (2012): "Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure," *Econometrica*, 80, 2805–2826.

Andrews, D. W. K. and X. Shi (2013): "Inference Based on Conditional Moment Inequalities," *Econometrica*, 81, 609–666.

Andrews, D. W. K. and G. Soares (2010): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometrica*, 78, 119–157.

Arcidiacono, P., E. Aucejo, P. Coate, and V. J. Hotz (2012): "Affirmative Action and University Fit: Evidence from Proposition 209," NBER working paper #18523.

Armstrong, T. B. (2012): "Asymptotically Exact Inference in Conditional Moment Inequality Models," Mimeo: Yale University.

——— (2014): "Weighted KS Statistics for Inference on Conditional Moment Inequalities," *Journal of Econometrics*, 181, 92–116.

Armstrong, T. B. and H. P. Chan (2012): "Multiscale Adaptive Inference on Conditional Moment Inequalities," Mimeo: Yale University and National University of Singapore.

Aucejo, E. M., F. A. Bugni, and V. J. Hotz (2015a): "The Effect of an Affirmative Action Ban on Graduation Rates at the University of California," Mimeo: London School of Economics and Duke University.

——— (2015b): "Online Supplemental Material to "Identification and Inference on Regressions with Missing Covariate Data"," Mimeo: London School of Economics and Duke University.

Beresteanu, A. and F. Molinari (2008): "Asymptotic Properties for a Class of Partially Identified Models," *Econometrica*, 76, 763–814.

Beresteanu, A., F. Molinari, and I. Molchanov (2011): "Sharp Identification Regions in Models with Convex Predictions," *Econometrica*, 79, 1785–1821.

Bontemps, C., T. Magnac, and E. Maurin (2012): "Set Identified Linear Models," *Econometrica*, 80, 1129–1155.

Bugni, F. A. (2010): "Bootstrap Inference in Partially Identified Models Defined by Moment Inequalities: Coverage of the Identified Set," *Econometrica*, 78, 735–753.

——— (2015): "A comparison of inferential methods in partially identified models in terms of error in coverage probability (Formerly circulated as "Bootstrap Inference in Partially Identified Models Defined by Moment Inequalities: Coverage of the Elements of the Identified Set")," *Econometric Theory*, Firstview, 1–56.

Bugni, F. A., I. A. Canay, and X. Shi (2015): "Specification Tests for Partially Identified Models defined by Moment Inequalities," *Journal of Econometrics*, 185, 259–282.

Canay, I. A. (2010): "E.L. Inference for Partially Identified Models: Large Deviations Optimality and Bootstrap Validity," *Journal of Econometrics*, 156, 408–425.

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and Confidence Regions for Parameter Sets in Econometric Models," *Econometrica*, 75, 1243–1284.

Chernozhukov, V., S. Lee, and A. M. Rosen (2013): "Intersection Bounds: Estimation and Inference," *Econometrica*, 81, 667–737.

Chesher, A. and A. Rosen (2014a): "Generalized Instrumental Variable Models," Mimeo: CeMMAP working paper CWP04/14.

——— (2014b): "An Instrumental Variable Random Coefficients Model for Binary Outcomes," *Econometrics Journal*, 17, S1–S19.

Chesher, A., A. Rosen, and K. Smolinski (2013): "An Instrumental Variable Model for Multiple Discrete Choice," *Quantitative Economics*, 4, 157–196.

Chetverikov, D. (2012): "Adaptive Test of Conditional Moment Inequalities," Unpublished manuscript.

Domínguez, M. and I. N. Lobato (2004): "Consistent Estimation of Models Defined by Conditional Moment Restrictions," *Econometrica*, 72, 1601–1615.

Galichon, A. and M. Henry (2006): "Inference in Incomplete Models," Mimeo: Ecole Polytechnique, Paris - Department of Economic Sciences and Pennsylvania State University.

——— (2011): "Set Identification in Models with Multiple Equilibria," *Journal of Econometrics*, 78, 1264–1298.

——— (2013): "Dilation Bootstrap: A methodology for constructing confidence regions with partially identified models," *Journal of Econometrics*, 177, 109–115.

Horowitz, J. L. and C. F. Manski (1995): "Identification and Robustness with Contaminated and Corrupted Data," *Econometrica*, 63, 281–302.

——— (1998): "Censoring of Outcomes and Regressors due to Survey Nonresponse: Identification and Estimation using Weights and Imputations," *Journal of Econometrics*, 84, 37–58.

——— (2000): "Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data," *Journal of the American Statistical Association*, 95, 77–88.

——— (2006): "Identification and estimation of statistical functionals using incomplete data," *Journal of Econometrics*, 132, 445–459.

Horowitz, J. L., C. F. Manski, M. Ponomareva, and J. Stoye (2003): "Computation of Bounds on Population Parameters When the Data Are Incomplete," *Reliable Computing*, 9, 419–440.

Imbens, G. and C. F. Manski (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72, 1845–1857.

Kim, K. (2008): "Set Estimation and Inference with Models Characterized by Conditional Moment Inequalities," Mimeo: Michigan State University.

Lewbel, A. (2002): "Estimation of Average Treatment Effects with Misclassification," *Econometrica*, 75, 537–551.

Lundberg, S. (1988): "Labor Supply of Husbands and Wives: A Simultaneous Equations Approach," *The Review of Economics and Statistics*, 70, 224–235.

Mahajan, A. (2006): "Identification and Estimation of Regression Models with Missclassification," *Econometrica*, 74, 631–665.

Manski, C. F. (2003): *Partial Identification of Probability Distributions*, Springer-Verlag.

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

Molinari, F. (2008): "Partial Identification of Probability Distributions with Misclassified Data," *Journal of Econometrics*, 144, 81–117.

Pakes, A., J. Porter, K. Ho, and J. Ishii (2014): "Moment Inequalities and Their Application," Forthcoming in Econometrica, Working Paper.

Pollard, D. (1990): *Empirical Processes: Theory and Applications*, NSF-CBMS Regional Conference Series in Probability and Statistics, Volume 2.

Ponomareva, M. (2010): "Inference in Models Defined by Conditional Moment Inequalities with Continuous Covariates," Mimeo: Northern Illinois University.

Romano, J. P. and A. M. Shaikh (2008): "Inference for Identifiable Parameters in Partially Identified Econometric Models," *Journal of Statistical Planning and Inference*, 138, 2786–2807.

——— (2010): "Inference for the Identified Set in Partially Identified Econometric Models," *Econometrica*, 78, 169–211.

Rosen, A. M. (2008): "Confidence Sets for Partially Identified Parameters that Satisfy a Finite Number of Moment Inequalities," *Journal of Econometrics*, 146, 107–117.

Stoye, J. (2009): "More on Confidence Intervals for Partially Identified Parameters," *Econometrica*, 77, 299–1315.

Online Supplemental Material to Aucejo et al. (2015)

Esteban M. Aucejo
Department of Economics
London School of Economics
e.m.aucejo@lse.ac.uk

Federico A. Bugni
Department of Economics
Duke University
federico.bugni@duke.edu

V. Joseph Hotz
Department of Economics
Duke University, NBER, and IZA
hotz@econ.duke.edu

June 12, 2015

Abstract

This supplement contains the proofs of Theorems 4.1 and 4.2 of Aucejo et al. (2015, Section 4), along with two intermediate results that are required for these proofs. All definitions required for this supplement can be found in Aucejo et al. (2015). Finally, the derivations in this supplement follow closely results in Andrews and Shi (2013) (hereafter, referred to as AS13).

S1 Results on inference

Proof of Theorem 4.1. The proof of this result follows closely the arguments in AS13 (Theorem 2(a)). Notice that Assumption A.2 implies their Assumptions S1-S2, Assumption A.5 implies the manageability of the stochastic processes implied by their Assumption M, and Assumption A.6 implies their Assumption GMS1.

Suppose that Eq. (4.6) does not hold. In this case, we can find a subsequence $\{a_n\}_{n\geq 1}$ of $\{n\}_{n\geq 1}$ and a sequence $\{(\theta_{a_n}, F_{a_n}) \in \bar{\mathcal{F}}_0\}_{n\geq 1}$ s.t. $P_{F_{a_n}}(\theta_{a_n} \notin CS_{a_n}) > \alpha \ \forall n \in \mathbb{N}$. By the compactness implicit in the definition of $\bar{\mathcal{F}}_0$, we can find a further subsequence $\{b_n\}_{n\geq 1}$ of $\{a_n\}_{n\geq 1}$ s.t. $\{(\theta_{b_n}, F_{b_n}) \in \bar{\mathcal{F}}_0\}_{n\geq 1} \in SubSeq(h_2)$ for some limiting variance-covariance kernel h_2, where $SubSeq(h_2)$ is as in Definition A.4. By this and Assumption A.5, Lemmas S2.1-S2.2 imply that:

$$\begin{pmatrix} v_{b_n,F_{b_n}}(\theta_{b_n},\cdot) \\ \hat{h}_{2,b_n,F_{b_n}}(\theta_{b_n},\cdot) \end{pmatrix} \overset{d}{\to} \begin{pmatrix} v_{h_2}(\cdot) \\ h_2(\cdot) \end{pmatrix}$$

as stochastic processes indexed by $(x,\nu) \in \mathcal{G}$. This and Assumptions A.2 and A.6 allow us to establish AS13 (Lemmas A2-A5). In turn, these can be used to contradict $P_{F_{b_n}}(\theta_{b_n} \notin CS_{b_n}) > \alpha \ \forall n \in \mathbb{N}$, thus concluding the proof. □

Proof of Theorem 4.2. The proof of this result follows closely the arguments in AS13 (Theorem 3), with the exception of certain steps. For the sake of completeness, we sketch the main steps of the proof and point out the differences with the one in AS13.

Consider the following derivation:

$$\begin{aligned} P_F(\theta \in CS_n) &= P_F(T_n(\theta) \leq c(\varphi_n(\theta,\cdot), \hat{h}_{2,n}(\theta,\cdot), 1-\alpha+\eta)+\eta) \\ &\leq P_F(T_n(\theta) \leq c(\mathbf{0}, \hat{h}_{2,n}(\theta,\cdot), 1-\alpha+\eta)+\eta) \\ &= P_F(n^{-\chi/2}T_n(\theta) \leq n^{-\chi/2}(c(\mathbf{0}, \hat{h}_{2,n}(\theta,\cdot), 1-\alpha+\eta)+\eta)), \end{aligned}$$

where the first line holds by definition of $\hat{c}_n(\theta, 1-\alpha)$, the second line holds by definition of $\varphi_n(\theta,\cdot)$ and $c(\cdot, \hat{h}_{2,n}(\theta,\cdot), 1-\alpha+\eta)$, combined with Assumptions A.2(b) and A.6, which imply that $\varphi_n(\theta,\cdot) \geq \mathbf{0}$, and in the last line χ is as in Assumption A.2(g). The proof is completed by showing that (a) $P_F(n^{-\chi/2}T_n(\theta) \geq C) \to 1$ for some $C > 0$ and (b) $c(\mathbf{0}, \hat{h}_{2,n}(\theta,\cdot), 1-\alpha+\eta) = O_p(1)$, which imply that $n^{-\chi/2}(c(\mathbf{0}, \hat{h}_{2,n}(\theta,\cdot), 1-\alpha+\eta)+\eta) = o_p(1)$. The proof of (b) is identical to the proof in AS13 (which requires our Assumptions A.2 and A.5). On the other hand, our proof of (a) is slightly different, and so we devote the remainder of this proof to develop this argument.

By definition, $\theta \notin \Theta_S(F)$ implies that $\exists j \leq p$ s.t. $E_F[M_j(Z,\theta,x,\nu)] < 0$ for some $(x,\nu) \in \mathcal{G}$. Under Assumptions A.2(c,e,f) and A.4, we can use the arguments in the proof of Theorem A.3 to define a set $A \subset \mathcal{G}$ with positive Lebesgue measure s.t. $E_F[M_j(Z,\theta,x,\nu)] \leq -\epsilon \ \forall (x,\nu) \in A$. As a consequence,

$$S(E_F[M(Z,\theta,x,\nu)], Var_F[M(Z,\theta,x,\nu)] + \lambda D_F(\theta)) \geq \eta \ \ \forall (x,\nu) \in A$$

for some $\delta > 0$. By Assumptions A.2(a) and A.3, this implies that:

$$\int_{(x,\nu)\in A} S(D_F^{-1/2}(\theta)E_F[M(Z,\theta,x,\nu)], h_{2,F}(\theta,x,\nu) + \lambda I_{p\times p})d\mu(x,\nu) \geq \eta\mu(A) > 0. \quad \text{(S1.1)}$$

To complete the proof, consider the following derivation:

$$\begin{aligned} n^{-\chi/2}T_n(\theta) &= n^{-\chi/2}\int_{(x,\nu)\in\mathcal{G}} S(v_{n,F}(\theta,x,\nu)+h_{1,n,F}(\theta,x,\nu),\hat{h}_{2,n,F}(\theta,x,\nu)+\lambda I_{p\times p})d\mu(x,\nu) \\ &= \int_{(x,\nu)\in\mathcal{G}} S(n^{-1/2}v_{n,F}(\theta,x,\nu)+D_F^{-1/2}(\theta)E_F[M(Z,\theta,x,\nu)],\hat{h}_{2,n,F}(\theta,x,\nu)+\lambda I_{p\times p})d\mu(x,\nu) \\ &\geq \int_{(x,\nu)\in A} S(n^{-1/2}v_{n,F}(\theta,x,\nu)+D_F^{-1/2}(\theta)E_F[M(Z,\theta,x,\nu)],\hat{h}_{2,n,F}(\theta,x,\nu)+\lambda I_{p\times p})d\mu(g) \\ &\overset{p}{\to} \int_{(x,\nu)\in A} S(D_F^{-1/2}(\theta)E_F[M(Z,\theta,x,\nu)],h_{2,F}(\theta,x,\nu)+\lambda I_{p\times p})d\mu(x,\nu) \geq \eta\mu(A) > 0, \quad \text{(S1.2)} \end{aligned}$$

where the first line holds by definition of $T_n(\theta)$, the second line holds by Assumption A.2(g) and by definition of $h_{1,n,F}(\theta,x,\nu)$, the third line holds by $A\subset\mathcal{G}$ and Assumption A.2(c), the convergence in the fourth line holds by the same argument described in the next paragraph, and the last expression is positive by Eq. (S1.1). By Eq. (S1.2), $P_F(n^{-\chi/2}T_n(\theta)\geq\eta\mu(A)/2)\to 1$ for some $\eta\mu(A)/2>0$, which implies the desired result.

To conclude the proof, it suffices to justify the convergence in the fourth line of Eq. (S1.2). For a fixed parameter $(\theta,F)\in\mathcal{F}$, Lemmas S2.1-S2.2 (see Section S2 in this supplement) imply that:

$$\begin{pmatrix} v_{n,F}(\theta,\cdot) \\ \hat{h}_{2,n,F}(\theta,\cdot) \end{pmatrix} \overset{d}{\to} \begin{pmatrix} v_{h_{2,F}}(\theta,\cdot) \\ h_{2,F}(\theta,\cdot) \end{pmatrix}$$

as stochastic processes indexed by $(x,\nu)\in\mathcal{G}$. In turn, this implies that:

$$\sup_{(x,\nu)\in\mathcal{G}} \left\| \begin{pmatrix} n^{-1/2}v_{n,F}(\theta,x,\nu) \\ \hat{h}_{2,n,F}(\theta,x,\nu) \end{pmatrix} - \begin{pmatrix} \mathbf{0} \\ h_{2,F}(\theta,x,\nu) \end{pmatrix} \right\| \overset{p}{\to} 0.$$

The convergence in the fourth line of Eq. (S1.2) is a result of this, the almost sure representation theorem, the bounded convergence theorem, and Assumption A.2(d). □

S2 Auxiliary results

Lemma S2.1. *Assume Assumption A.5 and that $\{(\theta_{k_n},F_{k_n})\in\bar{\mathcal{F}}_0\}_{n\geq1}\in SubSeq(h_2)$ for an arbitrary subsequence $\{k_n\}_{n\geq1}$ of $\{n\}_{n\geq1}$. Then,*

$$v_{k_n,F_{k_n}}(\theta_{k_n},\cdot)\overset{d}{\to} v_{h_2}(\cdot),$$

as stochastic processes indexed by $(x,\nu)\in\mathcal{G}$, where v_{h_2} is a $\mathbb{R}^p$-valued Gaussian process with zero mean and variance-covariance kernel $h_2(\cdot,\cdot)$ on $\mathcal{G}\times\mathcal{G}$.

Proof. This result follows from AS13 (Lemmas A1(a) and E3). We describe the main ideas behind these arguments for the sake of completeness. Throughout this proof, we replace the subsequence $\{k_n\}_{n\geq1}$ by the original sequence $\{n\}_{n\geq1}$ in order to simplify the notation.

Suppose that $\{(\theta_n,F_n)\in\bar{\mathcal{F}}_0\}_{n\geq1}\in SubSeq(h_2)$. By Pollard (1990, Theorem 10.2), the desired result is a consequence of the following conditions:

(1) $(\mathcal{G}, \rho)$ is a totally bounded pseudo-metric space, where ρ is the following pseudo-metric:

$$\rho^2((x,\nu),(\tilde{x},\tilde{\nu})) \;\equiv\; \lim_{n\to\infty}(Trace(Var_{F_n}[D_{F_n}^{-1/2}(\theta_n)(M(Z,\theta_n,x,\nu)-M(Z,\theta_n,\tilde{x},\tilde{\nu})]),$$

(2) The finite dimensional convergence holds, i.e., $\forall(a,L) \in \mathbb{R}^p/\mathbf{0}\times\mathbb{N}$ and $\forall\{(x_s,\nu_s)\}_{s=1}^L \subset \mathcal{G}$, $\{a'v_{n,F_n}(\theta_n,x_s,\nu_s)\}_{s=1}^L$ converges in distribution to an L-dimensional Gaussian distribution with zero mean and variance covariance matrix with (s_1,s_2) component given by $a'h_2((x_{s_1},\nu_{s_1}),(x_{s_2},\nu_{s_2}))a$.

(3) $\{v_{n,F_n}(\theta_n,x,\nu):(x,\nu)\in\mathcal{G}\}_{n\geq 1}$ is stochastically equicontinuous with respect to ρ.

To prove these conditions, AS13 use the Crámer-Wold device. In particular, AS13 (Lemma A1(a)) shows that these conditions hold if, for all $a\in\mathbb{R}^p/\mathbf{0}$, the following three conditions hold:

(a) $(\mathcal{G},\rho_a)$ is a totally bounded pseudo-metric space, where ρ_a is the following pseudo-metric:

$$\rho_a^2((x,\nu),(\tilde{x},\tilde{\nu})) \equiv \lim_{n\to\infty} Var_{F_n}[D_{F_n}^{-1/2}(\theta_n)a'(M(Z,\theta_n,x,\nu)-M(Z,\theta_n,\tilde{x},\tilde{\nu})], \tag{S2.1}$$

(b) The finite dimensional convergence holds, i.e., $\forall L$ and $\forall\{(x_s,\nu_s)\}_{s=1}^L\subset\mathcal{G}$, $\{a'v_{n,F_n}(\theta_n,x_s,\nu_s)\}_{s=1}^L$ converges in distribution to an L-dimensional Gaussian distribution with zero mean and variance covariance matrix with (s_1,s_2) component given by $a'h_2((x_{s_1},\nu_{s_1}),(x_{s_2},\nu_{s_2}))a$. This convergence uniquely determines a Gaussian distribution v_a concentrated on the space of uniformly $\rho_a(\cdot)$-continuous bounded functionals on $\mathcal{G}$, $U_{\rho_a}(\mathcal{G})$,

(c) $a'v_{n,F_n}(\theta_n,\cdot)\overset{d}{\to}v_a$.

To prove conditions (a)-(c), we rely on AS13 (Lemma E3), which extends Pollard (1990, Theorem 10.6, page 53) to triangular array stochastic processes. Fix $a\in\mathbb{R}^p/\mathbf{0}$ and $(x,\nu),(\tilde{x},\tilde{\nu})\in\mathcal{G}$ arbitrarily and define:

$$\begin{aligned} f_{a,n,i}(\omega,x,\nu) &\equiv n^{-1/2}a'D_{F_n}^{-1/2}(\theta_n)(M_n(Z_i,\theta_n,x,\nu)-E_{F_n}[M_n(Z_i,\theta_n,x,\nu)]),\\ \rho_{n,a}^2((x,\nu),(\tilde{x},\tilde{\nu})) &\equiv n\,E_{F_n}[f_{a,n,i}(\omega,x,\nu)-f_{a,n,i}(\omega,\tilde{x},\tilde{\nu})]^2. \end{aligned} \tag{S2.2}$$

By definition, notice that $a'v_{n,F_n}(\theta_n,x,\nu)=\sum_{i=1}^n f_{a,n,i}(\omega,x,\nu)$. AS13 (Lemma E3) show that conditions (a)-(c) hold provided that, $\forall a\in\mathbb{R}^p/\mathbf{0}$, the following results hold:

(i) $\{f_{a,n,i}(\omega,x,\nu):(x,\nu)\in\mathcal{G}\}_{i=1}^n$ is manageable with respect to some envelopes $\{F_{a,n,i}(\omega)\}_{i=1}^n$,

(ii) $\lim_{n\to\infty}E_{F_n}[f_{a,n,i}(\omega,x,\nu)f_{a,n,i}(\omega,\tilde{x},\tilde{\nu})]=a'h_2((x,\nu),(\tilde{x},\tilde{\nu}))a$ for all $(x,\nu),(\tilde{x},\tilde{\nu})\in\mathcal{G}$,

(iii) $\limsup_{n\to\infty}\sum_{i=1}^n E_{F_n}[F_{a,n,i}^2]<\infty$,

(iv) $\lim_{n\to\infty}\sum_{i=1}^n E_{F_n}[F_{a,n,i}^2 1[F_{a,n,i}>\varepsilon]]=0$ for all $\varepsilon>0$,

(v) The pseudo-metric ρ_a in Eq. (S2.1) satisfies $\rho_a((x,\nu),(\tilde{x},\tilde{\nu}))\equiv\lim_{n\to\infty}\rho_{n,a}((x,\nu),(\tilde{x},\tilde{\nu}))$ for all $(x,\nu),(\tilde{x},\tilde{\nu})\in\mathcal{G}$ and, for all deterministic sequences $\{(x_n,\nu_n)\in\mathcal{G}\}_{n\geq1}$ and $\{(\tilde{x}_n,\tilde{\nu}_n)\in\mathcal{G}\}_{n\geq1}$, $\rho_a((x_n,\nu_n),(\tilde{x}_n,\tilde{\nu}_n))\to0$ implies that $\rho_{n,a}((x_n,\nu_n),(\tilde{x}_n,\tilde{\nu}_n))\to0$,

The verification of these conditions is similar to that in AS13.

Condition (i). By Assumption A.5, $\{a'M(Z_i,\theta,x,\nu) : (x,\nu) \in \mathcal{G}\}_{i=1}^n$ is manageable with respect to the envelopes $\{a'M(Z_i,\theta)\}_{i=1}^n$. By the definitions in Eq. (S2.2) and AS13 (Lemma E1), it then follows that $\{f_{a,n,i}(\omega,x,\nu) : (x,\nu) \in \mathcal{G}\}_{i=1}^n$ is manageable with respect to envelopes $\{F_{a,n,i}(\omega)\}_{i=1}^n$ defined as follows:

$$F_{a,n,i}(\omega) \equiv n^{-1/2} a' D_{F_n}^{-1/2}(\theta_n)(M_n(Z_i,\theta_n) + E_{F_n}[M_n(Z_i,\theta_n)]).$$

Condition (ii)-(v). While the definitions of our stochastic processes and envelopes are slightly different from those in AS13, one can still complete this proof by using similar arguments to those in AS13 (Lemma E3). □

Lemma S2.2. *Assume Assumption A.5 and that $\{(\theta_{k_n}, F_{k_n}) \in \bar{\mathcal{F}}_0\}_{n\geq 1} \in SubSeq(h_2)$ for an arbitrary subsequence $\{k_n\}_{n\geq 1}$ of $\{n\}_{n\geq 1}$. Then,*

$$\sup_{(x_n,\nu_n),(\tilde{x}_n,\tilde{\nu}_n)\in\mathcal{G}} ||\hat{h}_{2,k_n,F_{k_n}}(\theta_{k_n},(x_n,\nu_n),(\tilde{x}_n,\tilde{\nu}_n)) - h_2((x_n,\nu_n),(\tilde{x}_n,\tilde{\nu}_n))|| \xrightarrow{p} 0.$$

Proof. This result follows from AS13 (Lemmas A1(b)). We describe the main ideas behind these arguments for the sake of completeness. Throughout this proof, we replace the subsequence $\{k_n\}_{n\geq 1}$ by the original sequence $\{n\}_{n\geq 1}$ in order to simplify the notation.

Consider the following derivation:

$$\sup_{(x,\nu),(\tilde{x},\tilde{\nu})\in\mathcal{G}} ||\hat{h}_{2,n,F_n}((x,\nu),(\tilde{x},\tilde{\nu})) - h_2((x,\nu),(\tilde{x},\tilde{\nu}))|| \leq \left\{ \begin{array}{c} \sup_{(x,\nu),(\tilde{x},\tilde{\nu})\in\mathcal{G}} ||\hat{h}_{2,n,F_n}((x,\nu),(\tilde{x},\tilde{\nu})) - h_{2,F_n}((x,\nu),(\tilde{x},\tilde{\nu}))|| \\ +\sup_{(x,\nu),(\tilde{x},\tilde{\nu})\in\mathcal{G}} ||h_{2,F_n}((x,\nu),(\tilde{x},\tilde{\nu})) - h_2((x,\nu),(\tilde{x},\tilde{\nu}))|| \end{array} \right\}.$$

The RHS is a sum of two terms. By $\{(\theta_n,F_n) \in \bar{\mathcal{F}}_0\}_{n\geq 1} \in SubSeq(h_2)$, the second term converges to zero. Hence, it suffices to show that the first term is $o_p(1)$.

For any $s_1, s_2 = 1,\ldots,p$, the (s_1,s_2)-component of $\hat{h}_{2,n,F_n}((x,\nu),(\tilde{x},\tilde{\nu}))$ is given by:

$$\begin{aligned} &\hat{h}_{2,n,F_n}((x,\nu),(\tilde{x},\tilde{\nu}))_{(s_1,s_2)} \\ &= n^{-1}\sigma_{s_1}^{-1}(\theta_n)\sigma_{s_2}^{-1}(\theta_n)\sum_{i=1}^n (M_{s_1}(Z_i,\theta_n,x,\nu) - \bar{M}_{n,s_1}(\theta_n,x,\nu))(M_{s_2}(Z_i,\theta_n,\tilde{x},\tilde{\nu}) - \bar{M}_{n,s_2}(\theta_n,\tilde{x},\tilde{\nu})) \\ &= n^{-1}\sum_{i=1}^n f_{n,i,s_1,s_2}^{mm}(\omega,(x,\nu),(\tilde{x},\tilde{\nu})) - \left(n^{-1}\sum_{i=1}^n f_{n,i,s_1}^m(\omega,x,\nu)\right)\left(n^{-1}\sum_{i=1}^n f_{n,i,s_2}^m(\omega,\tilde{x},\tilde{\nu})\right). \end{aligned}$$

where we have relied on the i.i.d. assumption implicit in $(\theta_n,F_n) \in \bar{\mathcal{F}}_0$ and the following definitions:

$$\begin{aligned} f_{n,i,s}^m(\omega,x,\nu) &\equiv M_s(Z_i,\theta_n,x,\nu) - E_{F_n}[M_s(Z_i,\theta_n,x,\nu)], \\ f_{n,i,s,\check{s}}^{mm}(\omega,(x,\nu),(\tilde{x},\tilde{\nu})) &\equiv f_{n,i,s}^m(\omega,x,\nu) \times f_{n,i,\check{s}}^m(\omega,\tilde{x},\tilde{\nu}). \end{aligned}$$

Notice that, by definition, $E_{F_n}[f_{n,i,s}^m(\omega,x,\nu)] = E_{F_n}[f_{n,i,\check{s}}^m(\omega,\tilde{x},\tilde{\nu})] = 0$ and $E_{F_n}[f_{n,i,s,\check{s}}^{mm}(\omega,(x,\nu),(\tilde{x},\tilde{\nu}))] =$

$h_{2,F_n}((x,\nu),(\tilde{x},\tilde{\nu}))_{(s,\check{s})}$. Based on this argument, the desired result follows from proving that $\forall s,\check{s}=1,\ldots,p$,

$$\sup_{(x,\nu)\in\mathcal{G}}\left\|n^{-1}\sum_{i=1}^{n}f^{m}_{n,i,s}(\omega,x,\nu)-E_{F_n}[f^{m}_{n,i,s}(\omega,x,\nu)]\right\|\xrightarrow{p}0,$$
$$\sup_{(x,\nu),(\tilde{x},\tilde{\nu})\in\mathcal{G}}\left\|n^{-1}\sum_{i=1}^{n}f^{mm}_{n,i,s,\check{s}}(\omega,(x,\nu),(\tilde{x},\tilde{\nu}))-E_{F_n}[f^{mm}_{n,i,s,\check{s}}(\omega,(x,\nu),(\tilde{x},\tilde{\nu}))]\right\|\xrightarrow{p}0.$$

To complete this task we rely on AS13 (Lemma E2), which extends Pollard (1990, Theorem 8.2) to triangular array stochastic processes. This result requires that, for arbitrary $s,\check{s}=1,\ldots,p$, we verify certain conditions on the following triangular array of processes:

(i) $\{\{f^{m}_{n,i,s}(\omega,x,\nu)\ :\ (x,\nu)\in\mathcal{G}\}_{i=1}^{n}\}_{n\geq1}$,

(ii) $\{\{f^{mm}_{n,i,s,\check{s}}(\omega,(x,\nu),(\tilde{x},\tilde{\nu}))\ :\ (x,\nu),(\tilde{x},\tilde{\nu})\in\mathcal{G}\}_{i=1}^{n}\}_{n\geq1}$.

Conditions for (i). By Assumption A.5, $\{M(Z_i,\theta,x,\nu):(x,\nu)\in\mathcal{G}\}_{i=1}^{n}$ is manageable with respect to the envelopes $\{M(Z_i,\theta)\}_{i=1}^{n}$. From this, it follows that $\{M_s(Z_i,\theta,x,\nu):(x,\nu)\in\mathcal{G}\}_{i=1}^{n}$ is manageable with respect to the envelopes $\{M_s(Z_i,\theta)\}_{i=1}^{n}$. By AS13 (Lemma E1), it then follows that $\{f^{m}_{n,i,s}(\omega,x,\nu):(x,\nu)\in\mathcal{G}\}_{i=1}^{n}$ is manageable with respect to envelopes $\{F_{n,i,s}(\omega)\}_{i=1}^{n}$ defined as follows:

$$F_{n,i,s}(\omega)\equiv\sigma_s^{-1}(\theta_n)(M_s(Z_i,\theta_n)+E_{F_n}[M_s(Z_i,\theta_n)]).\tag{S2.3}$$

To complete the argument, it suffices to show that $n^{-1}\sum_{i=1}^{n}E_{F_n}[F^{1+\eta}_{n,i,s}]\leq\check{K}$ for some $\check{K}<\infty$, $\eta>0$, and all $n\in\mathbb{N}$. For this purpose, consider the following derivation for $\eta=1+\delta$ with $\delta>0$ as in Definition A.1:

$$E_{F_n}[F^{2+\delta}_{n,i,s}]=E_{F_n}[(\sigma_s^{-1}(\theta_n)(M_s(Z_i,\theta_n)+E_{F_n}[M_s(Z_i,\theta_n)]))^{2+\delta}]\leq2^{2+\delta}E_{F_n}[|\sigma_s^{-1}(\theta_n)M_s(Z_i,\theta_n)|^{2+\delta}],$$

where the equality holds by Eq. (S2.3), the inequality holds by the convexity of $x^{2+\delta}$. The desired result then follows immediately from $(\theta_n,F_n)\in\bar{\mathcal{F}}_0$, as this implies that $F^{2+\delta}_{n,i,s}$ is i.i.d. and that $E_{F_n}[|\sigma_j^{-1}(\theta_n)M_{n,j}(Z,\theta_n)|^{2+\delta}]<K$ for all $j=1,\ldots,p$ and $n\in\mathbb{N}$.

Conditions for (ii). By our previous verification, $\{f^{m}_{n,i,s}(\omega,(x,\nu))\ :\ (x,\nu)\ \in\ \mathcal{G}\}_{i=1}^{n}$ is manageable with respect to envelopes $\{F_{n,i,s}(\omega)\}_{i=1}^{n}$ with $F_{n,i,s}(\omega)$ as in Eq. (S2.3) for $s=1,\ldots,p$. From this, $f^{mm}_{n,i,s,\check{s}}(\omega,(x,\nu),(\tilde{x},\tilde{\nu}))\equiv f^{m}_{n,i,s}(\omega,x,\nu)f^{m}_{n,i,\check{s}}(\omega,\tilde{x},\tilde{\nu})$, and the arguments in the proof of AS13 (Lemma A1(b)), it then follows that $\{f^{mm}_{n,i,s,\check{s}}(\omega,(x,\nu),(\tilde{x},\tilde{\nu})):(x,\nu),(\tilde{x},\tilde{\nu})\in\mathcal{G}\}_{i=1}^{n}$ is manageable with respect to envelopes $\{F_{n,i,s,\check{s}}(\omega)\}_{i=1}^{n}$ defined by:

$$F_{n,i,s,\check{s}}(\omega)\equiv\sigma_s^{-1}(\theta_n)\sigma_{\check{s}}^{-1}(\theta_n)(M_s(Z_i,\theta_n)+E_{F_n}[M_s(Z_i,\theta_n)])(M_{\check{s}}(Z_i,\theta_n)+E_{F_n}[M_{\check{s}}(Z_i,\theta_n)]).\tag{S2.4}$$

To complete the argument, it suffices to show that $n^{-1}\sum_{i=1}^{n}E_{F_n}[F^{2+\delta/2}_{n,i,s,\check{s}}]\leq\check{K}$ for some $\check{K}<\infty$, $\eta>0$, and all $n\in\mathbb{N}$. For this purpose, consider the following derivation for $\eta=1+\delta/2$ with $\delta>0$ as in Definition A.1:

$$\begin{aligned}E_{F_n}[F^{2+\delta/2}_{n,i,s,\check{s}}]&=E_{F_n}[(\sigma_s^{-1}(\theta_n)\sigma_{\check{s}}^{-1}(\theta_n)(M_s(Z_i,\theta_n)+E_{F_n}[M_s(Z_i,\theta_n)])(M_{\check{s}}(Z_i,\theta_n)+E_{F_n}[M_{\check{s}}(Z_i,\theta_n)]))^{2+\delta/2}]\\&\leq4^{2+\delta}E_{F_n}[|\sigma_s^{-1}(\theta_n)M_s(Z_i,\theta_n)||\sigma_{\check{s}}^{-1}(\theta_n)M_{\check{s}}(Z_i,\theta_n)|^{2+\delta/2}]\\&\leq4^{2+\delta}\{E_{F_n}[|\sigma_s^{-1}(\theta_n)M_s(Z_i,\theta_n)|^{2+\delta}]\}^{(2+\delta/2)/(2+\delta)}\{E_{F_n}[|\sigma_{\check{s}}^{-1}(\theta_n)M_{\check{s}}(Z_i,\theta_n)|^{2+\delta}]\}^{(2+\delta/2)/(2+\delta)},\end{aligned}$$

where the first line holds by Eq. (S2.4), the second line holds by the convexity of $x^{2+\delta/2}$, and the third line follows from Hölder's inequality. The desired result then follows immediately from $(\theta_n,F_n)\in\bar{\mathcal{F}}_0$, as this implies that $F_{n,i,s,\check{s}}^{2+\delta/2}$ is i.i.d. and that $E_{F_n}[|\sigma_j^{-1}(\theta_n)M_{n,j}(Z,\theta_n)|^{2+\delta}] < K$ for all $j = 1, \ldots, p$ and $n \in \mathbb{N}$. □

References

Andrews, D. W. K. and X. Shi (2013): "Inference Based on Conditional Moment Inequalities," *Econometrica*, 81, 609–666.

Aucejo, E. M., F. A. Bugni, and V. J. Hotz (2015): "Identification and Inference on Regressions with Missing Covariate Data," Mimeo: London School of Economics and Duke University.

Pollard, D. (1990): *Empirical Processes: Theory and Applications*, NSF-CBMS Regional Conference Series in Probability and Statistics, Volume 2.