NORTHWESTERN UNIVERSITY

Essays on Inference in Partially Identified Models

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Economics

By

Federico Andrés Bugni

EVANSTON, ILLINOIS

December 2008

ABSTRACT

Essays on Inference in Partially Identified Models

Federico Andrés Bugni

This dissertation is composed of three chapters, each contributing to different aspects of the literature of partially identified econometric models.

In the first chapter, I introduce a bootstrap procedure to perform inference in the class of partially identified econometric models defined by finitely many moment equalities and inequalities. I provide two different versions: one to cover each element of the identified set with a prespecified probability and the other to cover the identified set itself with a prespecified probability. I compare my bootstrap procedure, a competing asymptotic approximation and the subsampling procedure proposed by Chernozhukov, Hong and Tamer [**23**] in terms of the rate at which they achieve the desired coverage level. Under certain conditions, I show that inference based on my bootstrap and asymptotic approximation should eventually be more precise than inference based on subsampling. A Monte Carlo study confirms this finding in a small sample simulation.

In the second chapter, I adapt the specification test for functional data developed by Bugni, Hall, Horowitz and Neumann [**19**] to the presence of missing observations. By

using a worst case scenario approach, my method is able to extract all the information available in the data while being agnostic about the nature of the missing observations. Under the null hypothesis, my specification test will reject the null hypothesis with a probability that, in the limit, does not exceed the significance level of the test. Moreover, the power of the test converges to one whenever the distribution of the non-missing data conveys that the null hypothesis is false. The procedure is illustrated by using it to test the hypothesis that a sample of wage paths was generated by a specific equilibrium job search model.

The third chapter explores the causal relationship between the child labor legislation dictated by several U.S. states between 1880 and 1910 and the spectacular decrease in child occupation rates during this period. Previous literature has studied this connection using differencing techniques in binary choice models, which I show to be inadequate. I develop a model with multiple equilibria to analyze the labor market mechanism by which the legislation affects the household's decision to send their children to work. Based on this model, it is possible to establish if the legislation was effective and if it constituted a benign policy or not, which sheds new light to previous results.

Acknowledgements

I am grateful to my advisors, Joel Horowitz, Rosa Matzkin and Elie Tamer for their constant guidance and support. I have received advice and comments from Ivan Canay, Xiaohong Chen, Joe Ferrie, Enno Mammen, Charles Manski, Adam Rosen and Viktor Subbotin. I am especially thankful to my dear friends, Nenad Kos and Xun Tang, who have always been willing to engage in discussions and help me with many technical difficulties. The Eisner Fellowship and the Dissertation Year Fellowship provided me financial support during the latter part of the graduate program.

Graduate school provided me with many wonderful friends: Arnau Bages, Agustín Casas, Ivan Fau-Rufa, Nacho Franceschelli, Jon Gemus, Michel Janna-Gandur, Manuel Müller-Frank, Claudia Neri, Nenad Kos, Tim Lin, Emiliano Pagnotta, Niels Pedersen and Xun Tang. They have all contributed to make this whole experience enjoyable and unforgettable. My parents, Graciela and Gustavo Bugni, stood behind me every step of the way. Last, but not least, I am grateful to Silvia Glaubach, who has always been there for me. Without her encouragement, support and advice, I would have never survived this journey.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Bootstrap Inference on Partially Identified Models

1.1. Introduction

This paper contributes to the growing literature on inference in partially identified econometric models. A model is said to be *partially identified* or *set identified* when the sampling process and the maintained assumptions restrict the value of the parameter of interest to a set, called the *identified set*, which is smaller than the logical range of the parameter but potentially[1] larger than a single point. Partially identified models arise naturally in economic models when strong and usually unrealistic assumptions are traded by weaker and more credible assumptions. The literature on partially identified models in econometrics has been largely developed and popularized by Manski (See, for example, Manski [**40**] and Manski [**41**]).

The goal of this paper is twofold. The first objective is to introduce a novel bootstrap procedure to construct confidence sets for a wide class of partially identified models. In large samples, our bootstrap procedure achieves exactly the desired coverage probability. The second objective is to compare our bootstrap procedure with competing inferential procedures in terms of the rate of convergence of the error in the coverage probability, that is, in terms of the rate at which they achieve the desired coverage level. To the

---

[1] If the parameter of interest is restricted to a single point, the model is said to be point identified. Since a singleton is a special type of set, point identified models are a special case of partially identified models.

best of our knowledge, this is the first paper performing this type of comparison among competing inferential procedures for partially identified models.

The objective of this paper is to perform hypothesis tests in partially identified models. Based on the duality between hypothesis testing and confidence sets, the hypothesis testing problem can be translated into the construction of confidence sets that cover the object of interest with a minimum prespecified probability. Unlike the point identified literature, there are two possible objects of interest. On the one hand, the object of interest can be the identified set itself. A set $C_n(1-\alpha)$ is a confidence set for the identified set with level $(1-\alpha)$ if and only if the following property is satisfied,

$$\inf_{\mathbf{P}\in\mathcal{F}} \liminf_{n\to\infty} P\left(\Theta_I(\mathbf{P}) \subseteq C_n(1-\alpha)\right) \geq (1-\alpha) \tag{1.1}$$

where $\Theta_I(\mathbf{P})$ denotes the identified set for a certain distribution of observables $\mathbf{P}$ that belongs to a set of possible distributions $\mathcal{F}$. On the other hand, the object of interest can be each of the elements of the identified set. The rationale behind this approach is that if all parameters of the identified set are covered, then the true parameter that generated the observations will also be covered. A set $C_n(1-\alpha)$ is a confidence set for each element of the identified set with $(1-\alpha)$ level if and only if the following property is satisfied,

$$\inf_{\mathbf{P}\in\mathcal{F}} \inf_{\theta\in\Theta_I(\mathbf{P})} \liminf_{n\to\infty} P\left(\theta \in C_n(1-\alpha)\right) \geq (1-\alpha) \tag{1.2}$$

The distinction between these two constructions was pointed out by Imbens and Manski [**38**], who show that the confidence set for the identified set will also be a confidence set for each of its elements. In accordance with this distinction, we provide two versions of our bootstrap procedure: one that covers the identified set with a fixed probability and the other one that covers each element of the identified set with a fixed probability.

A confidence set is said to provide *consistent inference in level* if it satisfies the corresponding coverage requirement (condition (1.1) or (1.2), depending on which is the object of interest) with equality. In other words, a confidence set results in consistent inference in level if, in large samples, it achieves exactly the desired coverage level. This is a desirable property since it implies that the confidence set is not excessively large, which would result in unnecessary loss of statistical power of the underlying hypothesis test. We show that our bootstrap procedure provides consistent inference in level.

Our results on confidence regions for partially identified models build upon the criterion function approach introduced by Chernozhukov, Hong and Tamer [**23**] (henceforth, CHT). In their paper, they implement their inference using a resampling technique called subsampling. In essence, we provide a way to implement the criterion function approach in a wide class of econometric models using an alternative resampling technique, the bootstrap. Our bootstrap procedure differs qualitatively from replacing the subsampling method provided by CHT [**23**] with the bootstrap, that is, we do not merely propose a bootstrap analogue of their subsampling method. In fact, we show that a bootstrap analogue of their subsampling procedure would, in general, fail to be consistent in level. The difference between our bootstrap method and the bootstrap analogue of CHT [**23**]'s subsampling lies in the choice of the bootstrap criterion function, which is the key to our consistency result. Following similar techniques to those used to implement our bootstrap scheme, we also propose an asymptotic approximation to perform inference. In independent research, a similar asymptotic approximation has also been proposed by Soares [**60**], CHT [**23**] and Andrews and Soares [**5**].

There are currently many methods available to implement inference in partially identified econometric models. Given the choice of the criterion function, the researcher can implement inference using our bootstrap, our asymptotic approximation or subsampling. Since all these methods provide consistent inference in level (that is, they all manage to achieve the desired goal, asymptotically), an important basis of comparison is the rate at which the error in the coverage probability vanishes (that is, the rate at which this goal is achieved). If two methods have errors in the coverage probability that converge to zero at different rates, then the one that converges faster will eventually be more accurate than the one that converges slower. We show that, under certain conditions, our bootstrap and our asymptotic approximation both have error in the coverage probability that converges to zero at the same rate, which is a faster rate than the one obtained by using subsampling. Hence, under these conditions, our results imply that inference based on our bootstrap and our asymptotic approximation should eventually be more precise than inference based on subsampling. Our Monte Carlo simulation shows that this difference in accuracy can be important in applications of moderate sample sizes.

The rest of the paper is organized as follows. Section 1.2 reviews the literature of inference in partially identified models, and section 1.3 provides an introduction to the criterion function approach. As we have already anticipated, we can construct confidence sets for two different objects of interest and the remainder of the paper deals with these two separately. In section 1.4, we consider the construction of confidence sets for the identified set. Section 1.4.1 introduces our assumptions and provides examples of econometric models where they are satisfied. Our bootstrap procedure is introduced in section 1.4.2, where we also demonstrate its consistency in level and we analyze its error in coverage probability. In section 1.4.3, we consider the two competing inferential procedures: subsampling and asymptotic approximation, for which we also show consistency in level and analyze the error in coverage probability. These inferential methods are compared using a Monte Carlo simulation in section 1.4.4. Section 1.5 repeats this analysis for the construction of confidence sets for each element of the identified set. The structure of section 1.5 is similar to the one of section 1.4. We introduce the setup, we present the bootstrap procedure and show its properties, we perform the comparison with alternative inferential schemes and we provide a Monte Carlo simulation. Section 1.6 concludes the paper and provides directions for further research. The appendix collects all the proofs of the paper.

1.2. Literature review

There is now a growing literature on inference on partially identified (or set identified) parameters. As we mentioned in the introduction, the objective of this literature is to construct confidence sets that cover the object of interest with a prespecified probability.

The most natural way of constructing confidence sets is to expand the boundaries of an estimator of the identified set. For identified sets whose boundary is a functional of the observed data, this approach can be easy to implement. For examples of this approach, see Horowitz and Manski [**37**] or Imbens and Manski [**38**] (when sets are intervals) and Rosen

[**57**] (when sets are polyhedrons). Beresteanu and Molinari [**11**] extend this approach to more general settings by using developments on set valued random variables.

An alternative approach to constructing confidence sets for partially identified models is the criterion function approach, introduced by CHT [**23**]. The first step of this approach is to define a function, called *the criterion function*, that is minimized exclusively at the identified set. Confidence sets are generated by inverting the sample analogue of this function. This procedure is attractive because it can automatically handle problems that would be very hard to deal with a more direct approach. In order to implement their inference, CHT [**23**] propose a subsampling approximation and an asymptotic approximation. In relation to this approach, Manski and Tamer [**42**] provided a consistent estimator of the identified set based on criterion functions.

According to the literature on inference in point identified models, the rates of convergence of subsampling procedures are likely to be slow, relative (for example) to the asymptotic approximation or the bootstrap approximation. See, for example, Horowitz [**36**], Bickel, Götze and van Zwet [**14**], Politis and Romano [**53**] and Politis, Romano and Wolf [**54**]. Under certain conditions, we show that these results extend to a wide class of partially identified models. Moreover, we show that these rates of convergence determine the rate at which the error in the coverage probability vanishes. As a consequence, inferential methods with a faster rate of convergence are eventually more precise than inferential methods with a slower rate of convergence.

Andrews, Berry and Jia [**3**] consider games with discrete strategies, where the parameters are restricted by necessary conditions imposed by the Nash equilibrium. In these games, the parameters are usually partially identified due to the existence of multiple equilibria. They provide a method to perform inference that differs significantly from ours. Pakes, Porter, Ho and Ishii [**52**] consider estimation and inference for points in the boundaries of the identified set of partially identified models. Their inferential method has the advantage that it is simple to implement and does not depend on the unknown number of binding moment inequalities. In general, their method results in confidence sets whose asymptotic coverage may exceed the desired coverage, that is, *conservative inference*.

Romano and Shaikh ([**55**] and [**56**]) consider the general problem of constructing coverage regions using a subsampling stepdown control procedure that is comparable to CHT [**23**]'s subsampling construction. They formally show that the subsampling stepdown control procedure cannot be replaced by a bootstrap stepdown control procedure. Rosen [**58**] studies the problem of inference in partially identified models defined by one-sided moment inequalities, similar to the one studied in this paper. The limiting distribution of his test statistic depends on the number of inequalities that are binding at the current parameter value which, of course, is unknown. In order to overcome this difficulty, he replaces the unknown number of binding moment conditions by a known lower bound. The resulting test statistic is asymptotically pivotal and, thus, straightforward to implement but admittedly results in conservative inference.

Blundell, Gosling, Ichimura and Meghir [**16**] study the wage distribution in the labor force taking into account the selection problem generated by unemployment. Since they only observe wages for people who work, the distribution of wages in the labor force is partially identified. They propose the bootstrap as a method for inference but do not formally investigate its asymptotic validity or analyze its properties. Galichon and Henry

([**28**] and [**29**]) consider inference on a general class of partially identified models. To implement their inference, they develop a new bootstrap method, called dilation bootstrap, which differs considerably from the traditional bootstrap. Their inferential procedure presents significant computational advantages with respect to alternative methods, but, to the best of our knowledge, the proof of the consistency in level of the dilation bootstrap is under progress. In research independent to ours, Soares [**60**] constructs coverage sets for each parameter in the identified set for the type of econometric models considered in this paper. He develops an asymptotic approximation that is similar to the one proposed by CHT [**23**] and by this paper.

Andrews and Soares [**5**] study the power of the hypothesis tests in partially identified models defined by moment inequalities. They introduce an inference method called generalized moment selection (GMS), in which information about the slackness of the sample moment conditions is used to infer which population moment conditions are binding. Our bootstrap and our asymptotic approximation can be considered special cases of their GMS procedure. Their results indicate that GMS tests are more powerful than alternative competing inferential methods, such as subsampling. In this sense, their results provide an interesting complement to our work: they recommend using GMS to test hypothesis in partially identified models based on the power of the test and we do so based on the accuracy of the approximation.

Canay [**21**] studies the problem of inference on the parameters that compose the identified set and shows that a criterion function based on empirical likelihood has certain optimality properties. Stoye [**61**] revisits the analysis of Imbens and Manski [**38**] and reveals the importance of a superefficiency assumption.

For partially identified models whose identified set is an interval, Imbens and Manski [**38**] discuss the issue of uniformity of the inference. A confidence set provides robust inference if it provides consistent inference in level not only for a fixed probability distribution, but uniformly for a class of probability distributions[2]. They illustrate the problem by showing how confidence sets for partially identified models based on pointwise asymptotics provide very misleading results in the limiting case when the upper and lower bounds of the interval coincide and the parameter of interest becomes point identified. Romano and Shaikh ([**55**] and [**56**]) provide conditions under which their subsampling construction achieves uniform coverage. Soares [**60**], Andrews and Guggenberger [**4**] and Andrews and Soares [**5**] show that these concerns extend to the general class of partially identified models defined by moment inequalities. The reason is that, in this class of models, test statistics have pointwise asymptotic distributions that are discontinuous in the true distribution generating the data but such discontinuity is not present in the finite sample distribution. In research developed independently to ours, Soares [**60**] and Andrews and Soares [**5**] show how to construct confidence sets that provide uniform coverage using asymptotic approximations based on similar techniques to the ones proposed in this paper.

1.3. The criterion function approach

Suppose that the economic model is known up to a parameter θ, that belongs to a parameter space $\Theta \subseteq \mathbb{R}^{\eta}$. According to the model, the observations are sampled from a distribution $\mathbf{P}(\theta_0)$, where θ_0 denotes the true value of the parameter. Since the model

---

[2] Formally, this requires modifying the requirement of equations (1.1) or (1.2) so that $\inf_{\mathbf{P}\in\mathcal{F}}$ is computed *before* taking $\liminf_{n\to\infty}$. See, e.g., Andrews and Guggenberger [**4**] and Andrews and Soares [**5**].

is partially identified, the sampling distribution does not completely determine θ_0, but, rather, restricts it to a certain set, denoted $\Theta_I\left(\mathbf{P}\left(\theta_0\right)\right)$ or, more succinctly[3], Θ_I.

The objective of inference is to construct confidence sets that cover the object of interest with a prespecified probability. As mentioned in the introduction, the object of interest can be the identified set itself or can be each element of the identified set. The *criterion function approach* provides a very general procedure for both coverage objectives in a relatively simple way.

In the criterion function approach we define a non-negative function of the parameter space, denoted by Q, that equals zero if and only if θ belongs to the identified set. This function is referred to as *criterion function* since it provides a criterion that characterizes the identified set. We denote its sample analogue by Q_n. The basic idea of the criterion function approach is to construct a $(1-\alpha)$ confidence region of the object of interest, denoted $C_n\left(1-\alpha\right)$, using a lower level set of the sample analog of the criterion function, namely,

$$C_n\left(1-\alpha\right)=\left\{\theta\in\Theta : a_nQ_n\left(\theta\right)\leq c_n\left(\theta\right)\right\} \tag{1.1}$$

where $\left\{a_n\right\}_{n=1}^{+\infty}$ is a sequence of constants that makes the (asymptotic) distribution of $a_nQ_n\left(\theta\right)$ non-degenerate. $C_n\left(1-\alpha\right)$ will be a confidence set for the identified set or for each element of the identified set depending on how $c_n\left(\theta\right)$ is defined. On the one hand, $C_n\left(1-\alpha\right)$ is a confidence region for the identified set with level $(1-\alpha)$ (condition (1.1)) if, for every θ in the parameter space, $c_n\left(\theta\right)$ is equal to c_n, the $(1-\alpha)$ quantile of the distribution of $\sup_{\theta\in\Theta_I}a_nQ_n\left(\theta\right)$. On the other hand, $C_n\left(1-\alpha\right)$ is a confidence region for

---

[3] The dependence on $\mathbf{P}\left(\theta_0\right)$ can be dropped without risk of confusion because we consider inference for a fixed probability distribution.

each element of the identified set with level $(1-\alpha)$ (condition (1.2)) if for every θ in the parameter space, $c_n(\theta)$ is set to be the $(1-\alpha)$ quantile of the distribution of $a_nQ_n(\theta)$.

Once we have chosen the criterion function (and with it, its sample analogue), all that remains to construct the confidence sets is a way to approximate quantiles of the distribution of either $\sup_{\theta\in\Theta_I} a_nQ_n(\theta)$ or $a_nQ_n(\theta)$ for every θ in the parameter space. This approximation problem is non-standard precisely because the econometric model is partially identified.

For a general class of models, CHT [**23**] show how to construct these confidence sets using subsampling. For the particular class of models considered in this paper, several papers[4] show how to construct these confidence sets by simulation from an estimate of the asymptotic distribution. One of the contributions of this paper is to show how the bootstrap can be used to perform this construction.

1.4. Confidence sets for the identified set

We now consider the construction of confidence sets that cover the identified set with a minimum prespecified probability. Formally, our objective is to construct a random set that satisfies equation (1.1). The construction of confidence sets for the identified set is harder than the construction of confidence sets for each element of the identified set. Once the first problem has been solved and analyzed, the second problem can be solved and analyzed with similar techniques. Therefore, we devote the main body of the paper to the construction and analysis of the first problem and leave the second one for the final section.

---

[4] The asymptotic approximation procedure has been independently proposed by Soares [**60**], CHT [**23**], Andrews and Soares [**5**] and earlier versions of this paper.

The confidence sets considered in this section can be related to a hypothesis testing problem. Romano and Shaikh [**56**] show that a confidence region for the identified set can be interpreted as a test for the family of null hypotheses $H_\theta : \theta \in \Theta_I$ indexed by $\theta \in \Theta$ while controlling for the familywise error rate, that is, the probability of even one false rejection. We provide an analogous interpretation of this statement. Suppose that for a certain set $\mathcal{S}$ $(\mathcal{S} \subseteq \Theta)$, we want to test the null hypothesis $H_0 : \mathcal{S} \subseteq \Theta_I$ versus the alternative hypothesis $H_1 : \mathcal{S} \nsubseteq \Theta_I$, while keeping the probability of a false rejection to be less or equal than $(1-\alpha)$. A confidence set that satisfies condition (1.1) contains all sets $\mathcal{S} \subseteq \Theta$ that will fail to reject the aforementioned null hypothesis.

1.4.1. Setup

In this section, we introduce the assumptions that define our econometric model. We consider two separate set of assumptions. The first set of assumptions will be more general and will constitute what we call *the general model.* The second set of assumptions will be a particular subset of the first one and will give rise to what we refer to as *the conditionally separable model.* The reason to consider these two setups separately is that consistency in level can be obtained under the assumptions of the general model but results regarding rates of convergence require the stronger framework imposed by the conditionally separable model.

After introducing and explaining these assumptions, we provide examples of relevant economic models where these are satisfied.

1.4.1.1. General model. The following assumptions constitute our general model in the independent and identically distributed (i.i.d.) setting.

(A1) For the probability space $(\Omega, \mathcal{B}, \mathbf{P})$, let $Z : \Omega \to \mathcal{Z}$ be a random vector. We observe an i.i.d. sample $\mathcal{X}_n \equiv \{Z_i\}_{i=1}^n$.

(A2) The parameter space, denoted by Θ, is a compact and convex subset of a finite dimensional Euclidean space $\mathbb{R}^\eta$ ($\eta < +\infty$).

(A3) The identified set, denoted by Θ_I, is given by,

$$\Theta_I = \left\{ \theta \in \Theta : \left\{ \mathbb{E}\left(m\left(Z, \theta \right) \right) \leq \vec{0} \right\} \right\}$$

where $m(z, \theta) : \mathcal{Z} \times \Theta \to \mathbb{R}^J$ is a (jointly) measurable function and $\mathbb{E}(m(Z, \theta)) : \Theta \to \mathbb{R}^J$ is a lower semi-continuous function. Moreover, Θ_I is a proper subset of Θ.

(A4) For every $\theta \in \Theta$ and every $j = 1, 2, ..., J$, the variance of $m_j(Z, \theta)$ is positive and finite. For every $z \in \mathcal{Z}$, $\{(m(z, \theta) - \mathbb{E}(m(Z, \theta))) : \theta \in \Theta\}$ is a separable subset of $l_J^\infty(\Theta)$, the space of bounded functions that map Θ into $\mathbb{R}^J$. The empirical process associated to the random variable $m(Z, \theta)$, given by,

$$v_n(m_\theta) = n^{-1/2} \sum_{i=1}^n \left(m(Z_i, \theta) - \mathbb{E}(m(Z, \theta)) \right)$$

is stochastically equicontinuous, i.e., for any $\varepsilon > 0$,

$$\lim_{\eta \downarrow 0} \limsup_{n \to \infty} P^* \left(\sup_{\theta \in \Theta} \sup_{\{\theta' : \|\theta' - \theta\| \leq \eta\}} \| v_n(m_\theta) - v_n(m_{\theta'}) \| > \varepsilon \right) = 0$$

where $\|\cdot\|$ denotes Euclidean distance and P^* denotes the outer measure[5] with respect to P.

---

[5] Let $(\Omega, \mathcal{B}, \mathbf{P})$ be a probability space. For any arbitrary subset of Ω, denoted A, its outer measure is defined by $P^*(A) = \inf_{S \subseteq \mathcal{B}} \{P(S) : A \subseteq S\}$.

We briefly comment on some of the assumptions. Assumption (A1) requires that the sample is i.i.d.. The result of consistency of the bootstrap procedure proposed in this paper is based on laws of large numbers, central limit theorems and laws of iterated logarithm. Consistency of our bootstrap procedure can be generalized to non i.i.d. settings, provided that these results hold and, of course, that the resampling method is adequately adjusted.

Assumption (A3) defines the identified set as the intersection of finitely many weak moment inequalities. These weak inequalities are upper bounds of the expectation of a random function. Of course, we can trivially accommodate lower bounds by changing signs and we can accommodate equality restrictions by combining upper and lower bounds. Notice that assumption (A3) allows the identified set to be empty. A valuable feature of our inference procedure is that if the identified set is empty, then, eventually, our confidence set will be equal to the smallest possible confidence set[6], almost surely.

The present setup allows for econometric models defined by *conditional* moment conditions as long as the covariates have finite support[7]. To see why, suppose that the conditioning covariate X has finite support given by S_X and let the identified set be given by,

$$\Theta_I = \left\{ \theta \in \Theta : \bigcap_{x \in S_X} \left\{ \mathbb{E}\left(M\left(Y,\theta\right)|X=x\right) \leq 0 \right\} \right\}$$

where $M\left(y,\theta\right) : \mathcal{Y} \times \Theta \rightarrow \mathbb{R}^J$ is a jointly measurable function and for every $x \in S_X$, $\mathbb{E}\left(M\left(Y,\theta\right)|X=x\right) : \Theta \rightarrow \mathbb{R}^J$ is lower semi-continuous. This identified set can

---

[6] With the criterion function approach, the smallest possible confidence set is equal to $\hat{\Theta}_I\left(0\right) = \left\{\theta \in \Theta : a_n Q_n\left(\theta\right) \leq 0\right\}$.

[7] The methods proposed in this paper cannot directly handle covariates with infinite support. If this is the case, one can still use our techniques by partitioning the support of the continuous covariate into finitely many cells. In this process, some information will be lost, and so our method will result in conservative inference.

be equivalently reformulated in the form required by assumption (A3). By defining $Z = (Y, X) : \Omega \to \mathcal{Y} \times S_X$ and $m(Z, \theta) = M(Y, \theta) 1[X = x]$ we get,

$$\Theta_I = \left\{ \theta \in \Theta : \bigcap_{x \in S_X} \{\mathbb{E}(M(Y, \theta) 1[X = x]) \leq 0\} \right\}$$

which has the structure required by assumption (A3).

Assumption (A4) introduces the regularity conditions introduced to apply the law of iterated logarithm and the (Donsker) central limit theorem in our proofs.

1.4.1.2. Conditionally separable model. The following assumptions constitute our conditionally separable model in the i.i.d. setting.

(B1) For the probability space $(\Omega, \mathcal{B}, \mathbf{P})$, let $(X, Y) : \Omega \to \{S_X \times \mathbb{R}^J\}$ be a random vector, where the support of X, denoted S_X, is composed of K values (finite support). We observe an i.i.d. sample $\mathcal{X}_n \equiv \{X_i, Y_i\}_{i=1}^n$.

(B2) The parameter space, denoted by Θ, is a compact and convex subset of a finite dimensional Euclidean space $\mathbb{R}^\eta$ $(\eta < +\infty)$.

(B3) The identified set Θ_I is given by,

$$\Theta_I = \left\{ \theta \in \Theta : \bigcap_{k=1}^{K} \{\mathbb{E}(Y - M(\theta, x_k) | x_k) \leq 0\} \right\}$$

where , for each $x \in S_X$, $M(\theta, x) : \Theta \to \mathbb{R}^J$ is continuous. Moreover, Θ_I is a proper subset of Θ.

(B4) For every $x \in S_X$ and $j = 1, 2, ..., K$, the variance of $\{Y_j | X = x\}$ is positive and finite.

(B5) For every $x \in S_X$, $\{Y | X = x\}$ has finite fourth absolute moments.

This setup is a particular case of the general model that strengthens assumptions (A3) and (A4). In the presence of covariates, assumption (A3) of the general model defines the identified set as the intersection of moment inequalities of the form $\mathbb{E}\left(m\left(Z,\theta\right)|X=x\right)\leq 0$. The assumption (B3) strengthens this by requiring that the conditional expectation $\mathbb{E}\left(m\left(Z,\theta\right)|X=x\right)$ can be separated into the expectation of a random variable that does not involve θ, given by $\mathbb{E}\left(Y|X=x\right)$, and a conditionally non-stochastic term that involves θ, given by $M\left(\theta,x\right)$. Subtracting the (conditional) mean to the stochastic process $\left\{Y-M\left(\theta,X\right)|X=x_k\right\}$ results in a random vector (i.e., no dependency on the parameter θ), and so the separability and stochastic equicontinuity conditions required by (A4) are trivially satisfied.

1.4.1.3. Examples. We consider three examples of economic models that satisfy the assumptions of our frameworks.

Example 1. ***Inference on the mean with missing or censored data***

This example was first considered by Manski [**40**]*. Suppose that we are interested in performing inference on the mean of a random variable, denoted by* $\{Z|X\}$*, and assume that the support of* X *is composed of* K *values,* $S_X=\{x_k\}_{k=1}^K$. *Our parameter of interest is the following* K *dimensional vector:* $\theta=\left\{\mathbb{E}\left(Z|X=x_k\right)\right\}_{k=1}^K$.

In our random sample of observations of $\{Z|X\}$, *certain observations are missing (or are censored) and we are unwilling to make assumptions about the distribution of these observations. Let* U *denote the binary variable that takes value one if the observation is unobserved and zero otherwise. By the law of iterated expectations,*

$$\mathbb{E}\left(Z|X\right)=\mathbb{E}\left(Z|U=0,X\right)P\left(U=0|X\right)+\mathbb{E}\left(Z|U=1,X\right)\left(1-P\left(U=0|X\right)\right)$$

The observed sample will identify the mean conditional on being observed and the probability of observation, but will be silent about the mean conditional on being unobserved. Nevertheless, we may be able to establish logical lower and upper bounds for $\{Z|X\}$, *denoted by* $Z_L(X)$ *and* $Z_H(X)$, *respectively*[8]*. The identified set for our parameter of interest is,*

$$\Theta_I = \left\{ \theta \in \Theta : \bigcap_{k=1}^{K} \left\{ \begin{array}{c} \mathbb{E}\left(Z(1-U) + Z_L(X)U|X = x_k\right) - \theta_k \leq 0 \\ \theta_k - \mathbb{E}\left(Z(1-U) + Z_H(X)U|X = x_k\right) \leq 0 \end{array} \right\} \right\}$$

Under random sampling and regularity conditions, this model satisfies the assumptions of the conditionally separable framework.

Example 2. ***Inference on parametric models with missing or censored data***

Suppose that our model predicts that for a known function f,

$$\mathbb{E}\left(Z - f(X,\beta)|W\right) = 0$$

where Z *is the explained variable,* X *is the explanatory variable,* β *is the parameter of interest and* W *is an exogenous variable. Typical examples of this setup are linear index models, such as the linear model, the probit model or the logit model.*

Suppose that certain observations of the explained variable are missing (or censored). Let U *denote the binary variable that takes value one if the observation is unobserved and zero otherwise. By the law of iterated expectations,*

$$\mathbb{E}\left(Z - f(X,\beta)|W\right) = \left\{ \begin{array}{c} \mathbb{E}\left(Z - f(X,\beta)|W, U=0\right) P(U=0|W) + \\ +\mathbb{E}\left(Z - f(X,\beta)|W, U=1\right) P(U=1|W) \end{array} \right\}$$

---

[8] When the event has no logical lower (repectively, upper) bound, then $Z_L(X) = -\infty$ (repectively, $Z_H(X) = +\infty$).

Suppose that $\{Z|W\}$ *has logical lower and upper bounds, given by* $Z_L(W)$ *and* $Z_H(W)$, *respectively. Also assume that the support of* W *is given by finitely many values:* $S_W = \{w_k\}_{k=1}^{K}$. *Under this conditions, the identified set for the parameter of interest is given by,*

$$\Theta_I = \left\{ \beta \in \Theta : \bigcap_{k=1}^{K} \left\{ \begin{array}{c} -\mathbb{E}\left(Z(1-U) + Z_H(W)U - f(X,\beta)|W = w_k\right) \leq 0 \\ \mathbb{E}\left(Z(1-U) + Z_L(W)U - f(X,\beta)|W = w_k\right) \leq 0 \end{array} \right\} \right\}$$

Under random sampling and regularity conditions, this model satisfies the assumptions of the general framework. Moreover, it also satisfies the assumptions of the conditionally separable model when the explanatory variable X *is exogenous.*

Example 3. ***Multiplicity of equilibria in games***

Consider the two player static entry game with complete information considered by Tamer [**62**]. *In this model, two players* $(i = 1, 2)$ *have to simultaneously decide whether to enter a market* $(y_i = 1)$ *or not* $(y_i = 0)$.

The Nash Equilibrium strategy for player $i = 1, 2$ *is given by* $y_i = 1[y_i^* \geq 0]$, *where* y_i^* *denotes the profits of entering the market for player* i. *These profits are assumed to be given by* $y_i^* = x_i\beta_i + y_{-i}\Delta_i + u_i$, *where* $(x_1, x_2) \in \mathbb{R}^d$ *represents the vector of exogenous variables,* (u_1, u_2) *is a random vector of latent variables with conditional density* $f_u(\cdot|\Omega)$, *and* $\theta = (\beta_1, \beta_2, \Delta_1, \Delta_2, \Omega)$ *is the vector of parameters of interest.*

Given the structure of the model and under reasonable assumptions, Tamer [**62**] *shows that the model has multiple of equilibria. Without imposing any equilibrium selection* *assumptions, an implication of the model is that the identified set is given by,*

$$\Theta_I = \left\{ \theta \in \Theta : \bigcap_{x \in S_X} \left\{ \begin{array}{c} P\left((0,0)\,|x\right) = \Pi_1\left(x,\theta\right); \\ P\left((1,1)\,|x\right) = \Pi_2\left(x,\theta\right); \\ \Pi_3\left(x,\theta\right) \leq P\left((0,1)\,|x\right) \leq \Pi_4\left(x,\theta\right) \end{array} \right\} \right\}$$

where $\Pi_1\left(x,\theta\right), \Pi_2\left(x,\theta\right), \Pi_3\left(x,\theta\right)$ *and* $\Pi_4\left(x,\theta\right)$ *are known functions that are continuous on* θ*.*

Provided that the covariates have finite support, random sampling and regularity conditions, this model satisfies the assumptions of the conditionally separable framework. In Ciliberto and Tamer [**24**]*, this model is generalized to more than two players and applied to the airline industry. This generalization also satisfies all the assumptions of our conditionally separable framework.*

1.4.2. Bootstrap procedure

In this section, we introduce our bootstrap procedure to construct confidence regions for the identified set. As discussed in section 1.3, such confidence region can be constructed by approximating the $(1-\alpha)$ quantile of the distribution of,

$$\Gamma_n \equiv \sup_{\theta \in \Theta_I} a_n Q_n\left(\theta\right)$$

where Q_n is the sample analogue criterion function and $\{a_n\}_{n=1}^{+\infty}$ is a sequence of constants that makes the (asymptotic) distribution of $a_n Q_n\left(\theta\right)$ non-degenerate.

In order to implement any inferential procedure based on the criterion function approach, we need to complete certain steps. First, we need to define the criterion function for our problem. Second, we need to generate an estimator of the identified set. This estimator is not our final goal, but an intermediate step in our inference problem. Third, we need to define the resampling procedure that implements our inference.

1.4.2.1. Criterion function. By definition, a function $Q : \Theta \to \mathbb{R}$ is a valid criterion function if it is non-negative and takes value zero if and only if it is evaluated at a parameter in the identified set. The following lemma characterizes all possible criterion functions for the type of models considered in this paper.

Lemma 4. *Under assumption (A3), the function* $Q : \Theta \to \mathbb{R}$ *is a valid criterion function if and only if it is of the form,*

$$Q(\theta) = G_{\mathbf{P}}\left(\left\{[\mathbb{E}(m_j(Z,\theta))]_+\right\}_{i=1}^{J}\right)$$

where $[x]_+ = x1[x \geq 0]$ *and* $G_{\mathbf{P}} : \mathbb{R}_+^J \to \mathbb{R}$ *is a non-negative function such that* $G_{\mathbf{P}}(y) = 0$ *if and only if* $y = \vec{0}$.

The lemma implies that there are many possible criterion functions. The notation $G_{\mathbf{P}}$ reveals that, in principle, the criterion function could depend on the probability distribution $\mathbf{P}$. Also, notice that the criterion function need not be continuous. As we will soon show, if we assume certain properties about this function we will obtain desirable asymptotic results, such as consistency in level and rates of convergence. With this objective in mind, we consider the following assumption.

(CF) The population criterion function is given by one of the following two functions,

$$
\begin{aligned}
Q(\theta) &= \sum_{j=1}^{J} w_j \left[\mathbb{E}\left(m_j(Z,\theta)\right)\right]_+ \\
Q(\theta) &= \max\left\{w_j \left[\mathbb{E}\left(m_j(Z,\theta)\right)\right]_+\right\}_{j=1}^{J}
\end{aligned}
$$

where $\{w_j\}_{j=1}^{J}$ are (arbitrary) positive constants.

Throughout this paper, we will focus on criterion functions that satisfy assumption (CF), but some of our results will extend to more general criterion functions. In particular, we will sometimes refer to the following generalization of assumption (CF).

(CF') The population criterion function is given by $Q(\theta) = G\left(\left\{\left[\mathbb{E}\left(m_j(Z,\theta)\right)\right]_+\right\}_{j=1}^{J}\right)$, where $G : \mathbb{R}_+^J \to \mathbb{R}$ is a non-negative function that does not depend on $\mathbf{P}$, is strictly increasing in every coordinate, weakly convex, continuous, homogeneous of degree β and satisfies $G(y) = 0$ if and only if $y = 0$.

Clearly, assumption (CF) is a particular case of assumption (CF'). In the general model, consistency of the bootstrap is shown under assumption (CF) whereas in the conditionally separable model, consistency of the bootstrap is shown under assumption (CF'). Rates of convergence and error in the coverage probability in the conditionally separable model will differ depending on which assumption we use. In particular, the rates of convergence we can show under assumption (CF) are slightly better than those we can show under assumption (CF')[9]. Since the criterion function is a choice of the

---

[9] Under the conditional separable model and under assumption (CF), we will show that the bootstrap approximation has error in the coverage probability of order $n^{-1/2}$. In the same model but under assumption (CF'), the same techniques will show that the bootstrap procedure has error in the coverage probability of order $n^{-1/2}\ln n$ (corollary 44 in the appendix).

econometrician and since assumption (CF) allows us to prove better rates of convergence, we will restrict attention to it in the main text.

In the literature, alternative criterion functions have been considered. For example, CHT [**23**] and Andrews and Soares [**5**] consider dividing each expectation by its standard deviation, that is,

$$Q(\theta) = G\left(\left\{\left[\frac{\mathbb{E}(m_j(Z,\theta))}{\sigma(m_j(Z,\theta))}\right]_+\right\}_{j=1}^J\right)$$

This ensures that the criterion function is not affected by rescaling of the moment inequalities. We can also adopt this rescaling in our bootstrap procedure without affecting the consistency of the approximation or any of the rates of convergence results[10]. We prefer not to do so in the main text to keep the notation simple.

1.4.2.2. Estimation of the identified set. By definition, the identified set is the subset of the parameter space that satisfies $Q(\theta) = 0$. Therefore, the *analogy principle* suggests defining the estimate of the identified set as the collection of parameters that satisfy $Q_n(\theta) = 0$. In the context of sets defined by moment inequalities, this set estimate would be given by,

$$\hat{\Theta}_I^{AP} = \left\{\theta \in \Theta : \{\mathbb{E}_n(m_j(Z,\theta)) \leq 0\}_{j=1}^J\right\}$$

where for every $j = 1, 2, ..., J$, $\mathbb{E}_n(m_j(Z,\theta))$ denotes $n^{-1}\sum_{i=1}^n m_j(Z_i,\theta)$. This set estimate will be called *the analogy principle estimator.* Notice that it is possible that this estimator is empty.

The estimator of the identified set is an ingredient in the construction of confidence sets for the identified set. An estimator of the identified set is adequate for the purpose of inference if it allows us to construct confidence sets that are consistent in level. In settings

---

[10] This result is simple to show and it is omitted from the paper for the sake of brevity.

of practical relevance, the analogy principle estimator is not an adequate estimator for the purpose of inference.

This problem has been considered by CHT [**23**]. They propose an estimator of the identified based on two modifications of the sample criterion function. First, they redefine the sample criterion function so its minimum value is zero[11]. Second, in the spirit of Manski and Tamer [**42**], they estimate the identified set with the set of parameters whose value of the modified criterion function is less than an amount that converges to zero at a suitable rate. In order to perform inference in the class of partially identified models considered in this paper, it is sufficient to adopt only the second of these two modifications. Evading the first modification is computationally valuable, since we avoid solving additional optimization problems to approximate the distribution of the statistic.

Our estimate of the identified set is constructed as follows. Let $\{\tau_n\}_{n=1}^{+\infty}$ be a positive sequence such that $\tau_n/\sqrt{n} = o(1)$ and $\sqrt{\ln\ln n}/\tau_n = o(1)$ (almost surely). For example, $\tau_n = \ln n$ or $\tau_n = \ln\ln n$ satisfy these requirements. Our estimate is given by,

$$\hat{\Theta}_I(\tau_n) = \left\{\theta \in \Theta : \left\{\mathbb{E}_n\left(m_j\left(Z,\theta\right)\right) \leq \tau_n/\sqrt{n}\right\}_{j=1}^{J}\right\}$$

The requirements on the sequence $\{\tau_n\}_{n=1}^{+\infty}$ can be explained intuitively. If the identified set is non-empty, the requirement that τ_n is positive and satisfies $\sqrt{\ln\ln n}/\tau_n = o(1)$ implies that for almost all sample sequences, the set estimator will eventually include the identified set. Intuitively, we are artificially expanding the sample analogue estimate of the identified set in order to eventually include the identified set, almost surely. If

---

[11] To their initial choice of criterion function (denoted $Q_n(\theta)$), they define the modified criterion function to be $\tilde{Q}_n(\theta) = Q_n(\theta) - \inf_{\tilde{\theta}\in\Theta} Q_n\left(\tilde{\theta}\right)$. Thus, by construction, the infimum of the modified criterion function over the parameter space is zero.

we expand the set too much, that is, if τ_n increases too rapidly, we run into the risk of distorting the asymptotic results. This is avoided by requiring that $\tau_n/\sqrt{n} = o(1)$. The following lemma formalizes these findings.

Lemma 5. *Assume (A1)-(A4). Let* $\{\tau_n\}_{n=1}^{+\infty}$ *be a positive sequence such that* $\tau_n/\sqrt{n} = o(1)$ *and* $\sqrt{\ln\ln n}/\tau_n = o(1)$, *almost surely, and define* $\hat{\Theta}_I(\tau_n)$ *as follows,*

$$\hat{\Theta}_I(\tau_n) = \left\{\theta \in \Theta : \left\{\mathbb{E}_n(m_j(Z,\theta)) \leq \tau_n/\sqrt{n}\right\}_{j=1}^{J}\right\}$$

For a sequence of positive numbers $\{\varepsilon_n\}_{n=1}^{+\infty}$ *such that* $\varepsilon_n/\sqrt{n} = o(1)$ *and* $\tau_n/\varepsilon_n = o(1)$, *almost surely, and define* $\Theta_I(\varepsilon_n) = \left\{\theta \in \Theta : \{\mathbb{E}(m_j(Z,\theta)) \leq \varepsilon_n/\sqrt{n}\}_{j=1}^{J}\right\}$. *If the identified set is non-empty then,*

$$P\left(\liminf\left\{\Theta_I \subseteq \hat{\Theta}_I(\tau_n) \subseteq \Theta_I(\varepsilon_n)\right\}\right) = 1.$$

and if the identified set is empty then,

$$P\left(\liminf\left\{\hat{\Theta}_I(\tau_n) = \varnothing\right\}\right) = 1$$

When the identified set is non-empty, our set estimate will eventually be "sandwiched" between two sets, almost surely. These sets are the identified set and a sequence of sets that converges to the identified set. When the identified set is empty, our set estimate will eventually become empty, almost surely.

The restrictions on the sequence $\{\tau_n\}_{n=1}^{+\infty}$ provide little guidance on how to implement the estimator (and the inference based on it) in a finite sample setting. We will comment on this important practical question in the next subsection.

1.4.2.3. The procedure. We now introduce our bootstrap procedure to construct confidence sets for the identified set with a prespecified probability. We will actually propose two different procedures: one to be used if the model satisfies the assumptions of the general model and one to be used exclusively if the model satisfies the assumptions of the conditionally separable model.

Bootstrap procedure for the general model. The following bootstrap method is intended for the general model, and so, in particular, it can also be applied to the conditionally separable model. For this procedure, our main result will be the consistency in level.

(1) Choose a positive sequence $\{\tau_n\}_{n=1}^{+\infty}$ such that $\tau_n/\sqrt{n} = o(1)$ and $\sqrt{\ln\ln n}/\tau_n = o(1)$, almost surely,

(2) Estimate of the identified set with,

$$\hat{\Theta}_I(\tau_n) = \left\{\theta \in \Theta : \left\{\mathbb{E}_n\left(m_j(Z,\theta)\right) \leq \tau_n/\sqrt{n}\right\}_{j=1}^{J}\right\}$$

(3) Repeat the following for $s = 1, 2, ..., S$. Construct bootstrap samples of size n, by sampling randomly with replacement from the data. Denote the bootstrapped observations by $\{Z_i^*\}_{i=1}^{n}$ and for every $j = 1, 2, ..., J$, let $\mathbb{E}_n^*\left(m_j(Z,\theta)\right)$ denote $n^{-1}\sum_{i=1}^{n} m_j\left(Z_i^*,\theta\right)$. Compute,

$$\Gamma_n^* = \begin{cases} \sup\limits_{\theta\in\hat{\Theta}_I(\tau_n)} G\left(\left\{\begin{array}{c} \left[\sqrt{n}\left(\mathbb{E}_n^*\left(m_j(Z,\theta)\right) - \mathbb{E}_n\left(m_j(Z,\theta)\right)\right)\right]_+ * \\ *1\left[\mathbb{E}_n\left(m_j(Z,\theta)\right) \geq -\tau_n/\sqrt{n}\right] \end{array}\right\}_{j=1}^{J}\right) & \text{if } \hat{\Theta}_I(\tau_n) \neq \varnothing \\ 0 & \text{if } \hat{\Theta}_I(\tau_n) = \varnothing \end{cases}$$

(4) Let $\hat{c}_n^B(1-\alpha)$ be the $(1-\alpha)$ quantile of the bootstrapped distribution of Γ_n^*, approximated with arbitrary accuracy from the previous step. The bootstrap estimate of the $(1-\alpha)$ coverage region for the identified set is given by,

$$\hat{C}_n^B(1-\alpha) = \left\{\theta \in \Theta : \sqrt{n}Q_n(\theta) \leq \hat{c}_n^B(1-\alpha)\right\}$$

In order to implement our procedure, we need to specify the sequence $\{\tau_n\}_{n=1}^{+\infty}$ described in the first step. This sequence enters the procedure in two places. First, it enters in the estimation of the identified set (step 2) and, second, it enters in indicator term in the bootstrap criterion function (step 3)[12]. The restrictions on the rate of the sequence $\{\tau_n\}_{n=1}^{+\infty}$ in first step provide little guidance on how to choose this sequence in a practical application. In our Monte Carlo simulations, the sequences $\tau_n = \ln\ln n$ and $\tau_n = \ln n$ seemed to provide similar and satisfactory results. Based on these findings, it appears that the finite sample performance of our inferential method does not depend critically on the specific choice of this sequence.

The key to the consistency in level of our bootstrap procedure is the bootstrap analogue criterion function defined in step 3. In particular, it is essential to the consistency result that we introduce (a) the indicator function term $1\left[\mathbb{E}_n(m_j(Z,\theta)) \geq -\tau_n/\sqrt{n}\right]$ and (b) the recentering term (that is, subtracting the sample moment from the bootstrap sample moment). The indicator function term is similar to the parametric bootstrap proposed by Andrews [**2**]. The properties required on the sequence $\{\tau_n\}_{n=1}^{+\infty}$ imply that this term indicates whether the corresponding population moment inequality is binding, eventually,

---

[12] In principle, the sequence $\{\tau_n\}_{n=1}^{+\infty}$ in steps 2 and 3 could be two different sequences provided that they both satisfy the rate requirements in step 1. The formal arguments in the appendix allow these two sequences to differ. We restrict both sequences to coincide in order to simplify the notation.

almost surely[13]. Because of these two terms, this procedure differs qualitatively from the bootstrap version of the subsampling methods provided by CHT [**23**]. These differences are analyzed in the appendix (section A.2.1).

Bootstrap procedure for the conditionally separable model. In the conditionally separable model, we will be interested in obtaining rates of convergence (and with them, error in the coverage probability). In order to understand why we need to introduce a separate bootstrap method for the conditionally separable model, we need to distinguish between whether the design of the covariates is fixed or random. The design of the covariates refers to how the econometrician perceives the distribution of covariates in the sample. If the design of the covariates is fixed, then the distribution of the covariates is considered to be non-stochastic (or stochastic and conditioned upon) and if the design of the covariates is random, it is considered to be stochastic. Of course, the inference we perform is different depending on the case.

If the covariates are perceived as fixed, then the cell frequency of the covariates is a constant and we can show that the bootstrap procedure for the general model will deliver rates of convergence of order $n^{-1/2}$. If the covariates are perceived as random, then the cell frequency of the covariates is random. If this is the case, our arguments will only be able to show that the bootstrap procedure proposed in the previous section produces rates of order $n^{-1/2} \ln n \ln \ln n$ (instead of $n^{-1/2}$). Nevertheless, it is possible to design a bootstrap method which can achieve rates of convergence of order $n^{-1/2}$ independently of the design

---

[13] If the moment inequality is binding then, eventually, the corresponding indicator function will be equal to one, almost surely and if the moment inequality is not binding then, eventually, the corresponding indicator function will be equal to zero, almost surely.

of the covariates. This will be referred to as the bootstrap procedure *specialized for the conditionally separable case* and it consists of the following steps:

(1) Choose a positive sequence $\{\tau_n\}_{n=1}^{+\infty}$ such that $\tau_n/\sqrt{n} = o(1)$ and $\sqrt{\ln\ln n}/\tau_n = o(1)$, almost surely,

(2) Estimate of the identified set with,

$$\hat{\Theta}_I(\tau_n) = \left\{\theta \in \Theta : \left\{\hat{p}_k\left(\mathbb{E}_n(Y_j|x_k) - M_j(\theta, x_k)\right) \leq \tfrac{\tau_n}{\sqrt{n}}\right\}_{(j,k)=1}^{J\times K}\right\},$$

(3) Repeat the following for $s = 1, 2, ..., S$. Construct bootstrap samples of size n, by sampling randomly with replacement from the data. Denote the bootstrapped observations by $\{Y_i^*, X_i^*\}_{i=1}^n$, and for every $k = 1, 2, ..., K$, and $j = 1, 2, ..., J$ let $\mathbb{E}_n^*(Y_j|x_k) = n^{-1}\sum_{i=1}^n\{Y_{j,i}^*|X_i^* = x_k\}$ and $\hat{p}_k^* = n^{-1}\sum_{i=1}^n 1\left[X_i^* = x_k\right]$. Compute,

$$\Gamma_n^* = \begin{cases} \sup\limits_{\theta\in\hat{\Theta}_I(\tau_n)} G\left(\left\{\begin{array}{c} \left[\sqrt{n}\hat{p}_k^*\left(\mathbb{E}_n^*(Y_j|x_k) - \mathbb{E}_n(Y_j|x_k)\right)\right]_+ * \\ *1\left[\begin{array}{c}\hat{p}_k\left(\mathbb{E}_n(Y_j|x_k) - M_j(\theta, x_k)\right) \\ \geq -\tau_n/\sqrt{n}\end{array}\right]\end{array}\right\}_{(j,k)=1}^{J\times K}\right) & \text{if } \hat{\Theta}_I(\tau_n) \neq \varnothing \\ 0 & \text{if } \hat{\Theta}_I(\tau_n) = \varnothing \end{cases}$$

(4) Let $\hat{c}_n^B(1-\alpha)$ be the $(1-\alpha)$ quantile of the bootstrapped distribution of Γ_n^*, simulated with arbitrary accuracy from the previous step. The bootstrap estimate of the $(1-\alpha)$ coverage region for the identified set is given by,

$$\hat{C}_n^B(1-\alpha) = \left\{\theta \in \Theta : \sqrt{n}Q_n(\theta) \leq \hat{c}_n^B(1-\alpha)\right\}$$

If the covariates are fixed by design, the general bootstrap procedure and the one specialized for conditionally separable model are identical. When the design is random, these two methods differ in step 3. In particular, when the estimator for the identified set is non-empty, the argument inside the $[\cdot]_+$ function is a random vector rather than a random function. This is the key feature that allows us to obtain rates of convergence of order $n^{-1/2}$.

1.4.2.4. Consistency in level of the bootstrap approximation. In this section, we analyze the asymptotic properties of our bootstrap procedure. As a first step, we show that the distribution of the statistic of interest has a certain asymptotic representation. The statement of this theorem and its proof can be found in the appendix (theorem 36). In order to deduce properties of our bootstrap approximation, we establish that, conditional on the sample, our bootstrap approximation has an analogous asymptotic representation (theorem 38). These two theorems are the key to establish the remaining results of the section.

The following lemma characterizes the limiting distribution of the statistic of interest.

Lemma 6. *Assume (A1)-(A4) and (CF'). If the identified set is non-empty, then* $\lim_{m\to\infty} P\left(\Gamma_m \leq h\right) = P\left(H\left(\zeta\right) \leq h\right),$ *where* H *is the function and* ζ *is the random process described in theorem 36. If the identified set is empty, then for every* $n \in \mathbb{N}$, $P\left(\Gamma_n \leq h\right) = 1\left[h \geq 0\right]$.

Moreover, $\lim_{m\to\infty} P\left(\Gamma_m \leq h\right)$ *is continuous for all* $h \neq 0$.

The previous lemma indicates that the only possible discontinuity of the limiting distribution of the statistic of interest can[14] occur at zero. The possibility of a discontinuity

---

[14] We can construct examples which do and which do not have discontinuity at zero.

at zero is a consequence of undertaking the criterion function approach, since the criterion function maps all realizations where the restrictions imposed by the identified set are satisfied into zero. By choosing the criterion function adequately (with assumption (CF')), we make zero *the only possible* discontinuity point.

The following theorem shows that the bootstrap distribution converges pointwise to the limiting distribution of the statistic of interest for almost all sample sequences.

Theorem 7. *Assume (A1)-(A4) and (CF'). If the identified set is non-empty, then for any* $h \in \mathbb{R}$ *in the continuity set of the limiting distribution,*

$$\lim_{n\to\infty} \left| P\left(\Gamma_n^* \leq h | \mathcal{X}_n\right) - \lim_{m\to\infty} P\left(\Gamma_m \leq h\right) \right| = 0$$

almost surely. If the identified set is empty, then for any $h \in \mathbb{R}$*,*

$$\lim_{n\to\infty} \left| P\left(\Gamma_n^* \leq h | \mathcal{X}_n\right) - \lim_{m\to\infty} P\left(\Gamma_m \leq h\right) \right| = 0$$

almost surely.

The traditional definition of bootstrap consistency (see, e.g., Hall [**32**] or Horowitz [**36**]) requires that the conditional distribution of our bootstrap estimate converges to the limiting distribution of the statistic of interest, *uniformly over the real line.* In the case when the identified set is empty, we have shown that the limiting distribution is degenerate at zero and, eventually, our bootstrap approximation also becomes degenerate at zero, almost surely. In this case, uniform convergence is achieved by pointwise convergence at zero. In the case when the identified set is non-empty, the limiting distribution has a discontinuity at zero. Given this discontinuity, it is possible that the bootstrap approximation we propose fails to converge (pointwise) at zero.

To resolve this issue, our strategy will be to exclude the discontinuity point from our goal. Except on an arbitrarily small neighborhood around zero, we show that the bootstrap approximation is consistent. To this end, we introduce an alternative definition of consistency of bootstrap, namely the consistency of the bootstrap *on a set.*

Definition 8 (Bootstrap consistency on a set $\mathcal{S}$)**.** *The bootstrap estimate of the distribution of interest is consistent on a set* $\mathcal{S} \subseteq \mathbb{R}$ *if and only if for any* $\varepsilon > 0$*,*

$$\lim_{n\to\infty} P\left(\sup_{h\in\mathcal{S}}\left|P\left(\Gamma_n^* \leq h|\mathcal{X}_n\right) - \lim_{m\to\infty} P\left(\Gamma_m \leq h\right)\right| > \varepsilon\right) = 0$$

The definition of bootstrap consistency on a set weakens the traditional definition of bootstrap consistency. Instead of requiring uniform convergence over the real line, it requires uniform convergence over a certain set. This new definition of consistency will be good enough for our purposes if, with probability approaching one, $\mathcal{S}$ includes the quantile we are interested in approximating. As we will explain soon, for the purpose of hypothesis testing, it suffices to consider a compact subset of the real line that excludes zero. This leads to a result we call *bootstrap consistency on any set excluding zero.*

Theorem 9 (Bootstrap consistency on any set excluding zero)**.** *Assume (A1)-(A4) and (CF'). If the identified set is non-empty then, for any* $\mu > 0$,

$$P\left(\lim_{n\to\infty}\sup_{|h|\geq\mu}\left|P\left(\Gamma_n^* \leq h|\mathcal{X}_n\right) - \lim_{m\to\infty} P\left(\Gamma_m \leq h\right)\right| = 0\right) = 0$$

and if the identified set is empty then,

$$P\left(\liminf\left\{\sup_{h\in\mathbb{R}}\left|P\left(\Gamma_n^*\leq h|\mathcal{X}_n\right)-\lim_{m\to\infty}P\left(\Gamma_m\leq h\right)\right|=0\right\}\right)=1$$

In the case of non-empty identified sets, we have dealt with the discontinuity at zero by simply excluding this point from the analysis. But is this point important to obtain a bootstrap approximation at zero? The answer is negative. By theorem 36 and lemma 6, it follows that $\lim_{m\to\infty}P\left(\Gamma_m\leq 0\right)\leq 0.5$. Thus, by the lack of pointwise convergence at zero, we might be unable to approximate a quantile of the distribution that is to the left of the median of the asymptotic distribution. Since the purpose of the approximation is to construct confidence sets, we are typically interested in approximating the 90, 95 and 99 percentiles of the distribution. All these quantiles will map into values for which our consistency result holds, with probability approaching one.

Based on theorem 9, we can show consistency in level of our bootstrap confidence sets.

Corollary 10 (Consistency in level - bootstrap approximation)**.** *Assume (A1)-(A4) and (CF'). If the identified set is non-empty then, for any* $\alpha\in[0,0.5)$

$$\lim_{n\to\infty}P\left(\Theta_I\subseteq\hat{C}_n^B\left(1-\alpha\right)\right)=\left(1-\alpha\right)$$

1.4.2.5. Rates of convergence of the bootstrap approximation. For the rest of this section, we focus on the conditionally separable model, which is described by assumptions (B1)-(B5). For this framework, we can obtain precise rates of convergence of our bootstrap approximation to the finite sample distribution of the statistic of interest. The following result follows from the representation theorems 36 and 38 and the Berry-Esseén theorem.

Theorem 11 (Rate of convergence - bootstrap approximation)**.** *Assume (B1)-(B5) and (CF) and choose the bootstrap procedure to be the one specialized for the conditionally separable model. If the identified set is non-empty then, for any* $\mu > 0$*,*

$$\sup_{|h| \geq \mu} \left| P\left(\Gamma_n^* \leq h | \mathcal{X}_n\right) - P\left(\Gamma_n \leq h\right) \right| = O_p\left(n^{-1/2}\right)$$

and if the identified set is empty then,

$$P\left(\liminf \left\{ \sup_{h \in \mathbb{R}} \left| P\left(\Gamma_n^* \leq h | \mathcal{X}_n\right) - P\left(\Gamma_n \leq h\right) \right| = 0 \right\}\right) = 1$$

We proposed a bootstrap approximation to the distribution of interest in order to construct confidence sets with a prespecified coverage probability. Since this constitutes an approximation, our confidence sets will not have exactly the desired coverage probability. The difference between the desired coverage and the actual coverage is referred to as the *error in the coverage probability* (ECP). By the consistency in level result of the previous subsection, the error in the coverage probability of our bootstrap approximation converges in probability to zero. In the conditionally separable model, theorem 11 can be used to provide an upper bound of the rate at which this convergence occurs. We state this result as a corollary of theorem 11.

Corollary 12 (ECP - bootstrap approximation)**.** *Assume (B1)-(B5), (CF) and choose the bootstrap procedure to be the one specialized for the conditionally separable model. If the identified set is non-empty then, for any* $\alpha \in [0, 0.5)$,

$$\left| P\left(\Theta_I \subseteq \hat{C}_n^B\left(1-\alpha\right)\right) - \left(1-\alpha\right) \right| = O\left(n^{-1/2}\right)$$

and if the identified set is empty then, for any $\alpha \in [0,1]$,

$$P\left(\liminf\left\{\hat{C}_n^B\left(1-\alpha\right)=\hat{\Theta}_I\left(0\right)\right\}\right)=1$$

In terms of coverage, the only relevant case is the non-empty identified set, since the empty set is trivially covered by any confidence set. According to the previous corollary, in the conditionally separable case, the error in the coverage probability converges to zero at a rate of order $n^{-1/2}$.

1.4.3. Alternative procedures

In previous sections, we proposed a bootstrap scheme to perform inference in partially identified models. We showed that it is consistent in level and we characterized its error in coverage probability. In this subsections, we consider two alternative inferential methods.

1.4.3.1. Subsampling. One can consider different subsampling procedures in order to approximate the distribution of interest. For example, one can use the subsampling scheme proposed by CHT [**23**] or one can consider the subsampling analogue of the bootstrap procedure proposed in preceding sections. The basic difference between the two is that the subsampling analogue of our bootstrap procedure will include a recentering term and an indicator function term $1\left[\mathbb{E}_n\left(m_j\left(Z,\theta\right)\right)\geq-\tau_n/\sqrt{n}\right]$ whereas CHT [**23**]'s subsampling will not have such terms[15]. When the identified set is non-empty, the difference between these two will converge in probability to zero[16].

---

[15] The recentering term does not affect the consistency of the approximation, since it converges in probability to zero. Nevertheless, its rate of convergence (in probability) is of order $(b_n/n)^{1/2}$, where b_n denotes the subsampling size. This rate is relatively slow, compared to the one obtained by the fact that we are drawing samples without replacement. We conjecture that this produces even slower rates of convergence than the one obtained by a subsampling procedure that does include recentering.

[16] See lemma 54 in the appendix for the proof.

Our subsampling procedure is as follows:

(1) Choose a positive sequence $\{\tau_n\}_{n=1}^{+\infty}$ such that $\tau_n/\sqrt{n} = o(1)$ and $\sqrt{\ln\ln n}/\tau_n = o(1)$, almost surely,

(2) Estimate of the identified set with,

$$\hat{\Theta}_I(\tau_n) = \left\{\theta \in \Theta : \left\{\mathbb{E}_n(m_j(Z,\theta)) \leq \tau_n/\sqrt{n}\right\}_{j=1}^{J}\right\}$$

(3) Repeat the following step for $s = 1, 2, ..., S$. Construct a subsample of size b_n (with $b_n \to \infty$ and $b_n/n = o(1)$) by sampling randomly without replacement from the data. Denote these observations by $\left\{Z_i^{SS}\right\}_{i=1}^{b_n}$ and for every $j = 1, 2, ..., J$, let $\mathbb{E}^{SS}_{b_n,n}(m_j(Z,\theta))$ denote $b_n^{-1}\sum_{i=1}^{b_n} m_j\left(Z_i^{SS},\theta\right)$. Compute,

$$\Gamma^{SS}_{b_n,n} = \begin{cases} \sup\limits_{\theta\in\hat{\Theta}_I(\tau_n)} G\left(\left\{\begin{array}{c}\left[\sqrt{b_n}\begin{pmatrix}\mathbb{E}^{SS}_{b_n,n}(m_j(Z,\theta))+\\ -\mathbb{E}_n(m_j(Z,\theta))\end{pmatrix}\right]_+ * \\ *1\left[\mathbb{E}_n(m_j(Z,\theta))\geq -\tau_n/\sqrt{n}\right]\end{array}\right\}_{j=1}^{J}\right) & \text{if } \hat{\Theta}_I(\tau_n)\neq\varnothing \\ 0 & \text{if } \hat{\Theta}_I(\tau_n)=\varnothing \end{cases}$$

(4) Let $\hat{c}^{SS}_{b_n,n}(1-\alpha)$ be the $(1-\alpha)$ quantile of the distribution $\Gamma^{SS}_{b_n,n}$, simulated with arbitrary accuracy from the previous step. The subsampling estimate of the $(1-\alpha)$ coverage region for the identified set is given by,

$$\hat{C}^{SS}_{b_n,n}(1-\alpha) = \left\{\theta \in \Theta : \sqrt{n}Q_n(\theta) \leq \hat{c}^{SS}_{b_n,n}(1-\alpha)\right\}$$

If the model is conditionally separable, we can choose to use a subsampling procedure specialized for the framework. In this case, the expression for $\Gamma^{SS}_{b_n,n}$ in step 3 would be replaced by,

$$\Gamma_{b_n,n}^{SS} =$$

$$= \begin{cases} \sup\limits_{\theta \in \hat{\Theta}_I(\tau_n)} G\left(\left(\left\{ \begin{array}{c} \left[\sqrt{b_n} \hat{p}_k^{SS} \left(\begin{array}{c} \mathbb{E}_{b_n,n}^{SS}\left(Y_j | x_k\right) + \\ -\mathbb{E}_n\left(Y_j | x_k\right) \end{array} \right) \right]_+ * \\ *1\left[\hat{p}_k \left(\mathbb{E}_n\left(Y_j | x_k\right) - M_{j,k}(\theta) \right) \geq -\tau_n / \sqrt{n} \right] \end{array} \right\}_{(j,k)=1}^{J \times K} \right) \right) & \text{if } \hat{\Theta}_I(\tau_n) \neq \varnothing \\ 0 & \text{if } \hat{\Theta}_I(\tau_n) = \varnothing \end{cases}$$

where, for each $k = 1, 2, ..., K$, and $j = 1, 2, ..., J$, $\mathbb{E}_{b_n,n}^{SS}(Y_j | x_k) = b_n^{-1} \sum_{i=1}^{b_n} \{Y_{j,i}^{SS} | X_i^{SS} = x_k\}$ and $\hat{p}_k^{SS} = b_n^{-1} \sum_{i=1}^{b_n} 1\left[X_i^{SS} = x_k\right]$.

We can establish theorems along the lines of theorem 9 for the subsampling approximation (theorem 46 in the appendix). Based on these results, we can demonstrate the consistency in level of the subsampling approximation.

Corollary 13 (Consistency in level - subsampling approximation)**.** *Assume (A1)-(A4) and (CF) and let* $\{b_n\}_{n=1}^{+\infty}$ *be a positive sequence such that* $b_n \to \infty$ *and* $b_n/n = o(1)$*. If the identified set is non-empty then, for any* $\alpha \in [0, 0.5)$*,*

$$\lim_{n \to \infty} P\left(\Theta_I \subseteq \hat{C}_{b_n,n}^{SS}(1-\alpha)\right) = (1-\alpha)$$

We can also establish the rate of convergence that can be used to find the error in the coverage probability of the subsampling approximation.

Corollary 14 (ECP - subsampling approximation)**.** *Assume (B1)-(B5), (CF), that the distribution of* $\{Y | X = x_k\}_{k=1}^{K}$ *is strongly non-lattice and let* $\{b_n\}_{n=1}^{+\infty}$ *be a positive sequence such that* $b_n \to \infty$ *and* $b_n/n = o(1)$*. If the identified set is non-empty then, for* *any* $\alpha \in [0, 0.5)$,

$$\left|P\left(\Theta_I \subseteq \hat{C}^{SS}_{b_n,n}(1-\alpha)\right) - (1-\alpha)\right| = O\left(b_n/n + b_n^{-1/2}\right)$$

If the identified set is empty then, for any $\alpha \in [0, 1]$*,*

$$P\left(\liminf\left\{\hat{C}^{SS}_{b_n,n}(1-\alpha) = \hat{\Theta}_I(0)\right\}\right) = 1$$

The corollary establishes an upper bound on the rate at which the error in coverage probability of the subsampling approximation converges to zero. In terms of coverage, the relevant case is the non-empty identified set. In this case, this upper bound depends on the choice of the subsampling size, reflecting the usual trade-off when we choose the subsampling size: increasing subsampling size increases the precision of the averages within a subsample but decreases the total number of subsamples available. The choice of b_n that minimizes this upper bound is $b_n = O\left(n^{2/3}\right)$, which results in error in the coverage probability of order $n^{-1/3}$.

Under certain conditions, we can establish that this rate constitutes not just an upper bound on the error in the coverage probability, but also a lower bound. We now describe the arguments but the formal derivations are provided in the appendix (section A.2.6). Under the assumptions of corollary 14 and using the asymptotic expansion in Babu and Singh [**6**], we show that the conditional distribution of our subsampling approximation has the following asymptotic representation,

$$\underbrace{P\left(\hat{\Gamma}^{SS}_{b_n,n} \leq h|\mathcal{X}_n\right)}_{\text{Subsampling approx.}} = \underbrace{P\left(\Gamma_n \leq h\right)}_{\text{Exact distribution}} + K_1(h)\, b_n^{-1/2} + K_2(h)\,\tfrac{b_n}{n} + o_p\left(\tfrac{b_n}{n} + b_n^{-1/2}\right) \tag{1.1}$$

uniformly over $h \geq \mu$ (for any $\mu > 0$), where $K_1(h)$ and $K_2(h)$ are two non-random functions given in the appendix (lemma 47). From this equation, if follows that, for any $h \geq \mu$, the absolute value of the difference between the subsampling approximation and the exact finite sample distribution is minimized by choosing subsampling size $b_n = C(h)\, n^{2/3}$, where $C(h)$ minimizes $\left| K_1(h)\, C(h)^{-1/2} + K_2(h)\, C(h) \right|$, subject to $C(h) > 0$. If $K_1(h)$ and $K_2(h)$ share the sign, then the convergence rate of the approximation cannot be faster than $n^{-1/3}$.

For the purpose of inference, we will be interested in values of h in a neighborhood of the $(1-\alpha)$ quantile of the the limiting distribution, which we denote $c_\infty(1-\alpha)$. Typically, the $(1-\alpha)$ level of interest is greater than 0.72 (usually: 90%, 95% or 99%) and in the appendix (lemma 47), we show that for all $(1-\alpha) \in (0.72, 1)$, $K_2(c_\infty(1-\alpha))$ is positive. Therefore, if $K_1(c_\infty(1-\alpha))$ is also positive, then the subsampling approximation converges to the distribution of interest at exactly the rate $n^{-1/3}$ (see corollaries 50 and 52 in the appendix). The conditions under which $K_1(c_\infty(1-\alpha))$ is positive involve restrictions on the moments of $\{Y|X = x_k\}_{k=1}^{K}$ that, to the best of our knowledge, lack intuitive interpretation. In the case that $K_1(c_\infty(1-\alpha))$ is non-positive, it might be possible to set the right hand side of equation (1.1) to be $o_p\left(n^{-1/3}\right)$ by a particularly judicious choice of $C(h)$. However, this approach would not be very practical, since it requires careful empirical selection of the subsampling size based on computation of $K_1(c_\infty(1-\alpha))$ and $K_2(c_\infty(1-\alpha))$. In practice, based on asymptotic approximation of equation (1.1), the subsampling size is likely to be chosen as $b_n = Cn^{2/3}$ for a fixed $C > 0$. In this case, unless $K_1(c_\infty(1-\alpha))\, C^{-1/2} + K_2(c_\infty(1-\alpha))\, C = 0$, the subsampling approximation will also converge to the exact distribution at a rate of $n^{-1/3}$.

According to previous sections, the bootstrap delivers error in the coverage probability of order $n^{-1/2}$. Hence, in the conditionally separable model and under certain conditions, the error in the coverage probability of the bootstrap is eventually smaller than the error in the coverage probability produced by subsampling.

1.4.3.2. Asymptotic approximation. Theorem 36 shows that the limiting distribution of the statistic of interest converges weakly to a continuous function of a tight Gaussian process with a certain variance-covariance function. An asymptotic approximation can be constructed by replacing the unknown Gaussian process by a consistent estimate. This procedure will be shown to be consistent in level and, if we restrict attention to the separable framework, will be shown to have the same upper bound on the rate of convergence as our bootstrap procedure.

Formally, we consider the following steps:

(1) Choose a positive sequence $\{\tau_n\}_{n=1}^{+\infty}$ such that $\tau_n/\sqrt{n} = o(1)$ and $\sqrt{\ln\ln n}/\tau_n = o(1)$, almost surely,

(2) Estimate of the identified set with,

$$\hat{\Theta}_I(\tau_n) = \left\{\theta \in \Theta : \left\{\mathbb{E}_n\left(m_j(Z,\theta)\right) \leq \tau_n/\sqrt{n}\right\}_{j=1}^{J}\right\}$$

(3) Repeat the following step for $s = 1, 2, ..., S$. Simulate a zero-mean Gaussian process where, for each $\{\theta_1, \theta_2\} \subseteq \Theta$, its covariance function is given by,

$$\hat{\Sigma}(\theta_1,\theta_2) = \mathbb{E}_n\left[\left(m(Z,\theta_1) - \mathbb{E}_n\left(m(Z,\theta_1)\right)\right)\left(m(Z,\theta_2) - \mathbb{E}_n\left(m(Z,\theta_2)\right)\right)'\right]$$

Denote this Gaussian process by $\hat{Z} : \Omega_n \to l_\infty^J(\Theta)$. Compute,

$$\Gamma_n^{AA} = \begin{cases} \sup\limits_{\theta \in \hat{\Theta}_I(\tau_n)} G\left(\left\{\left[\hat{Z}_j(\theta)\right]_+ * 1\left[\mathbb{E}_n(m_j(Z,\theta)) \geq -\tau_n/\sqrt{n}\right]\right\}_{j=1}^J\right) & \text{if } \hat{\Theta}_I(\tau_n) \neq \varnothing \\ 0 & \text{if } \hat{\Theta}_I(\tau_n) = \varnothing \end{cases}$$

(4) Let $\hat{c}_n^{AA}(1-\alpha)$ be the $(1-\alpha)$ quantile of the distribution Γ_n^{AA} simulated with arbitrary accuracy from the previous step. The subsampling estimate of the $(1-\alpha)$ coverage region for the identified set is given by,

$$\hat{C}_n^{AA}(1-\alpha) = \left\{\theta \in \Theta : \sqrt{n} Q_n(\theta) \leq \hat{c}_n^{AA}(1-\alpha)\right\}$$

If the model is conditionally separable then, in step 3, we simulate from a zero-mean normal vector with variance-covariance matrix given by,

$$\hat{\Sigma} = \mathbb{E}_n\left[\left(\left\{1(X = x_k)\left[Y_j - \mathbb{E}_n(Y_j|x_k)\right]\right\}_{(j,k)=1}^{J\times K}\right)\left(\left\{1(X = x_k)\left[Y_j - \mathbb{E}_n(Y_j|x_k)\right]\right\}_{(j,k)=1}^{J\times K}\right)'\right]$$

Following the steps we used for the bootstrap approximation, we can prove consistency in level for the asymptotic approximation (theorem 56 in the appendix).

Corollary 15 (Consistency in level - asymptotic approximation)**.** *Assume (A1)-(A4) and (CF'). If the identified set is non-empty then, for any* $\alpha \in [0, 0.5)$

$$\lim_{n\to\infty} P\left(\Theta_I \subseteq \hat{C}_n^{AA}(1-\alpha)\right) = (1-\alpha)$$

Moreover, we can also establish the rate of convergence (theorem 57 in the appendix) which can be used to find the error in the coverage probability of the asymptotic approximation.

Corollary 16 (ECP - asymptotic approximation)**.** *Assume (B1)-(B4) and (CF). If the identified set is non-empty then, for any* $\alpha \in [0, 0.5)$,

$$\left| P\left(\Theta_I \subseteq \hat{C}_n^{AA}(1-\alpha)\right) - (1-\alpha) \right| = O\left(n^{-1/2}\right)$$

If the identified set is empty then, for any $\alpha \in [0, 1]$,

$$P\left(\liminf \left\{ \hat{C}_n^{AA}(1-\alpha) = \hat{\Theta}_I(0) \right\}\right) = 1$$

For the conditionally separable model, the bootstrap and the asymptotic approximation have the same (upper bound) of the order of the error in the coverage probability. In other words, the bootstrap does not seem to be providing asymptotic refinements[17]. This is not surprising because the statistic of interest is not asymptotically pivotal. Finally, notice that the asymptotic approximation is implemented by simulation and therefore, requires exactly the same amount of computation as the bootstrap approximation.

1.4.4. Monte Carlo simulations

In order to evaluate the finite sample behavior of the different inferential methods, consider the following binary choice model with missing data. Suppose that we are interested in the decision of individuals between two mutually exclusive and exhaustive choices: choice

---

[17] In order to obtain asymptotic refinements, one could consider a computer intensive procedure called prepivoting or bootstrap iteration. The basic idea is to perform a bootstrap procedure on the bootstrap estimates. This has the effect of performing an approximation which includes one additional term in the Edgeworth expansion. This procedure was introduced by Beran [**9**] and Beran [**10**], formally analyzed by Hall and Martin [**33**] and described in Hall [**32**] and Horowitz [**36**]. The study of the validity of the prepivoting procedure in this setting is out the scope of this paper.

0 or choice 1. Let Y denote this choice, which is assumed to be generated by,

$$Y = 1\left[X\beta \geq \varepsilon\right]$$

where X is a vector of observable vector of explanatory variables with support denoted by $S\left(X\right)$, ε is an unobservable explanatory variable and β denotes the parameters of interest. Assume that $\varepsilon \sim N\left(0,1\right)$ independent of X, which implies that we adopt the probit model. Therefore,

$$P\left(Y = 1|X = x\right) = \mathbb{E}\left(Y|X = x\right) = \Phi\left(x\beta\right)$$

where Φ denotes the standard normal CDF.

Now suppose that we observe the covariates for every respondent but, for some respondents, we do not get to observe the choice. Denote by W the indicator function that takes value one if the choice is observed and zero otherwise. An i.i.d. sample will identify the distribution of the covariates, the distribution of choices conditional on the choice being observed and the probability of observing a response. The identified set is given by,

$$\Theta_I = \left\{\beta \in \Theta : \bigcap_{x \in S(X)} \left\{\mathbb{E}\left(YW|x\right) \leq \Phi\left(x\beta\right) \leq \mathbb{E}\left(YW + \left(1 - W\right)|x\right)\right\}\right\}$$

We consider the following four Monte Carlo designs which differ in the definition of the support of X and in the value of $\mathbb{E}\left(YW|x\right)$ and $\mathbb{E}\left(W|x\right)$ for every x in the support of X. The designs are described in table 1.1.

For all simulations we will sample $N = 600$ observations, with 100 observations for the first covariate, 200 observations for the second covariate and 300 observations for the third

		Covariate values		
		$x_1 = (1,0)$	$x_2 = (0,1)$	$x_3 = (1,1)$
Design 1	$\mathbb{E}\left(YW \mid x\right)$	$\Phi(-0.5)$	$\Phi(-0.5)$	$\Phi(-0.5)$
	$\mathbb{E}\left(W \mid x\right)$	$2\Phi(-0.5)$	$2\Phi(-0.5)$	$2\Phi(-0.5)$
		$x_1 = (1,0)$	$x_2 = (0,1)$	$x_3 = (1,1)$
Design 2	$\mathbb{E}\left(YW \mid x\right)$	$\Phi(-0.5)$	$\Phi(-0.5)$	$\Phi(-1)$
	$\mathbb{E}\left(W \mid x\right)$	$2\Phi\left(-0.5\right)$	$2\Phi\left(-0.5\right)$	$\Phi(-1) + \Phi(-0.5)$
		$x_1 = (1,0)$	$x_2 = (0,1)$	$x_3 = (-1,0)$
Design 3	$\mathbb{E}\left(YW \mid x\right)$	$\Phi(-0.5)$	$\Phi(-0.5)$	$\Phi(-0.5)$
	$\mathbb{E}\left(W \mid x\right)$	$\Phi(-0.5) + \Phi\left(0\right)$	$2\Phi(-0.5)$	$\Phi(-0.5) + \Phi\left(0\right)$
		$x_1 = (1,0)$	$x_2 = (0,1)$	$x_3 = (-1,0)$
Design 4	$\mathbb{E}\left(YW \mid x\right)$	$\Phi(-0.5)$	$\Phi(0)$	$\Phi(-0.5)$
	$\mathbb{E}\left(W \mid x\right)$	$\Phi(-0.5) + \Phi(0.1)$	$\Phi\left(0\right) + \Phi\left(-1\right)$	$\Phi(-0.5) + \Phi\left(0.1\right)$

Table 1.1. Monte Carlo designs

covariate. For each value of the covariate, we sample $\{Y|X\}$ and $\{W|X\}$ independently from a Bernoulli distribution with the mean specified by the table. We treat the design as random.

For our criterion function we choose $G\left(y\right) = \sum_{k=1}^{K} \sum_{j=1}^{J} \left[y_{j,k}\right]_{+}$, which satisfies assumption (CF). Each of the numbers presented in the table are the average of the result of 1000 Monte Carlo simulations. In each simulation, the distribution of the bootstrap, subsampling and asymptotic approximation are approximated from (the same) 200 Monte Carlo draws.

In order to implement our bootstrap and our asymptotic approximation we need to specify the sequence $\{\tau_n\}_{n=1}^{+\infty}$. We conducted simulations with $\tau_n = \ln\ln n$ and $\tau_n = \ln n$ and both specifications provide similar and satisfactory results. From this experience, we conjecture that the results are relatively robust to the choice of the sequence $\{\tau_n\}_{n=1}^{+\infty}$. For the sake of brevity, we only report results with $\tau_n = \ln \ln n$.

1.4.4.1. Design 1. The identified set is described by a pair of moment inequalities for each of the three covariate values. Combining these restrictions, the identified set is described by figure 1.1.

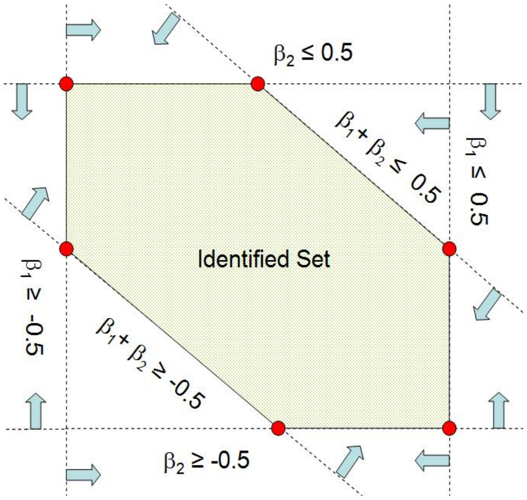

Figure 1.1. Identified set for first Monte Carlo design

The distinctive characteristic of this design is that the identified set has non-empty interior everywhere and that the boundaries of the identified set are defined by, at most, two constrains satisfied with equality. As a consequence, *in this particular case*, we can obtain consistent inference using bootstrap, subsampling and asymptotic approximation *even* if we set $\tau_n = 0$.

We present the result of constructing coverage sets in table 1.2. The first six rows correspond to subsampling procedures. Rows one to four correspond to different versions of the subsampling proposed by CHT [**23**][18]. Rows five and six correspond to the subsampling procedure proposed in section 1.4.3.1. Any of these subsampling procedures require specifying the subsampling size b_n. For the sake of brevity, we show the results for $b_n = n/2$ and $b_n = n/3$ but the results for other choices of subsampling size produced qualitatively similar results. According to CHT [**23**] and the analysis in section 1.4.3.1, the subsampling procedures with $\tau_n = \ln \ln n$ produce consistent inference in level. Rows seven and eight correspond to the naive bootstrap considered in section A.2.1, that is, the inferential procedure that results from replacing the subsampling method proposed by CHT [**23**] with the bootstrap. Recall that, in general, the naive bootstrap will produce inconsistent inference. Rows nine and ten correspond to our bootstrap procedure and our asymptotic approximation, respectively. According to the theoretical results in sections 1.4.2.4 and 1.4.3.2, both inferential schemes generate consistent inference in level. For each empirical coverage, we perform a two sided hypothesis test of whether the empirical coverage coincides with the desired coverage, and the result is represented by stars in the usual way[19].

We first analyze the results of subsampling. When $\tau_n = 0$, both subsampling procedures coincide and given the characteristics of this particular design, should produce consistent inference in level. Our simulations reveal that these procedures result in significant undercoverage in small samples. When $\tau_n = \ln \ln n$, both subsampling procedures should be consistent in level.

---

[18] When implementing this bootstrap analogue, we have respected the way CHT [**23**] estimate the identified set. Also, in their paper, they envision the possibility of performing iterations on their procedure. Our results were obtained with only one iteration.

[19] One star means significant at 10% level, two stars mean significant at 5% level and three stars mean significant at 1% level.

Procedure	Empirical coverage 75%	90%	95%	99%
CHT's Subsampling ($b_n = n/2, \tau_n = 0$)	47.9%***	65.8%***	76.1%***	88.0%***
CHT's Subsampling ($b_n = n/3, \tau_n = 0$)	57.9%***	78.0%***	85.3%***	94.4%***
CHT's Subsampling ($b_n = n/2, \tau_n = \ln\ln n$)	100%***	100%***	100%***	100%***
CHT's Subsampling ($b_n = n/3, \tau_n = \ln\ln n$)	100%***	100%***	100%***	100%***
Our Subsampling ($b_n = n/2, \tau_n = \ln\ln n$)	47.5%***	66.3%***	75.9%***	87.9%***
Our Subsampling ($b_n = n/3, \tau_n = \ln\ln n$)	57.7%***	77.5%***	85.9%***	94.7%***
Naive bootstrap ($\tau_n = 0$)	72.5%*	89.3%	94.2%	98.8%
Naive bootstrap ($\tau_n = \ln\ln n$)	100%***	100%***	100%***	100%***
Our bootstrap ($\tau_n = \ln\ln n$)	74.9%	89.8%	95.4%	99.0%
Our asymptotic approximation ($\tau_n = \ln\ln n$)	74.2%	89.5%	95.0%	99.6%

Table 1.2. Results of first Monte Carlo design

Even though the subsampling scheme proposed by CHT [**23**] is consistent, the results in section A.2.1 in the appendix hint that it could suffer from overcoverage in small samples (due to what we refer as the expansion problem). This is confirmed by our simulations.

The subsampling procedure of section 1.4.3.1 suffers from undercoverage. The poor performance of the subsampling procedures can be attributed to slow rates of convergence. If we implement the naive bootstrap with $\tau_n = 0$ we obtain an accurate approximation. This is a consequence of the simple structure of the current setup and, as we will show, does not hold in general. In turn, when the naive bootstrap is implemented with $\tau_n = \ln\ln n$, we suffer from overcoverage. This is a result of the expansion problem, described in section A.2.1 in the appendix.

The final two procedures correspond to the procedures proposed in this paper. As the table reveals, both approximations with $\tau_n = \ln\ln n$ achieve a very satisfactory performance.

1.4.4.2. Design 2. The identified set in this design is described by figure 1.2.

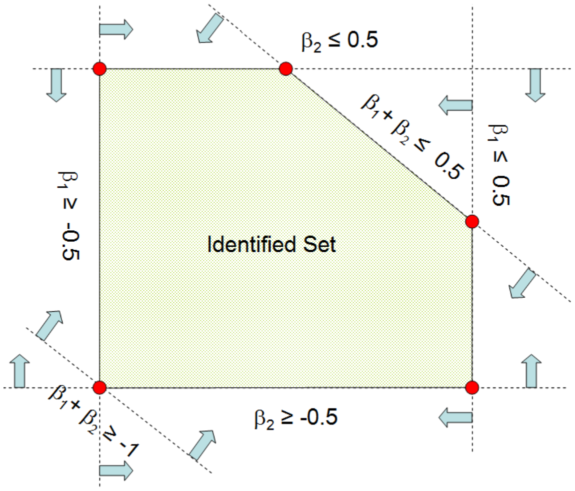

Figure 1.2. Identified set for the second Monte Carlo design

As in the previous design, the identified set has non-empty interior everywhere. The difference with respect to the previous design is that there is one point in the identified set, the point $(\beta_1, \beta_2) = (-0.5, -0.5)$, where one of the restrictions, namely $\beta_1 + \beta_2 \geq 0$, is both irrelevant and satisfied with equality. By the arguments provided in section A.2.1 in the appendix, it is not hard to see that the naive bootstrap procedure will *not* produce consistent inference (no matter how τ_n is chosen).

The failure of the naive bootstrap in this design is related to the boundary problems studied by Andrews [**2**]. The intuition is as follows. The identified set includes the point $(\beta_1, \beta_2) = (-0.5, -0.5)$, which happens to satisfy, with equality, three of the restrictions that define the identified set. In turn, the sample analogue estimate of the identified set will, almost surely, never include any point where three of these restrictions are satisfied with equality. Hence, the binding/non-binding structure of the identified set will almost never coincide with the binding/non-binding structure of its sample analogue.

	Empirical coverage			
Procedure	75%	90%	95%	99%
CHT's subsampling ($b_n = n/2, \tau_n = 0$)	43.7%***	63.9%***	72.7%***	87.3%***
CHT's subsampling ($b_n = n/3, \tau_n = 0$)	55.2%***	72.1%***	81.7%***	92.6%***
CHT's subsampling ($b_n = n/2, \tau_n = \ln \ln n$)	100%***	100%***	100%***	100%***
CHT's subsampling ($b_n = n/3, \tau_n = \ln \ln n$)	100%***	100%***	100%***	100%***
Our subsampling ($b_n = n/2, \tau_n = \ln \ln n$)	43.4%***	64.3%***	73.3%***	88.3%***
Our subsampling ($b_n = n/3, \tau_n = \ln \ln n$)	55.6%***	74.7%***	84.3%***	93.8%***
Naive bootstrap ($\tau_n = 0$)	70.1%***	88.0%**	93.5%**	98.6%
Naive bootstrap ($\tau_n = \ln \ln n$)	100%***	100%***	100%***	100%***
Our bootstrap ($\tau_n = \ln \ln n$)	75.5%	91.6%*	95.9%	99.0%
Our asymptotic approximation ($\tau_n = \ln \ln n$)	75.0%	91.8%*	95.4%	99.0%

Table 1.3. Results of second Monte Carlo design

The results are presented in table 1.3. The subsampling procedures have a mediocre finite sample behavior. If we set $\tau_n = 0$, all subsampling procedures suffer from undercoverage. When $\tau_n = \ln \ln n$, the subsampling procedure proposed by CHT [**23**] suffers from overcoverage whereas the subsampling procedure of section 1.4.3.1 still suffers from undercoverage. Again, we attribute the bad performance of these subsampling schemes to their slow convergence rates.

By the arguments in section A.2.1 of the appendix, the naive bootstrap with $\tau_n = \ln \ln n$ suffers from overcoverage. On the other hand, the naive bootstrap with $\tau_n = 0$ suffers from undercoverage as a consequence of boundary problems.

Our bootstrap and our asymptotic approximation exhibit a very satisfactory performance.

1.4.4.3. Design 3. Figure 1.3 depicts the situation of the identified set.

This design differs from the previous two in that the identified set has empty interior and, with positive probability, the analogy principle estimate of the identified set will be

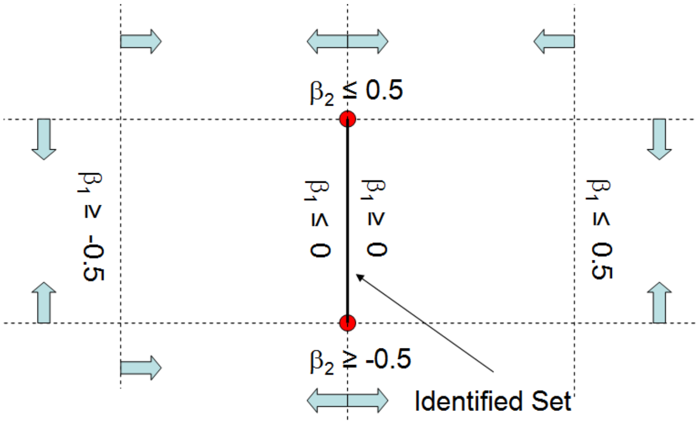

Figure 1.3. Identified set for third Monte Carlo design

	Empirical coverage			
Procedure	75%	90%	95%	99%
CHT's subsampling ($b_n = n/2, \tau_n = 0$)	34.3%***	40.4%***	43.0%***	44.3%***
CHT's subsampling ($b_n = n/3, \tau_n = 0$)	37.8%***	42.9%***	43.9%***	45.4%***
CHT's subsampling ($b_n = n/2, \tau_n = \ln\ln n$)	100%***	100%***	100%***	100%***
CHT's subsampling ($b_n = n/3, \tau_n = \ln\ln n$)	100%***	100%***	100%***	100%***
Our subsampling ($b_n = n/2, \tau_n = \ln\ln n$)	45.6%***	45.6%***	45.6%***	45.7%***
Our subsampling ($b_n = n/3, \tau_n = \ln\ln n$)	45.6%***	45.6%***	45.7%***	45.7%***
Naive bootstrap ($\tau_n = 0$)	45.6%***	45.7%***	45.7%***	45.7%***
Naive bootstrap ($\tau_n = \ln\ln n$)	100%***	100%***	100%***	100%***
Our bootstrap ($\tau_n = \ln\ln n$)	76.2%	89.9%	95.7%	98.8%
Our asymptotic approximation ($\tau_n = \ln\ln n$)	76.4%	90.5%	95.8%	98.9%

Table 1.4. Results of third Monte Carlo design

empty. This illustrates why we need to expand our estimate with the positive sequence $\{\tau_n\}_{n=1}^{+\infty}$ in order to generate an estimator of the identified set that is adequate for inference. By arguments in section A.2.1 of the appendix, the introduction of this sequence will produce inconsistencies in the naive bootstrap, because of the expansion problem. Even if we obtain a non-empty sample estimate of identified set with $\tau_n = 0$, the naive bootstrap will *not* lead to consistent inference due to the boundary problems.

The results are given in table 1.4. As usual, the subsampling procedures have a mediocre finite sample behavior. If we set $\tau_n = 0$, all subsampling procedures suffer from undercoverage. When $\tau_n = \ln \ln n$, the subsampling procedure proposed by CHT [**23**] suffers from overcoverage whereas the subsampling procedure described in section 1.4.3.1 still suffers from undercoverage. As usual, we suspect that the bad performance is associated to the slow convergence rates.

By the arguments in section A.2.1 of the appendix, the naive bootstrap with $\tau_n = \ln \ln n$ should result in overcoverage. This is confirmed by the finite sample behavior. The difference between this design and the previous ones is that if we set $\tau_n = 0$, then, with probability 0.5, this estimated set will be empty, which leads to undercoverage, as shown in the table.

Our bootstrap and our asymptotic approximation procedures combined with $\tau_n = \ln \ln n$, we obtain a very satisfactory finite sample performance.

1.4.4.4. Design 4. In this case, the identified set is empty or, equivalently, the model is misspecified. Since the identified set is empty, the empirical coverage is trivially 100%. Therefore, in this design we will compare the relative sizes of the coverage sets for different inferential methods. In order to achieve this task, we need to define a measure of size of the coverage sets generated by the different inferential methods. For any confidence set $C_n \subseteq \Theta$, we consider the following function,

$$\Pi\left(C_n\right) = \begin{cases} \sup_{\theta \in C_n} \left\{\sqrt{n} Q_n\left(\theta\right)\right\} & \text{if } C_n \neq \varnothing \\ 0 & \text{if } C_n = \varnothing \end{cases}$$

Procedure	Π-size of confidence set 75%	90%	95%	99%
CHT's subsampling ($b_n = n/2, \tau_n = 0$)	0.03	0.05	0.05	0.07
CHT's subsampling ($b_n = n/3, \tau_n = 0$)	0.04	0.05	0.06	0.08
CHT's subsampling ($b_n = n/2, \tau_n = \ln\ln n$)	1.98	2.12	2.22	2.38
CHT's subsampling ($b_n = n/3, \tau_n = \ln\ln n$)	1.81	1.98	2.09	2.30
Our subsampling ($b_n = n/2, \tau_n = \ln\ln n$)	0.14	0.16	0.17	0.19
Our subsampling ($b_n = n/3, \tau_n = \ln\ln n$)	0.13	0.15	0.16	0.18
Naive bootstrap ($\tau_n = 0$)	0.20	0.23	0.24	0.37
Naive bootstrap ($\tau_n = \ln\ln n$)	2.66	2.88	3.01	3.25
Our bootstrap ($\tau_n = \ln\ln n$)	0.54	0.74	0.87	1.11
Our asymptotic approximation ($\tau_n = \ln\ln n$)	0.54	0.75	0.88	1.11

Table 1.5. Results of fourth Monte Carlo design

The function Π constitutes a metric for the size of confidence sets generated by the criterion function approach, which is the case in all the inferential procedures analyzed in this paper. Given any pair of confidence sets constructed using the criterion function approach, denoted C_n and C_n', either $C_n' \subseteq C_n$ or $C_n' \subseteq C_n$ (or both) and, moreover, if $C_n \subseteq C_n'$ then $\Pi(C_n) \leq \Pi(C_n')$ and if $C_n' \subseteq C_n$ then $\Pi(C_n') \leq \Pi(C_n)$.

In table 1.5, we compare the average sizes of the coverage sets for the design, using the average value of the function Π over the Monte Carlo trials. In this case, it only makes sense to compare those methods that are known to produce consistent inference. These are: the subsampling proposed by CHT [**23**] with $\tau_n = \ln\ln n$ (rows three and four), the subsampling of section 1.4.3.1 (rows five and six), our bootstrap (row nine) and our asymptotic approximation (row ten).

The subsampling proposed by CHT [**23**] results in confidence sets that are relatively big and the subsampling proposed in section 1.4.3.1 results in confidence sets that are relatively small. Our bootstrap procedure and our asymptotic approximation generate confidence sets in between these two.

1.5. Confidence sets for each element of the identified set

In this section, we consider the problem of constructing a confidence set that covers each element of the identified set with a minimum prespecified probability. Formally, we are after the construction of a confidence set $C_n(1-\alpha)$ that satisfies condition (1.2).

This construction involves the inversion of simple hypothesis tests. For each point in the parameter space, we perform a hypothesis test, that results in either rejection or lack of rejection. The confidence set for each element of the identified set consists of all the points in the parameter space that are not rejected.

The confidence sets considered in this section can be related to a hypothesis testing problem. Suppose that for a certain parameter value $\theta \in \Theta$, we want to test the null hypothesis $H_0 : \theta \in \Theta_I$ versus the alternative hypothesis $H_1 : \theta \notin \Theta_I$, while keeping the probability of a false rejection to be less or equal than $(1-\alpha)$. A confidence set that satisfies condition (1.2) contains all parameters $\theta \in \Theta$ for which we fail to reject the aforementioned null hypothesis. In partially identified models, we can only restrict the true parameter that generated the data to the identified set. Therefore, a parameter value θ is a candidate to be the true parameter if and only if it belongs to the identified set. In this sense, the hypothesis $H_0 : \theta \in \Theta_I$ can be interpreted as the hypothesis that θ is a candidate for the true parameter value.

1.5.1. Setup

The following set of assumptions conform our econometric model in the i.i.d. setting.

- (C1) For the probability space $(\Omega, \mathcal{B}, \mathbf{P})$, let $Z : \Omega \rightarrow \mathcal{Z}$ be a random vector. We observe an i.i.d. sample $\mathcal{X}_n \equiv \{Z_i\}_{i=1}^{n}$.
- (C2) The parameter space, denoted by Θ, is a compact and convex subset of a finite dimensional Euclidean space $\mathbb{R}^{\eta}$ ($\eta < +\infty$).
- (C3) The identified set, denoted by Θ_I, is given by,

$$\Theta_I = \left\{\theta \in \Theta : \left\{\mathbb{E}\left(m\left(Z, \theta\right)\right) \leq \vec{0}\right\}\right\}$$

  where for each $\theta \in \Theta$, $m\left(z, \theta\right) : \mathcal{Z} \rightarrow \mathbb{R}^{J}$ is a measurable function. Moreover, Θ_I is a proper subset of Θ.
- (C4) For every $\theta \in \Theta$ and every $j = 1, 2, ..., J$, the variance of $m_j\left(Z, \theta\right)$ is positive and finite.
- (C5) For every $\theta \in \Theta$, $m\left(Z, \theta\right)$ has finite fourth absolute moments.

Assumptions (C1)-(C4) constitute a weaker set of assumptions than the ones conforming the general model of section 1.4.1 (assumptions (A1)-(A4)). When the objective is to construct confidence sets for each point in the identified set, these assumptions will deliver consistency in level of our inferential procedures.

When we add assumption (C5) to assumptions (C1)-(C4), we obtain results regarding rates of convergence and error in the coverage probability. As opposed to the problem of construction of confidence set for the identified set, we do not need to assume that the model is conditionally separable to obtain these results. The intuition for this is as follows. In this problem, we are performing individual hypothesis tests for each value in the parameter space. Once the value of the parameter is fixed, we are dealing with random vectors rather than random processes. Hence, under the appropriate moment conditions, we obtain rates of convergence similar to those obtained in the conditionally separable model.

The formal arguments that deliver consistency and rates of convergence are analogous to the ones used for the conditionally separable framework. Hence, by the same reasoning, the bootstrap procedure that is introduced next could be adapted to work on non i.i.d. random settings without affecting the consistency result.

For the type of confidence sets considered in this section, we can generalize the class of criterion functions used. In particular, under criterion functions that satisfy assumption (CF'), we will be able to show consistency in level of our bootstrap approximation. Moreover, dividing the sample moment conditions by the sample standard deviations in the sample analogue criterion function will not affect consistency in level or the rates of convergence.

1.5.2. Bootstrap procedure

We now introduce the bootstrap procedure to construct a confidence region for each element of the identified set. As discussed in section 1.3, such confidence region can be constructed by approximating the $(1-\alpha)$ quantile of the distribution of

$$\Gamma_n(\theta) \equiv a_n Q_n(\theta)$$

for each θ in the parameter space, where Q_n is the sample analogue criterion function and $\{a_n\}_{n=1}^{+\infty}$ is a sequence of constants that makes the (asymptotic) distribution of $a_n Q_n(\theta)$ non-degenerate.

For this purpose, we propose the following bootstrap scheme.

(1) Choose a positive sequence $\{\tau_n\}_{n=1}^{+\infty}$ such that $\tau_n/\sqrt{n} = o(1)$ and $\sqrt{\ln \ln n}/\tau_n = o(1)$, almost surely,

(2) Estimate of the identified set with,

$$\hat{\Theta}_I(\tau_n) = \left\{\theta \in \Theta : \left\{\mathbb{E}_n(m_j(Z,\theta)) \leq \tau_n/\sqrt{n}\right\}_{j=1}^{J}\right\}$$

(3) Repeat the following step for $s = 1, 2, ..., S$ and for every $\theta \in \Theta$. Construct bootstrap samples of size n, by sampling randomly with replacement from the data. Denote the bootstrapped sample by $\{Z_i^*\}_{i=1}^n$ and for every $j = 1, 2, ..., J$ let $\mathbb{E}_n^*(m_j(Z,\theta))$ denote $n^{-1}\sum_{i=1}^n m_j(Z_i^*,\theta)$. Compute,

$$\Gamma_n^*(\theta) = 1\left[\theta \in \hat{\Theta}_I(\tau_n)\right] * G\left(\left\{\begin{array}{c} \left[\sqrt{n}\left(\mathbb{E}_n^*(m_j(Z,\theta)) - \mathbb{E}_n(m_j(Z,\theta))\right)\right]_+ * \\ *1\left[\sqrt{n}\mathbb{E}_n(m_j(Z,\theta)) \geq -\tau_n\right] \end{array}\right\}_{j=1}^{J}\right)$$

(4) Let $\hat{c}_n^B(\theta, 1-\alpha)$ be the $(1-\alpha)$ quantile of the bootstrapped distribution of $\Gamma_n^*(\theta)$, simulated with arbitrary accuracy from the previous step,

(5) The bootstrap estimate of the $(1-\alpha)$ coverage region for each element in the identified set is given by,

$$\hat{C}_n^B(1-\alpha) = \left\{\theta \in \Theta : \sqrt{n}Q_n(\theta) \leq \hat{c}_n^B(\theta, 1-\alpha)\right\}$$

The asymptotic approximation analogue of this procedure was also independently introduced by Soares [**60**], CHT [**23**] and Andrews and Soares [**5**].

It should be noted that the bootstrap procedure we advocate differs qualitatively from replacing the subsampling scheme of CHT [**23**] with the bootstrap. As in the previous section, we refer to the latter resampling scheme as the naive bootstrap. In section A.3.1 of the appendix, we show that, in general, the naive bootstrap results in inconsistent inference. The reason for the inconsistency can be directly related to the inconsistency of the bootstrap on the boundary, studied by Andrews [**2**].

1.5.3. Properties of the bootstrap approximation

The formal analysis of the bootstrap can be established with representation theorems along the lines used to construct confidence sets for the identified set. For the sake of brevity, these theorems are stated and proved in the appendix. These representation theorems characterize the limiting distribution and prove bootstrap consistency excluding zero, which allows us to deduce the consistency in level of the bootstrap approximation.

Corollary 17 (Consistency in level - bootstrap approximation)**.** *Assume (C1)-(C4) and (CF'). If* θ *belongs to the boundary of* Θ_I *then, for any* $\alpha \in [0, 0.5)$,

$$\lim_{n\to\infty} P\left(\theta \in \hat{C}_n^B(1-\alpha)\right) = 1-\alpha$$

If θ *belongs to the interior of* Θ_I *then, for any* $\alpha \in [0, 1]$,

$$\lim_{n\to\infty} P\left(\theta \in \hat{C}_n^B(1-\alpha)\right) = 1$$

If θ does not belong to Θ_I then, for any $\alpha \in [0,1]$,

$$\lim_{n\to\infty} P\left(\theta \notin \hat{C}_n^B(1-\alpha)\right) = 1$$

Moreover, by adding assumption (C5), we can obtain rates of convergence, which allows us to deduce the error in the coverage probability.

Corollary 18 (ECP - bootstrap approximation)**.** *Assume (C1)-(C5) and (CF). If θ belongs to the boundary of Θ_I, then, for any $\alpha \in [0,0.5)$,*

$$\left|P\left(\theta \in \hat{C}_n^B(1-\alpha)\right) - (1-\alpha)\right| = O\left(n^{-1/2}\right)$$

If θ belongs to the interior of Θ_I, then, for any $\alpha \in [0,1]$,

$$\left|P\left(\theta \in \hat{C}_n^B(1-\alpha)\right) - 1\right| = O\left(n^{-1}\right)$$

If θ does not belong to Θ_I, then, for any $\alpha \in [0,1]$,

$$P\left(\theta \in \hat{C}_n^B(1-\alpha)\right) = O\left(n^{-1}\right)$$

If we replace assumption (CF) with assumption (CF') then the only result that changes is that when θ belongs to the boundary of Θ_I, for any $\alpha \in [0,0.5)$,

$$\left|P\left(\theta \in \hat{C}_n^B(1-\alpha)\right) - (1-\alpha)\right| = O\left(n^{-1/2}\ln n\right)$$

For any parameter on the boundary of the identified set, the probability of belonging to the confidence set converges to the desired level with rates of order $n^{-1/2}$ (under (CF)).

Outside the boundary of the identified set, the coverage probability converges to the desired value at a faster rate.

Exactly as in sections 1.4.3.1 and 1.4.3.2, we can propose an asymptotic approximation and a subsampling approximation. These inferential schemes are completely analogous to the bootstrap procedure and are presented in the appendix (sections A.3.6 and A.3.5). Using the same techniques, we can study the rate of convergence and the error in the coverage probability for these alternative methods. Such analysis reveals the following:

a) The bootstrap and the asymptotic approximation have the same (upper bound) on the order of the error in the coverage probability. The bootstrap does not provide refinements.

b) When the parameter belongs to the boundary of the identified set, the subsampling approximation provides an upper bound on the rate of convergence of order $n^{-1/3}$ (by choosing $b_n = O\left(n^{2/3}\right)$). In this case, we can also establish an asymptotic expansion similar to equation (1.1) which implies conditions under which the rate of convergence obtained by subsampling is no better than $n^{-1/3}$. Outside of the boundary of the identified set, subsampling achieves the same rates of convergence as both the bootstrap and the asymptotic approximation.

In summary, our bootstrap and our asymptotic approximation both provide the similar rates of convergence and, under certain conditions, both are orders of magnitude better than the ones obtained by subsampling.

In order to evaluate the finite sample behavior of the inferential methods, we construct confidence sets that cover each element of the identified set for each of the designs of section

1.4.4. The results of these simulations are in line with the results of section 1.4.4 and provided in the appendix (section A.3.7)

1.6. Conclusion

This paper contributes to the growing literature of inference in partially identified or set identified econometric models. We build on the criterion function approach to set inference proposed by Chernozhukov, Hong and Tamer [**23**].

The first contribution of this paper is to introduce a novel bootstrap procedure to construct coverage sets for a wide class of partially identified models. The models considered are those defined by finitely many moment inequalities and equalities, which includes many applications of economic interest.

In the context of inference in partially identified models, there are two possible goals. The first one is to construct a confidence set for the identified set and the second one is to construct a confidence set for each element of the identified set. These two constructions are related to two different and relevant hypothesis testing problems. In order to satisfy both of these objectives, we provide two distinct versions of our bootstrap procedure. Asymptotically, the coverage level provided by our confidence sets converges to the desired coverage level or, equivalently, our procedure is shown to be consistent in level (i.e., not conservative). This constitutes an advantage relative to other inferential procedures that have been proposed in the literature.

Our bootstrap method is shown to be qualitatively different from replacing the subsampling procedure proposed by Chernozhukov, Hong and Tamer [**23**] with the bootstrap.

Performing this replacement will not result in consistent inference due to several problems[20]. Our bootstrap procedure avoids these problems by an adequate definition of the bootstrap criterion function.

The second contribution of our paper is to analyze the rate at which each of the competing inferential methods achieve the desired coverage probability, also known as the error in the coverage probability. Under certain assumptions, we derive the error in the coverage probability of our bootstrap approximation, our asymptotic approximation[21] and a subsampling approximation like the one proposed by Chernozhukov, Hong and Tamer [**23**].

We show that our bootstrap approximation and our asymptotic approximation have error in the coverage probability of (at most) order $n^{-1/2}$. Under certain conditions, we show that the error in the coverage probability of the subsampling approximation converges to zero at a rate of $n^{-1/3}$. As a consequence, under these conditions, our bootstrap and our asymptotic approximation should eventually provide inference that is more precise that the competing subsampling approximation.

Monte Carlo simulations reveal that our bootstrap approximation and our asymptotic approximation have a satisfactory finite sample performance. By considering different setups, the examples show how our inferential methods are not affected by neither boundary nor expansion problems. Moreover, the simulations show that our bootstrap

---

[20] These two problems are described in detail in the supplementary appendix (sections A.2.1 and A.3.1). The first problem is related to the inconsistency of the bootstrap in the boundary, studied by Andrews [**2**] and the second problem is what we refer to as the expansion problem.

[21] As we mentioned earlier, this approximation was independently introduced by Soares [**60**], Andrews and Soares [**5**], Chernozhukov, Hong and Tamer [**23**] and working paper vesions of this paper.

and our asymptotic approximation exhibit a much better finite sample performance than subsampling, in accordance to the results regarding error in the coverage probability.

There are various extensions of this paper to be considered for further research. An important extension would be to allow for continuous covariates, which requires modifying the formal arguments in non trivial ways. Another important extension would be to study the robustness of the bootstrap method proposed in this paper, that is, whether the results obtained by our inferential method hold uniformly over a relevant class of probability distributions. For confidence sets for each element of the identified set, this problem is treated in detail by Andrews and Soares [**5**]. By using the same arguments, we can establish that the inference provided by our bootstrap procedure is also uniformly consistent in level.

CHAPTER 2

Specification Test for Missing Functional Data

2.1. Introduction

Economic data are frequently generated by stochastic processes that can be modeled as occurring in continuous time. The data may then be treated as realizations of random functions (functional data). Examples include wage paths and asset prices or returns. In this case, economic theory may provide a parametric model for the data, that is, a stochastic process that is known up to a finite dimensional parameter that may be the true process that generated the data. In such cases, a natural research question is whether the parametric model is the right model for the data, that is, whether there is a parameter value for which the model is the data generating process. This type of hypothesis test is referred to as a *specification test*.

In a recent paper, Bugni, Hall, Horowitz and Neumann [**19**] (hereafter, referred to as BHHN [**19**]) developed the first method for carrying out a specification test for functional data. Their contribution constitutes the generalization of the Cramér-von Mises[1] specification test to the distribution of random functions that depend on an unknown finite-dimensional parameter vector. Their procedure contributes to the literature by introducing functional data approaches to specification testing in econometrics and by

---

[1] BHHN [**19**] also implement the functional data analogue of the Kolmogorov-Smirnov specification test for functional data. Nevertheless, the Cramér-von Mises test is preferred to the Kolmogorov-Smirnov test since it tends to be more powerful in finite-dimensional settings and it is easier to compute in the infinite-dimensional setting.

developing parametric bootstrap methods that facilitate the use of techniques based on integration over functional spaces.

One weakness of the specification test in BHHN [**19**] is that it does not allow for the existence of missing observations. Both the theoretical results and the empirical implementation of the test require the econometrician to observe a sample of independent and identically distributed functions. This does not only forbid functions to be missing, but it also forbids functions from being unobserved in certain periods, that is, from having missing sections. Unfortunately, this is a strong restriction: missing data is a pervasive problem in most data samples and functional data samples are no exception. The particular feature of functional data is that observations can present missing sections, rather than being completely unobserved.

One might wonder if the specification test developed by BHHN [**19**] can still be applied to a functional data sample with missing observations by eliminating any observations that present missing sections. There are two reasons why this procedure should be avoided. First, the results of this test cannot be extrapolated to the distribution of the data unless we assume that the observed data is a representative sample of the general data, that is, unless missing data is *missing at random*[2]. If the assumption fails, our test results will be contaminated by sample selection bias, which invalidates our test results. Second, in the specific case of functional data, eliminating observations that have some missing sections will eliminate valuable information contained in their non-missing sections.

---

[2] This assumption will fail when there is a selection process deciding which observations are missing and which are not. See, for example, Heckman [**35**] and Manski [**40**]. As explained by Manski [**40**] and Manski [**41**], this assumption cannot be tested, precisely because the corresponding hypothesis depends on the unobserved data.

The objective of this paper is to provide a specification test that can be applied to functional data which is allowed to have missing observations. In order to deal with the missing data problem, we adopt a worst case scenario approach in the spirit of Manski [**40**] and Manski [**41**], which is able to extract all the possible information about the observed data and still be agnostic about the nature of the unobserved data. This approach has the advantage of producing correct conclusions independently of the true distribution of the missing data. Unfortunately, this approach has an unavoidable cost. Without assumptions about the nature of the missing data, the test statistic is *partially or set identified*, that is, it can only be restricted to an interval, even asymptotically. In practice, this implies that it is possible that the hypothesis test is inconclusive, that is, it is not possible to reject or to not reject the null hypothesis. This inconclusive outcome can happen both under the null hypothesis and under the alternative hypothesis.

The remaining of the paper is organized as follows. Section 2.2 describes the hypothesis test developed in BHHN [**19**]. Section 2.3 studies the identification problem posed by the missing data, which is the basis of our hypothesis test. In section 2.4, we introduce our hypothesis test and analyze its theoretical properties. Monte Carlo evidence is presented in section 2.5 and the empirical application to the NLSY79 data is shown in section 2.6. Section 2.7 concludes the paper. All the proofs of the paper are collected in the appendix.

2.2. The BHHN specification test

In this section, we briefly describe the BHHN [**19**] specification test. The observables of the economic phenomenon of interest are random functions distributed according to the data generating process denoted by X. Each realization of X is assumed to belong to $L_2(\mathcal{D})$, almost surely, where $L_2(\mathcal{D})$ denotes the space of square integrable functions defined on the space $\mathcal{D}$. The econometrician observes a random sample of size n of these functions, denoted by $\mathcal{X}_n$. The econometrician conjectures that the data generating process behaves according to a certain process, denoted by Y_θ, which is known up to a finite-dimensional parameter θ that belongs to a parameter space Θ. We assume that for all $\theta \in \Theta$, Y_θ also belongs to $L_2(\mathcal{D})$, almost surely. The objective of the hypothesis test is to decide whether the model $\{Y_\theta : \theta \in \Theta\}$ is a correct specification for X, or not. Formally, the hypotheses of the test are as follows,

$$\begin{aligned} &H_0 : \exists \theta \in \Theta, \text{ such that } X \text{ and } Y_\theta \text{ are equally distributed} \\ &H_1 : \nexists \theta \in \Theta, \text{ such that } X \text{ and } Y_\theta \text{ are not equally distributed} \end{aligned} \tag{2.1}$$

For any non-stochastic function $x \in L_2(\mathcal{D})$ and for any $\theta \in \Theta$, the cumulative distribution function of X and Y_θ are defined as follows,

$$\begin{aligned} F_X(x) &= P\left(X(t) \le x(t), \forall t \in \mathcal{D}\right) \\ F_Y(x|\theta) &= P\left(Y_\theta(t) \le x(t), \forall t \in \mathcal{D}\right) \end{aligned}$$

Under the null hypothesis, there exists a parameter value $\theta \in \Theta$ such that $F_X(x) = F_Y(x|\theta)$ for all $x \in L_2(\mathcal{D})$ and, under the alternative hypothesis, no such parameter value exists.

Let μ be a bounded and non-degenerate measure[3] on $L_2(\mathcal{D})$. As in BHHN [**19**], we can measure distance between the distributions of X and Y_θ with functional-data analogue of

---

[3] For example, this can be the Gaussian process described in BHHN [**19**].

the Cramér-von Mises two-sample statistic, given by the following integral,

$$T\left(X, Y_\theta\right) = \int \left(F_X\left(x\right) - F_Y\left(x|\theta\right)\right)^2 \mu\left(dx\right) \tag{2.2}$$

We assume that there is a unique parameter value that minimizes $T\left(X, Y_\theta\right)$, which we denote by θ_0. The minimized value of $T\left(X, Y_\theta\right)$ allows us to reexpress the hypotheses of our test: under the null hypothesis, $T\left(X, Y_{\theta_0}\right) = 0$, and under the alternative hypothesis, $T\left(X, Y_{\theta_0}\right) > 0$.

The hypothesis test developed by BHHN [**19**] is implemented by estimating the parameter θ_0, replacing cumulative distribution functions by their sample analogues and computing integrals by Monte Carlo integration methods. The asymptotic distribution of the test statistic is approximated using the bootstrap. Formally, the test involves the following sequence of steps.

(1) Estimate the parameter θ_0 root-n-consistently[4] and denote the estimate by $\hat{\theta}_0$.

(2) Compute the sample test statistic, $\hat{T}\left(X, Y_{\hat{\theta}_0}\right)$, which is given by,

$$\hat{T}\left(X, Y_{\hat{\theta}_0}\right) = \tfrac{1}{V}\sum_{j=1}^{V}\left(\hat{F}_X\left(Z_j\right) - \hat{F}_Y\left(Z_j|\hat{\theta}_0\right)\right)^2$$

---

[4] Since the model is known up to a finite dimensional parameter, one could use maximum likelihood estimation. Another possibility is to use the following estimator,

$$\hat{\theta}_0 = \underset{\theta\in\Theta}{\arg\min}\left\{\hat{T}\left(X, Y_\theta\right)\right\}$$

where $\hat{T}\left(X, Y_\theta\right)$ is the sample test statistic described in the second step.
Both estimators are examples of extremum estimators and can be shown to be root-n-consistent under mild regularity conditions. See, e.g., Amemiya [**1**].

where $\{Z_j\}_{j=1}^{V}$ is a random sample[5] of μ and, for every $x \in L_2(\mathcal{D})$, $\hat{F}_X(x)$ and $\hat{F}_Y\left(x|\hat{\theta}_0\right)$ are the sample analogue of the distribution functions, given by,

$$\hat{F}_X(x) = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} 1\left(X^i(t) \le x(t), \forall t \in \mathcal{D}\right)$$
$$\hat{F}_Y\left(x|\hat{\theta}_0\right) = \tfrac{1}{m}\textstyle\sum_{i=1}^{m} 1\left(Y^i_{\hat{\theta}_0}(t) \le x(t), \forall t \in \mathcal{D}\right)$$

where $\left\{Y^i_{\hat{\theta}_0}\right\}_{i=1}^{m}$ is a random sample[6] of $Y_{\hat{\theta}_0}$.

(3) For $s = 1, 2, ..., S$, repeat the following two steps,
   (a) Construct a bootstrap sample of size n of $Y_{\hat{\theta}_0}$, and denote it by $\mathcal{X}_n^*$. Estimate the parameter from the bootstrap sample, denoting the estimate by $\hat{\theta}_0^*$.
   (b) Compute the simulated test statistic, denoted $\hat{T}\left(X^*, Y_{\hat{\theta}_0^*}\right)$.

(4) Denote by $t^*_{\hat{\theta}_0}(1-\alpha)$ the $(1-\alpha)$ quantile of the simulated values of $n\hat{T}\left(X^*, Y_{\hat{\theta}_0^*}\right)$.

(5) Decide the outcome of the test in the following way,

Outcome	Decision
$t^*_{\hat{\theta}_0}(1-\alpha) < n\hat{T}\left(X, Y_{\hat{\theta}_0}\right)$	Reject H_0
$n\hat{T}\left(X, Y_{\hat{\theta}_0}\right) \le t^*_{\hat{\theta}_0}(1-\alpha)$	Do not reject H_0

According to the results in BHHN [**19**], the test has the right level under the null hypothesis, is consistent under a fixed alternative hypothesis and has non-trivial power against sequence of local alternative hypotheses whose distance from the null hypothesis is $O\left(n^{-1/2}\right)$. The test exhibits excellent performance in Monte Carlo simulations.

---

[5] The sample size V is chosen so that $\hat{T}\left(X, Y_{\hat{\theta}_0}\right)$ is an arbitrarily good approximation of $\int\left(\hat{F}_X(x) - F_Y\left(x|\hat{\theta}_0\right)\right)^2 d\mu(x)$.

[6] The sample size m is chosen so that $\hat{F}_Y\left(x|\hat{\theta}\right)$ is an arbitrarily good approximation of $F_Y\left(x|\hat{\theta}\right)$.

2.3. Identification analysis for missing functional data

We now consider how missing data in the a functional data sample affects the BHHN [**19**] specification test. It does so in two ways. First, missing data may affect our ability to consistently estimate the parameter θ_0 (step 1). This will certainly be the case our estimator is obtained by maximum likelihood method based on the value of all the observations in the interval $\mathcal{D}$. If we cannot compute the estimate, we cannot compute the test statistic (step 2) and we also will be unable to simulate the critical value (step 3). Second, missing data will forbid us from identifying the distribution of the observables, which we denoted by F_X.

The first problem may be avoided if we manage to estimate the parameter (root-n) consistently in spite of the missing data problem. For example, suppose that our sample consists of observations of sample paths of an economic phenomenon over of two years and we suffer from sample attrition exclusively during the second year. It may be possible to estimate the parameter using exclusively the information of the first year, where the sample is completely observed. In comparison, the second problem is unavoidable. If we are unwilling to make assumptions about the nature of the missing data, any period of unobserved data for functions in our sample implies that the distribution of the data is unidentified.

For most of the analysis of the paper, we will assume that the first problem can be avoided, that is, we will assume there is a root-n consistent estimator of θ_0, and so we will focus the analysis on providing an answer to the second problem[7].

---

[7] In section 2.4.4 we consider how the results can be modified when a root-n consistent estimator of θ_0 is unavailable.

The derivation of the test will proceed as follows. The first step will be to derive the identified set of the distribution of the sample. The second step will be to use this set to derive worst case scenario bounds for the test statistic. In the first two steps, we will assume that we know the population from where the observed data is sampled (of course, missing data is still unobserved) and, as a consequence, we compute the population version of these worst case scenario bounds. In the final step, we replace use sample analogue estimation and Monte Carlo integration to obtain estimates of the worst case scenario bounds, which allows us to implement a specification test for missing data.

2.3.1. Identified set for the cumulative distribution function

This section characterizes the identified set for the cumulative distribution function of the observables when there is missing functional data.

We will assume that the random sample consists of functions of time, whose paths are observed over an interval denoted by $\mathcal{D}$. Suppose that this interval can be divided into K periods, which we denote by $\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_K$. These periods are defined so that, in every period, every function in the sample is either observed (not missing) or unobserved (missing). In other words, no function can be partly observed and unobserved in any of these periods. Since there are K periods where a function can be either observed or unobserved, there are potentially 2^K missing data patterns[8]. Let the integer valued variable π indicate the pattern of missing data, so $\pi \in \left\{1, 2, ..., 2^K\right\}$. For each possible pattern of missing data, we can split the function into observed and unobserved sections. Under the missing data pattern π, denote by $O(\pi)$ and $U(\pi)$ those periods that are

---

[8] In applications, it is possible that the number of missing data patterns is much smaller. For example, if missing data is caused by permanent sample attrition, there can only be K possible missing data patterns.

observed and unobserved, respectively. Without loss of generality, the possible missing data patterns can be labeled so that $\pi = 1$ denotes the case when there is no missing data, and therefore, $O(1) = \mathcal{D}$ and $U(1) = \varnothing$. For example, consider a two-period model, i.e. $K = 2$. In this case, there are four possible missing data patters, which are described in the following table.

π	Description of the pattern	$O(\pi)$	$U(\pi)$
1	Not missing in period 1 and not missing in period 2	$\mathcal{D}$	$\varnothing$
2	Not missing in period 1 and missing in period 2	$\mathcal{D}_1$	$\mathcal{D}_2$
3	Missing in period 1 and not missing in period 2	$\mathcal{D}_2$	$\mathcal{D}_1$
4	Missing in period 1 and missing in period 2	$\varnothing$	$\mathcal{D}$

By the law of total probability, we can rewrite $F_X(x)$ as,

$$F_X(x) = \sum_{j=1}^{2^K} P(X(t) \leq x(t), \forall t \in \mathcal{D} | \pi = j) P(\pi = j) \tag{2.1}$$

For any $j \in \{1, 2, ..., 2^K\}$, consider the following derivation,

$$\begin{aligned}
&P(X(t) \leq x(t), \forall t \in \mathcal{D} | \pi = j) = \\
&= P\left(\cap_{i=1}^{K} \{X(t) \leq x(t), \forall t \in \mathcal{D}_i\} | \pi = j\right) \\
&= P(\{X(t) \leq x(t), \forall t \in O(j)\} \cap \{X(t) \leq x(t), \forall t \in U(j)\} | \pi = j) \\
&= \left\{ \begin{array}{c} P(\{X(t) \leq x(t), \forall t \in U(j)\} | \pi = j, \{X(t) \leq x(t), \forall t \in O(j)\}) * \\ *P(\{X(t) \leq x(t), \forall t \in O(j)\} | \pi = j) \end{array} \right\}
\end{aligned}$$

To simplify notation, denote $F_X(x, O(j) | \pi = j) = P(\{X(t) \leq x(t), \forall t \in O(j)\} | \pi = j)$ and, by convention, $F_X(x, O(1) | \pi = 1) = F_X(x | \pi = 1)$. Using the previous derivation in equation 2.1, we deduce that,

(2.2)

$$F_X(x) = F_X(x|\pi = 1) P(\pi = 1) + \\ + \sum_{j=2}^{2^K} \left\{ \begin{array}{c} P(\{X(t) \leq x(t), \forall t \in U(j)\} | \pi = j, \{X(t) \leq x(t), \forall t \in O(j)\}) * \\ * F_X(x, O(j) | \pi = j) P(\pi = j) \end{array} \right\}$$

From a random sample with missing data we can identify the frequency of each missing data pattern (that is, $P(\pi = j)$ for every $j = 1, 2, ..., 2^K$) and the distribution of the random sample of functions where these functions are observed (that is, $F_X(x, O(j) | \pi = j)$ for every $j = 1, 2, ..., 2^K$). Missing data generates an identification problem because we know nothing about the (conditional) distribution of the random sample of functions where these functions are not observed (that is, about the expression $P(\{X(t) \leq x(t), \forall t \in U(j)\} | \pi = j, \{X(t) \leq x(t), \forall t \in O(j)\})$ for all $j = 2, 3, ..., 2^K$). We obtain worst case scenario bounds for the distribution of the data by imposing logical bounds to these expressions.

Lemma 19. *For any $x \in L_2(\mathcal{D})$, the worst case scenario bounds for $F_X(x)$ are given by,*

$$F_X(x|\pi = 1) P(\pi = 1) \leq F_X(x)$$

$$F_X(x) \leq F_X(x|\pi = 1) P(\pi = 1) + \sum_{j=2}^{2^K} F_X(x, O(j) | \pi = j) P(\pi = j)$$

Moreover, these bounds are sharp.

The worst case scenario bounds described by lemma 19 are sharp in the sense that for a fixed $x \in L_2(\mathcal{D})$, the value of $F_X(x)$ cannot be restricted any further.

In addition to the restrictions imposed by lemma 19, the distribution of the data needs to satisfy the restrictions imposed by the fact that it is a cumulative distribution function. Denote by $x_{-\infty}$ and $x_{+\infty}$ the functions that map every element of $\mathcal{D}$ into minus and plus infinity, respectively, i.e., for every $t \in \mathcal{D}$, $x_{-\infty}(t) = -\infty$ and $x_{+\infty}(t) = +\infty$. We denote by Γ the set of functions that map $L_2(\mathcal{D})$ into $\mathbb{R}$ that satisfies the defining properties of a cumulative distribution function, that is, if $F \in \Gamma$, then $F : L_2(\mathcal{D}) \to [0,1]$, and (i) $\forall x_1, x_2 \in L_2(\mathcal{D})$ such that $x_1(t) \leq x_2(t)$ for every $t \in \mathcal{D}$, then $F(x_1) \leq F(x_2)$, (ii) F is right continuous, (iii) $\lim_{x \to x_{-\infty}} F(x) = 0$ and (iv) $\lim_{x \to x_{+\infty}} F(x) = 1$.

The following lemma characterizes the identified set for the cumulative distribution function of the data.

Lemma 20. *Define* $F_X^L : L_2(\mathcal{D}) \to \mathbb{R}$ *and* $F_X^H : L_2(\mathcal{D}) \to \mathbb{R}$ *as follows. For every* $x \in L_2(\mathcal{D})$,

$$\begin{aligned} F_X^L(x) &= F_X(x|\pi = 1) P(\pi = 1) \\ F_X^H(x) &= F_X(x|\pi = 1) P(\pi = 1) + \sum_{j=2}^{2^K} F_X(x, O(j) | \pi = j) P(\pi = j) \end{aligned}$$

The identified set for F_X, *denoted* $\mathcal{H}(F_X)$, *is given by,*

$$\mathcal{H}(F_X) = \left\{ \Gamma \cap \left\{ G : F_X^L \leq G \leq F_X^H \right\} \right\}$$

Lemma 19 implies that, for every $x \in L_2(\mathcal{D})$, F_X is constrained by a lower bound and an upper bound, denoted, respectively, by F_X^L and F_X^H. Since F_X is a cumulative distribution function, F_X belongs to the set Γ. Lemma 20 states that the identified set for F_X is only composed of all the cumulative distribution functions that are restricted by the worst case scenario bounds imposed by lemma 19.

It should be noted that not every mapping from $L_2(\mathcal{D})$ into $\mathbb{R}$ that satisfies the worst case scenario bounds imposed by lemma 19 is a cumulative distribution function. For example, the lower worst case scenario bound, F_X^L, satisfies these bounds but is not a cumulative distribution function as $\lim_{x \to x+\infty} F_X^L(x) = P(\pi = 1)$, which is less than one whenever there is missing data.

2.3.2. Bounds for the test statistic

In this section, we use the identified set for the cumulative distribution function of the data to develop worst case scenario bounds for the population version of the test statistic. The following theorem presents the result.

Theorem 21. *Let* $T_L(X, Y_{\theta_0})$ *and* $T_H(X, Y_{\theta_0})$ *be defined as follows,*

$$\begin{aligned} T_L(X, Y_{\theta_0}) &= \inf_{G \in \mathcal{H}(F_X)} \int (G(x) - F_Y(x|\theta_0))^2 \mu(dx) \\ T_H(X, Y_{\theta_0}) &= \sup_{G \in \mathcal{H}(F_X)} \int (G(x) - F_Y(x|\theta_0))^2 \mu(dx) \end{aligned}$$

The population version of the test statistic satisfies the following worst case scenario bounds,

$$T_L(X, Y_{\theta_0}) \leq T(X, Y_{\theta_0}) \leq T_H(X, Y_{\theta_0})$$

Moreover, these bounds are sharp.

The presence of missing data opens a gap between the worst case lower bound of the test statistic and the worst case upper bound of the test statistic. This gap reflects our ignorance about the distribution of the missing data. The worst case scenario bounds for the population test statistic provided in the theorem are sharp, in the sense that the possible values for the test statistic cannot be restricted any further[9]. In this sense, these bounds represent the best we can offer with the available information.

In the hypothetical case in which we know the population of the observed data, computing the sharp worst case scenario bounds is complicated because they require calculation of infimum or supremum on $\mathcal{H}(F_X)$, which is a set of functions. To circumvent this computational problem, we can consider *alternative worst case scenario bounds* that are easier to compute. Instead of computing infimum or supremum over the set $\mathcal{H}(F_X)$, these alternative bounds are the result of restricting to the following strict superset of $\mathcal{H}(F_X)$,

$$\mathcal{H}'(F_X) = \left\{G : F_X^L \leq G \leq F_X^H\right\}$$

Essentially, $\mathcal{H}'(F_X)$ ignores the restriction imposed by the fact that the distribution of the data needs to satisfy the defining properties of a cumulative distribution function.

---

[9] Formally, there is some distribution of the missing data such that the resulting test statistic is arbitrarily close to both the upper or the lower bound.

Theorem 22. *Let* $T_L'(X, Y_{\theta_0})$ *and* $T_H'(X, Y_{\theta_0})$ *be defined as follows,*

$$\begin{aligned} T_L'(X, Y_{\theta_0}) &= \inf_{G \in \mathcal{H}'(F_X)} \int (G(x) - F_Y(x|\theta_0))^2 \mu(dx) \\ T_H'(X, Y_{\theta_0}) &= \sup_{G \in \mathcal{H}'(F_X)} \int (G(x) - F_Y(x|\theta_0))^2 \mu(dx) \end{aligned}$$

where $\mathcal{H}'(F_X) = \left\{G : F_X^L \leq G \leq F_X^H\right\}$. *The population version of the test statistic satisfies the following worst case scenario bounds,*

$$T_L'(X, Y_{\theta_0}) \leq T(X, Y_{\theta_0}) \leq T_H'(X, Y_{\theta_0})$$

Moreover, these bounds are equivalent to those which result of imposing the alternative worst case scenario bounds individually for every $x \in L_2(\mathcal{D})$ *(i.e., the ones derived in lemma 19). Consequently, the worst case bounds can be computed as follows,*

$$\begin{aligned} T_L'(X, Y_{\theta_0}) &= \int \left\{ \begin{array}{c} 1\left[F_Y(x|\theta_0) < F_X^L(x)\right] \left(F_X^L(x) - F_Y(x|\theta_0)\right)^2 + \\ +1\left[F_Y(x|\theta_0) > F_X^H(x)\right] \left(F_X^H(x) - F_Y(x|\theta_0)\right)^2 \end{array} \right\} \mu(dx) \\ T_H'(X, Y_{\theta_0}) &= \int \max\left\{ \left(F_X^L(x) - F_Y(x|\theta_0)\right)^2, \left(F_X^H(x) - F_Y(x|\theta_0)\right)^2 \right\} \mu(dx) \end{aligned}$$

The only advantage of the alternative worst case scenario bounds with respect to the sharp ones is that they have a simple closed form expression. The potential disadvantage of the alternative bounds is that they may not be sharp, that is, they might not exhaust all the information contained in the data. The following theorem refers to this disadvantage.

Theorem 23. *The alternative worst case scenario lower bound coincides with the sharp worst case scenario lower bound,*

$$T_L'(X, Y_{\theta_0}) = T_L(X, Y_{\theta_0})$$

The alternative worst case scenario upper bound is greater or equal than the sharp worst case scenario upper bound,

$$T_H'(X, Y_{\theta_0}) \geq T_H(X, Y_{\theta_0})$$

This weak inequality might or might not be strict.

The result shows that alternative worst case scenario lower bound is sharp but the alternative worst case scenario upper bound may or may not be sharp.

2.3.3. Alternative identifying assumptions

Until now, we have considered the identified set of the distribution of the data and the test statistic without making any assumptions about the nature of the missing data. In certain situations, the econometrician might be willing to introduce a priori information about the distribution of the missing data which can restrict the distribution of the data. This is the content of the following lemma.

Lemma 24. *Suppose that for every $x \in L_2(\mathcal{D})$ and $j \in \left\{2, ..., 2^K\right\}$ we assume that,*

$$B_j^L(x) \leq F_X(x, U(j) | \pi = j, \{X(t) \leq x(t), \forall t \in O(j)\}) \leq B_j^H(x)$$

Define $R_X^L : L_2(\mathcal{D}) \to \mathbb{R}$ *and* $R_X^H : L_2(\mathcal{D}) \to \mathbb{R}$ *in the following way. For every* $x \in L_2(\mathcal{D})$,

$$\begin{aligned} R_X^L(x) &= F_X(x|\pi=1)\,P(\pi=1) + \sum_{j=2}^{2^K} B_j^L(x)\,F_X(x, O(j)\,|\pi=j)\,P(\pi=j) \\ R_X^H(x) &= F_X(x|\pi=1)\,P(\pi=1) + \sum_{j=2}^{2^K} B_j^H(x)\,F_X(x, O(j)\,|\pi=j)\,P(\pi=j) \end{aligned}$$

The restricted identified set for F_X, *denoted* $\mathcal{H}_R(F_X)$, *is given by,*

$$\mathcal{H}_R(F_X) = \left\{\Gamma \cap \left\{G : R_X^L \leq G \leq R_X^H\right\}\right\}$$

The previous lemma shows how a priori information about the missing data can be incorporated to obtain a new identified set for the cumulative distribution function of the data[10]. This can be used to establish the worst case scenario bounds for the test statistic with additional information. We do so in the following theorem.

Theorem 25. *Suppose that for every* $x \in L_2(\mathcal{D})$ *and* $j \in \left\{2, ..., 2^K\right\}$ *we assume that,*

$$B_j^L(x) \leq F_X(x, U(j)\,|\pi=j, \{X(t) \leq x(t), \forall t \in O(j)\}) \leq B_j^H(x)$$

These bounds will determine a restricted identified set for the distribution of the data, denoted $\mathcal{H}_R(F_X)$, *and given in lemma 24. Let* $T_L^R(X, Y_{\theta_0})$ *and* $T_H^R(X, Y_{\theta_0})$ *be defined as*

---

[10] Notice that the identified set derived without any a priori information (lemma 20) is the special case of the one derived in lemma 24 when for every $x \in L_2(\mathcal{D})$ and for every $j \in \left\{2, ..., 2^K\right\}$, $B_j^L(x) = 0$ and $B_j^H(x) = 1$.

follows,

$$\begin{aligned} T_L^R\left(X, Y_{\theta_0}\right) &= \inf_{G \in \mathcal{H}_R(F_X)} \int \left(G\left(x\right) - F_Y\left(x|\theta_0\right)\right)^2 \mu\left(dx\right) \\ T_H^R\left(X, Y_{\theta_0}\right) &= \sup_{G \in \mathcal{H}_R(F_X)} \int \left(G\left(x\right) - F_Y\left(x|\theta_0\right)\right)^2 \mu\left(dx\right) \end{aligned}$$

Then, the population version of the test statistic satisfies the following worst case scenario bounds,

$$T_L^R\left(X, Y_{\theta_0}\right) \leq T\left(X, Y_{\theta_0}\right) \leq T_H^R\left(X, Y_{\theta_0}\right)$$

Moreover, these bounds are sharp.

In the next subsections, we provide examples of additional information about the missing data that results in restricted worst case scenario bounds.

2.3.3.1. Example 1: Missing at random. Missing a random is an extreme assumption that delivers an extreme result: point identification of the distribution of the data.

Condition 26 (Missing at random)**.** *Observations are randomly selected into the different missing data patterns. As a consequence, the unobserved data are distributed in the same way as the observed data, and so, for every* $j \in \left\{2, ..., 2^K\right\}$,

$$F_X\left(x, U\left(j\right)|\pi = j, \left\{X\left(t\right) \leq x\left(t\right), \forall t \in O\left(j\right)\right\}\right) =$$

$$= F_X\left(x, U\left(j\right)|\pi = 1, \left\{X\left(t\right) \leq x\left(t\right), \forall t \in O\left(j\right)\right\}\right)$$

where $\pi = 1$ *represents the subsample with no missing data.*

Under this assumptions, the restricted worst case scenario bounds collapse to a unique value, which is the population value of the test statistic,

$$T_L^R\left(X, Y_{\theta_0}\right) = T_H^R\left(X, Y_{\theta_0}\right) = \int \left(F_X\left(x\right) - F_Y\left(x|\theta_0\right)\right)^2 \mu\left(dx\right)$$

This assumption is underlying every study where missing observations are ignored or eliminated from the sample. In particular, it is implicit in the empirical application conducted in BHHN [**19**].

2.3.3.2. Example 2: Stochastic domination. In certain functional data settings, the econometrician might be willing to assume that the distribution of the unobserved functions first order stochastically dominates (or is dominated by) the distribution of observed functions. In this case, the test statistic can be restricted to a strict subset of the sharp worst case scenario bounds.

Usually, the stochastic domination assumption can be motivated by an assumption about how data becomes missing. As an example, consider a sample constituted by wage paths for a cross section of individuals. If we are willing to assume that a part of the wage path that is missing is likely to be caused by unemployment, then this can imply that the distribution of wages that are unobserved dominates the distribution of wages that are observed. A similar example occurs when the sample is constituted by a cross section of stock prices paths, where missing data might be caused by bankruptcy.

Condition 27 (Stochastic domination)**.** *The distribution of unobserved functions first order stochastically dominates (or is dominated by) the distribution of the observed functions. Formally, if the distribution of unobserved functions dominates the distribution of* *the observed functions, then, for every* $j \in \left\{2, ..., 2^K\right\}$,

$$F_X\left(x, U\left(j\right) | \pi = j, \left\{X\left(t\right) \leq x\left(t\right), \forall t \in O\left(j\right)\right\}\right) \geq$$

$$\geq F_X\left(x, U\left(j\right) | \pi = 1, \left\{X\left(t\right) \leq x\left(t\right), \forall t \in O\left(j\right)\right\}\right)$$

where $\pi = 1$ *represents the subsample with no missing data. If the distribution of unobserved functions is dominated by the distribution of the observed functions, then, the direction of the previous inequality is reversed.*

If the distribution of unobserved functions dominates the distribution of the observed functions, we deduce the following worst case scenario bounds for the distribution of the data at any $x \in L_2\left(\mathcal{D}\right)$ and $j \in \left\{2, ..., 2^K\right\}$,

$$B_j^L\left(x\right) \leq F_X\left(x, U\left(j\right) | \pi = j, \left\{X\left(t\right) \leq x\left(t\right), \forall t \in O\left(j\right)\right\}\right) \leq B_j^H\left(x\right)$$

where $B_j^L\left(x\right) = F_X\left(x, U\left(j\right) | \pi = 1, \left\{X\left(t\right) \leq x\left(t\right), \forall t \in O\left(j\right)\right\}\right)$ and $B_j^H\left(x\right) = 1$. If, instead, the distribution of the observed data is dominated by the distribution of the unobserved data, then for $x \in L_2\left(\mathcal{D}\right)$ and $j \in \left\{2, ..., 2^K\right\}$, the previous restriction holds with $B_j^L\left(x\right) = 0$ and $B_j^H\left(x\right) = F_X\left(x, U\left(j\right) | \pi = 1, \left\{X\left(t\right) \leq x\left(t\right), \forall t \in O\left(j\right)\right\}\right)$. Theorem 25 provides the sharp restricted worst case scenario bounds for the test statistic, which will be narrower than the unrestricted bounds.

2.4. Specification test for missing functional data

This section utilizes the identification analysis of section 2.3 to develop our specification test for missing functional data.

2.4.1. Assumptions

In order to derive properties of the proposed hypothesis test, we consider the following set of assumptions,

(A1) The observed data, denoted by $\mathcal{X}_n = \{X_1, X_2, ...X_n\}$, is the result of missing data affecting an independent and identically distributed random sample from the population whose cumulative distribution function is F_X.

(A2) (i) θ_0 is uniquely defined as follows,

$$\theta_0 = \underset{\theta \in \Theta}{\arg\min} \{T(X, Y_\theta)\}$$

(ii) $\hat{\theta}$ is a estimator for θ_0 that has the following asymptotic representation,

$$n^{1/2}\left(\hat{\theta} - \theta_0\right) = n^{-1/2}\sum_{i=1}^{n}\Omega(X_i) + o_p(1)$$

where Ω is a p-vector valued function such that $\mathbb{E}(\Omega(X)) = 0$ and $cov(\Omega(X))$ is non-singular and $\int \Omega(X)' \Omega(X) \mu(dx) < \infty$.

(A3) $\partial F_Y(\cdot|\theta)/\partial\theta$ exists for all θ in an open set $\mathcal{O}$ that contains θ_0. Moreover,

$$\sup_{\theta \in \mathcal{O}} \int \frac{\partial F_Y(x|\theta)}{\partial \theta'} \frac{\partial F_Y(x|\theta)}{\partial \theta} d\mu(x) < \infty$$

and

$$\lim_{\varepsilon \to 0} \int \sum_{i,j=1,2,\dots,p} \sup_{\|\theta - \theta_0\| \leq \varepsilon} \left| \begin{array}{c} \left[\frac{\partial F_Y(x|\theta)}{\partial \theta_i} - \frac{\partial F_Y(x|\theta_0)}{\partial \theta_i}\right] * \\ * \left[\frac{\partial F_Y(x|\theta)}{\partial \theta_j} - \frac{\partial F_Y(x|\theta_0)}{\partial \theta_j}\right] \end{array} \right| d\mu(x) = 0$$

where $\|\theta - \theta_0\|$ denotes the Euclidean distance between θ and θ_0.

(A4) μ is the measure induced by the following Gaussian process,

$$Z(t) = \sum_{k=1}^{\infty} \rho_k N_k \phi_k(t)$$

where, for all $k \in \mathbb{N}$, $0 < |\rho_k| \leq Ck^{-d}$ for some constants $C < \infty$ and $d > 1$, $\{N_k\}_{k=1}^{+\infty}$ is a sequence of independent standard normal random variables and $\phi_k(t) = \sqrt{2}\sin(k\pi t)$.

These assumptions are exactly the assumptions imposed by BHHN [**19**]. In particular, notice that assumption (A2)-(ii) implies that missing data does not preclude our ability to estimate the parameter θ_0 in a root-n-consistently. As we mentioned in section 2.3, this is possible when, for example, all the functions in the random sample are observed in a certain period (typically, the first one).

2.4.2. Implementation of the test

In order to implement the test, we replace cumulative distribution functions by their sample analogues. Our specification test for missing functional data is given by the following steps,

(1) Estimate θ_0 using an estimation procedure that is root-n consistent under the presence of missing data. Denote this estimate $\hat{\theta}_0$.

(2) Estimate the sharp upper and lower bounds for the cumulative distribution function of the data, denoted, respectively, by $\hat{F}_X^L$ and $\hat{F}_X^H$. For every $x \in L_2(\mathcal{D})$,

these are given by,

$$\begin{aligned} \hat{F}_X^L(x) &= \hat{F}_X(x|\pi=1)\,\hat{P}(\pi=1) \\ \hat{F}_X^H(x) &= \hat{F}_X(x|\pi=1)\,\hat{P}(\pi=1) + \sum_{j=2}^{2^K} \hat{F}_X(x, O(j)\,|\pi=j)\,\hat{P}(\pi=j) \end{aligned}$$

where, for every $x \in L_2(\mathcal{D})$, and for every $j = 1, 2, ..., K$, the expressions $\hat{P}(\pi=j)$ and $\hat{F}_X(x, O(j)\,|\pi=j)$ are the sample analogue estimators, given by,

$$\hat{P}(\pi=j) = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} 1(\pi^i = 1)$$

$$\hat{F}_X(x, O(j)\,|\pi=j) = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} 1\left(X^i(t) \le x(t), \forall t \in O(j)\right) * \\ *1(\pi^i = 1) \end{pmatrix} \left(\hat{P}(\pi=j)\right)^{-1}$$

and, by convention, $\hat{F}_X(x|\pi=1) = \hat{F}_X(x, O(1)\,|\pi=1)$.

(3) Estimate the identified set for the cumulative distribution function of the data, denoted by $\hat{\mathcal{H}}(F_X)$, and given by,

$$\hat{\mathcal{H}}(F_X) = \left\{\Gamma \cap \left\{G : \hat{F}_X^L \le G \le \hat{F}_X^H\right\}\right\}$$

(4) Estimate the sharp worst case scenario lower and upper bounds, given by,

$$\begin{aligned} \hat{T}_L\left(X, Y_{\hat{\theta}_0}\right) &= \inf_{G \in \hat{\mathcal{H}}(F_X)} \tfrac{1}{V}\sum_{j=1}^{V}\left(G(Z_j) - \hat{F}_Y\left(Z_j|\hat{\theta}_0\right)\right)^2 \\ \hat{T}_H\left(X, Y_{\hat{\theta}_0}\right) &= \sup_{G \in \hat{\mathcal{H}}(F_X)} \tfrac{1}{V}\sum_{j=1}^{V}\left(G(Z_j) - \hat{F}_Y\left(Z_j|\hat{\theta}_0\right)\right)^2 \end{aligned}$$

where $\{Z_j\}_{j=1}^{V}$ is a random sample of μ.

(5) For $s = 1, 2, ..., S$, repeat the following two steps,

(a) Construct a bootstrap sample of size n of $Y_{\hat{\theta}_0}$, and denote it by $\mathcal{X}_n^*$. Estimate the parameter from the bootstrap sample, denoting the estimate by $\hat{\theta}_0^*$.

(b) Compute the simulated test statistic, denoted $\hat{T}\left(X^*, Y_{\hat{\theta}_0^*}\right)$.

(6) Denote by $t_{\hat{\theta}_0}^*(1-\alpha)$ the $(1-\alpha)$ quantile of the simulated values of $n\hat{T}\left(X^*, Y_{\hat{\theta}_0^*}\right)$.

(7) Decide the outcome of the test in the following way,

Outcome	Decision
$t_{\hat{\theta}_0}^*(1-\alpha) < n\hat{T}_L\left(X, Y_{\hat{\theta}_0}\right) \leq n\hat{T}_H\left(X, Y_{\hat{\theta}_0}\right)$	Reject H_0
$n\hat{T}_L\left(X, Y_{\hat{\theta}_0}\right) \leq n\hat{T}_H\left(X, Y_{\hat{\theta}_0}\right) \leq t_{\hat{\theta}_0}^*(1-\alpha)$	Do not reject H_0
$n\hat{T}_H\left(X, Y_{\hat{\theta}_0}\right) \leq t_{\hat{\theta}_0}^*(1-\alpha) < n\hat{T}_L\left(X, Y_{\hat{\theta}_0}\right)$	Inconclusive

Notice how the existence of a root-n-consistent estimator of θ_0 allows us to simulate the distribution of the test statistic under the null even under the presence of missing data.

Recall from section 2.3.2 that missing data opens a gap between the population lower and upper worst case scenario bound, reflecting our ignorance about the missing data. The gap in the population test statistic gets mapped into a gap of its sample analogue and, as a consequence, the hypothesis test has an inconclusive region. This is an undesired but unavoidable consequence of having missing data and imposing no assumptions regarding their distribution.

As we argued in section 2.3.2, the population sharp worst case scenario bounds are hard to compute since they require solving an optimization problem in a functional space. The estimation of the bounds is obtained by Monte Carlo integration and hence, instead of solving an optimization problem in functional spaces, we need to solve a finite dimensional (but possibly large) optimization problem. This is shown in the following lemma.

Lemma 28. *Let* $\{Z_j\}_{j=1}^{V}$ *denote the random of* μ *in the fourth step of our procedure. Consider the consider the following set,*

$$\hat{\mathcal{S}} = \left\{ \begin{array}{c} s \in \mathbb{R}^V : \forall j \in \{1,2,...V\}\,,\ \hat{F}_X^L(Z_j) \leq s_j \leq \hat{F}_X^H(Z_j) \\ \forall j,k \in \{1,2,...V\}\,,\ \textit{if } Z_j(t) \leq Z_k(t)\ \ \forall t \in \mathcal{D} \Longrightarrow s_j \leq s_k \end{array} \right\}$$

Then,

$$\begin{array}{rcl} \hat{T}_L\left(X, Y_{\hat{\theta}_0}\right) & = & \min\limits_{g \in \hat{\mathcal{S}}} \frac{1}{V} \sum\limits_{j=1}^{V} \left(g_j - \hat{F}_Y\left(Z_j|\hat{\theta}_0\right)\right)^2 \\ \hat{T}_H\left(X, Y_{\hat{\theta}_0}\right) & = & \max\limits_{g \in \hat{\mathcal{S}}} \frac{1}{V} \sum\limits_{j=1}^{V} \left(g_j - \hat{F}_Y\left(Z_j|\hat{\theta}_0\right)\right)^2 \end{array}$$

The previous lemma shows that computing the estimate of the lower and upper sharp worst scenario bounds amounts to solving a V-dimensional optimization problem subject to boundaries and monotonicity constraints. The optimization problem in lemma 28 has a quadratic objective function and linear inequality constrains and the challenge in solving it lies entirely in the dimension of the problem, which grows with V, the quality of our Monte Carlo approximation[11].

Even though the optimization problem described in lemma 28 can be solved numerically, its implementation is slow and exposes the researcher to obtaining incorrect solutions if he computes local optima with global optima. In order to avoid these complications,

---

[11] This problem can be solved by MATLAB's FMINCON function. As starting values of the optimization problem, we suggest the use the estimate of the alternative worst case scenario bounds, described in lemma 29.

we might consider estimating the alternative worst case scenario bounds, defined in theorem 22. In order to implement a hypothesis test based on these bounds, we replace the estimate of the sharp identified set for the cumulative distribution function of the data in step 3 with the estimate of its strict superset, given by,

$$\hat{\mathcal{H}}'(F_X) = \left\{G : \hat{F}_X^L \leq G \leq \hat{F}_X^H\right\}$$

The rest of the test proceeds exactly as before. In particular, in step 4, we will obtain the following estimate of the alternative worst case scenario bounds for the test statistic,

$$\begin{aligned} \hat{T}_L\left(X, Y_{\hat{\theta}_0}\right) &= \inf_{G \in \hat{\mathcal{H}}'(F_X)} \tfrac{1}{V} \sum_{j=1}^{V} \left(G(Z_j) - \hat{F}_Y\left(Z_j|\hat{\theta}_0\right)\right)^2 \\ \hat{T}_H\left(X, Y_{\hat{\theta}_0}\right) &= \sup_{G \in \hat{\mathcal{H}}'(F_X)} \tfrac{1}{V} \sum_{j=1}^{V} \left(G(Z_j) - \hat{F}_Y\left(Z_j|\hat{\theta}_0\right)\right)^2 \end{aligned}$$

The following lemma provides an explicit formula for these bounds.

Lemma 29. *Let* $\{Z_j\}_{j=1}^V$ *denote the random of* μ *in the fourth step of our procedure. The estimates for the alternative worst case scenario bounds are given by,*

$$\begin{aligned} \hat{T}'_L\left(X, Y_{\hat{\theta}_0}\right) &= \tfrac{1}{V} \sum_{j=1}^{V} \left\{ \begin{array}{c} 1\left[\hat{F}_Y\left(Z_j|\hat{\theta}_0\right) < \hat{F}_X^L(Z_j)\right]\left(\hat{F}_X^L(Z_j) - \hat{F}_Y\left(Z_j|\hat{\theta}_0\right)\right)^2 + \\ +1\left[\hat{F}_Y\left(Z_j|\hat{\theta}_0\right) > \hat{F}_X^H(Z_j)\right]\left(\hat{F}_X^H(Z_j) - \hat{F}_Y\left(Z_j|\hat{\theta}_0\right)\right)^2 \end{array} \right\} \\ \hat{T}'_H\left(X, Y_{\hat{\theta}_0}\right) &= \tfrac{1}{V} \sum_{j=1}^{V} \max\left\{\left(\hat{F}_X^L(Z_j) - \hat{F}_Y\left(Z_j|\hat{\theta}_0\right)\right)^2, \left(\hat{F}_X^H(Z_j) - \hat{F}_Y\left(Z_j|\hat{\theta}_0\right)\right)^2\right\} \end{aligned}$$

Recall that theorem 23 indicated that, at the population level, the sharp worst case scenario lower bound coincided with the alternative worst case scenario lower bound and the sharp worst case scenario upper bound was smaller or equal to the alternative worst case scenario upper bound. The following result shows the same relationship holds for the estimates of these bounds.

Theorem 30. *The estimate of the alternative worst case scenario lower bound coincides with the estimate of the sharp worst case scenario lower bound,*

$$\hat{T}_L' \left(X, Y_{\hat{\theta}_0} \right) = \hat{T}_L \left(X, Y_{\hat{\theta}_0} \right)$$

The estimate of the alternative worst case scenario upper bound is greater or equal than estimate of the sharp worst case scenario upper bound,

$$\hat{T}_H' \left(X, Y_{\hat{\theta}_0} \right) \geq \hat{T}_H \left(X, Y_{\hat{\theta}_0} \right)$$

This weak inequality might or might not be strict.

Based on the previous result, the following corollary follows.

Corollary 31. *The following results are true,*

1. *The test based on the estimate of the sharp worst case scenario bounds rejects if and only if the test based on the alternative worst case scenario bound rejects.*
2. *If the test based on the estimate of the alternative worst case scenario bounds does not reject then the test based on the sharp worst case scenario bounds will also not reject.*
3. *If the test based on the estimate of the of the sharp worst case scenario bounds is inconclusive then the test based on the alternative worst case scenario bounds will also be inconclusive.*

2.4.3. Properties of the test

As in any other hypothesis test, the properties of the hypothesis test depend on whether the null hypothesis is true or false (that is, whether $T\left(X, Y_{\theta_0}\right)$ is zero or positive). In the presence of missing data, the true value of the test statistic also depends on whether the (population) worst case scenario bounds we use are zero or positive. The following table describes all the possibilities,

	H_0 is true	H_0 is false
$T_L\left(X, Y_{\theta_0}\right) = 0$, $T_H\left(X, Y_{\theta_0}\right) = 0$	case 1	impossible
$T_L\left(X, Y_{\theta_0}\right) = 0$, $T_H\left(X, Y_{\theta_0}\right) > 0$	case 2	case 3
$T_L\left(X, Y_{\theta_0}\right) > 0$, $T_H\left(X, Y_{\theta_0}\right) > 0$	impossible	case 4

The columns of the table represent the unknown truth that we are interested in learning and the rows represent the truth that can be identified from the population affected by the missing data. The first row (case 1) corresponds to the case when the null hypothesis is true and this can be learnt from the observed population. The last row (case 4) corresponds to the case when the null hypothesis is false and this can be learnt from the population. Finally, the middle row (cases 2 and 3) represents the situation where we cannot decide if the null hypothesis is true or not, even if we knew the data generating process of the observed data.

The table implicitly assumes that we are using the sharp worst case scenario bounds given in theorem 21 (hence, the classification of tows based on $T_L\left(X, Y_{\theta_0}\right)$ and $T_H\left(X, Y_{\theta_0}\right)$).

If other worst case scenario bounds are utilized[12], then we will still have the four cases described by the table, except that the relevant population version for the worst case scenario bounds will be replaced by these other bounds.

The first two theorems refer to the behavior of the test under the null hypothesis, which occurs in cases 1 and 2.

Theorem 32. *Let assumptions (A1)-(A4) hold and suppose that the null hypothesis is true. Then,*

$$\limsup_{n\to\infty} P\left(t^*_{\hat{\theta}_0}(1-\alpha) < n\hat{T}_L\left(X, Y_{\hat{\theta}_0}\right) \leq n\hat{T}_H\left(X, Y_{\hat{\theta}_0}\right)\right) \leq \alpha$$

Theorem 32 implies that the level of the test is correct but it may result in conservative inference. Recall from section 2.2 that when there is no missing data, the upper and lower bounds collapse and coincide with the test statistic in BHHN [**19**], which results in a non-conservative hypothesis test. Hence, our hypothesis test is conservative solely due to the presence of missing data.

Before stating further results about the hypothesis test, we establish the following intermediate lemma, which is the key to the subsequent results.

Lemma 33. *Let assumptions 1-4 hold. Then, the estimates of the sharp (alternative, restricted) worst case scenario bounds are consistent for the population sharp (alternative, restricted) worst case scenario bounds.*

---

[12] For example, we could use the alternative worst case scenario bounds derived in theorem 22 or, if additional information about the missing data is available, we could use the restricted worst case scenario bounds derived in theorem 25.

Based on this lemma, we can establish two additional conclusions. The first one is an undersirable feature of our hypothesis test.

Theorem 34. *Let assumptions (A1)-(A4) hold, suppose that the null hypothesis is true and* $T_H\left(X, Y_{\theta_0}\right) > 0$ *(case 2). Then,*

$$\lim_{n\to\infty} P\left(n\hat{T}_L\left(X, Y_{\hat{\theta}_0}\right) \leq n\hat{T}_H\left(X, Y_{\hat{\theta}_0}\right) \leq t^*_{\hat{\theta}_0}\left(1-\alpha\right)\right) = 0$$

Theorem 34 indicates that if the null hypothesis is true but the population worst case scenario bounds do not contain this information, then the probability of making the correct decision (not rejecting) converges to zero. The next result constitutes a desirable feature of our test.

Theorem 35. *Let assumptions (A1)-(A4) hold, suppose that the null hypothesis is false and* $T_L\left(X, Y_{\theta_0}\right) > 0$ *(case 4). Then,*

$$\lim_{n\to\infty} P\left(t^*_{\hat{\theta}_0}\left(1-\alpha\right) < n\hat{T}_L\left(X, Y_{\hat{\theta}_0}\right) \leq n\hat{T}_H\left(X, Y_{\hat{\theta}_0}\right)\right) = 1$$

Theorem 35 shows that whenever the null hypothesis is false and the worst case scenario contain this information (that is, the population upper bound is zero), then the probability of making the right decision (rejecting) converges to one. In other words, the test is consistent against fixed alternative hypothesis that can be discovered from information in the population.

In order to provide a complete characterization of the properties of the hypothesis test, we should consider the behavior of the bounds when the null hypothesis is false but this is undetectable from the worst case scenario bounds (case 3). This would require studying the properties of the estimate of the lower worst case scenario bound under the alternative hypothesis which is out of the scope of this paper.

2.4.4. Analysis when root-n consistent estimator of θ_0 is unavailable

As we described in the section 2.3, missing data might create two possible identification problems to the BHHN [**19**] test. The first problem occurs because missing data can destroy the point identification of the parameter θ_0, which would hinder the possibility of estimating it in a consistently. The second problem occurs because missing data destroys the point identification of the distribution of the data. We have argued that for certain applications a root-n consistent estimator of θ_0 might be available even under the presence of missing data. On the other hand, any observations with missing data will result in the loss of point identification of the distribution of the data. In this sense, we argued that the first identification problem might be avoidable, whereas the second problem is inevitable. Based on this distinction, we have constructed our specification test under the assumption that a root-n consistent estimator exists (assumption 2.(ii)) and we have focused our analysis on the second identification problem.

This section briefly considers how our analysis changes when missing data destroys the point identification of θ_0. Assume now that the available data restricts the parameter to a certain set $\Theta_I(\mathbf{P})$ (where $\Theta_I(\mathbf{P}) \subset \Theta$), that is, θ_0 is *partially identified* or *set identified* and $\Theta_I(\mathbf{P})$ is referred to as the *identified set*.

The identified set can be estimated consistently using recent developments by Manski and Tamer [**42**] and Chernozhukov, Hong and Tamer [**23**]. Using a worst case scenario approach, we can define the identified set by $\Theta_I(\mathbf{P}) = \{\theta \in \Theta : T_H(X, Y_\theta) \leq 0\}$ and then, the estimate of the identified set would be given by,

$$\hat{\Theta}_I = \left\{ \theta \in \Theta : n\hat{T}_H \left(X, Y_\theta \right) \leq \varepsilon_n \right\}$$

where $\{\varepsilon_n\}_{n=1}^{+\infty}$ is a sequence that converges to zero at a suitable rate.

In a conceptual manner, it is not hard to extend our analysis of section 2.3 to the case when the parameter is partially identified. In the procedure proposed in previous sections, both the worst case scenario bounds and the distribution of the test statistic depend on the parameter value. Consequently, if the parameter of interest can only be restricted to a set, we should run a hypothesis test for each value of the estimate of the identified set. If the individual test for every parameter value in the estimate of the identified set is rejected, then we reject the null hypothesis. If the individual test for every parameter value in the estimate of the identified set is not rejected, then we do not to reject the null hypothesis. In any other case, the test is inconclusive. As expected, the lack of identification of the parameter implies that the hypothesis test becomes even less informative. Moreover, if performing the hypothesis test for only one parameter value is already a computationally demanding task, doing it for a set of parameter values seems to be computationally prohibitive. The study of the properties of the test considered in this subsection is out of the scope of the current paper.

2.5. Monte Carlo simulations

In this section, we develop two Monte Carlo experiments to study the performance of our specification test. In the first experiment, observations are conjectured to be the wage paths simulated from the Burdett-Mortensen labor market model. In the second experiment, observations are conjectured to be return paths of stock prices specified by the Black-Scholes model.

2.5.1. Simulations from the Burdett-Mortensen model

For our first Monte Carlo simulations, we consider a two sector version of the Burdett-Mortensen labor market model. For a detailed description of the model, see Burdett and Mortensen [**20**], Mortensen [**47**] and Bowlus, Kiefer and Neumann [**17**].

In this model, firms can be classified into two groups: low productive firms, with productivity P_L and high productive firms, with productivity P_H. In order to produce the homogeneous good, firms need to form a match with workers. This matching process is affected by frictions: it takes time for unemployed workers and for vacant firms to discover each other and agree to produce.

We now describe the dynamics of the model from the point of view of the worker. At each point in time, workers in this economy can be employed (matched with a firm) or unemployed (unmatched). At a Poisson rate λ_0, unemployed workers receive a job offer with a wage distributed according to an endogenous offer distribution, denoted by F. In equilibrium, unemployed workers will only receive offers that are higher than their reservation wage, denoted r, and will hence be immediately accepted. Employed workers receive two types of shocks. First, at a Poisson rate λ_1, they receive a new job offer, which they will only accept if it represents an improvement to the current wage rate. Second, at a Poisson rate δ, they receive a shock that destroys their current match and leaves them immediately unemployed.

Firms choose the wage offer to maximize the profits of production. In equilibrium, firms will be indifferent between every wage in the support of the endogenous distribution F: lower wage means higher profits when the job offer is accepted but also means that the job position will remain vacant for a longer period of time.

In our simulations, an observation is the wage path of an individual over three years. For the same model simulation, we generate two sets of data. In the first data, the benchmark, there is no missing data. In the second set of data, 10% of the sample is not observed exclusively on the third period. The first year of data is observed for every individual and so the parameter of the model will be estimated exclusively based on this information. In order to implement the test we use the following test parameter values: $n = 2000$, $m = 1000$, $k = 4$, $T = 159$ (159 months, or 3 years), $S = 1000$ and $V = 200$. In order to be able to compute 1000 Monte Carlo replications for each missing data pattern and for each of the hypothesis in a fast manner, we implement our hypothesis test using alternative worst case scenario bounds.

2.5.1.1. Simulations under the null. The parameter values for our simulations under the null are the following: $\lambda_0 = 0.03$, $\lambda_1 = 0.01$, $\delta = 0.0035$, $r = 100$, $P_1 = 500$, $P_2 = 1000$. Moreover, We assume that half of the firms are low productivity firms and the other half are high productivity firms.

Table 2.1 describes the results of 1000 simulations. For each significance level, we compute the percentage of simulations where the test results in rejection, lack of rejection or inconclusive.

Our simulations indicate that relatively few missing observations with any sort of missing data pattern implies that the test is almost completely inconclusive.

Sample	$\alpha = 10\%$			$\alpha = 5\%$			$\alpha = 1\%$		
	Rej.	No Rej	Inc.	Rej.	No rej	Inc.	Rej.	No rej	Inc.
No MD	6.8%	93.2%	0%	2.6%	97.4%	0%	0.7%	99.3%	0%
With MD	0.5%	0%	99.5%	0.1%	0%	99.9%	0%	0%	100%

Table 2.1. Results of simulations under the null

Sample	$\alpha = 10\%$			$\alpha = 5\%$			$\alpha = 1\%$		
	Rej.	No rej	Inc.	Rej.	No rej	Inc.	Rej.	No rej	Inc.
No MD	87.1%	12.9%	0%	82.3%	17.7%	0%	67.4%	32.6%	0%
With MD	0.5%	0%	99.5%	0.4%	0%	99.6%	0%	0%	100%

Table 2.2. Results of simulations under the first alternative hypothesis

2.5.1.2. Simulations under the alternative. We consider two alternative hypothesis from the Burdett-Mortensen model. These alternative hypothesis are inspired by features present in actual labor markets that are ignored by the model.

In the first alternative hypothesis, we allow for factors other than the wage level to affect the quality of the job. In the previously described model, a job (offer) is completely characterized by the wage level. This is obviously a simplification as, in reality, jobs are described by a vector of characteristics, where only one of them being the wage. As a consequence of this simplification, employed workers will only accept a new job offer if the new wage is higher than the current wage and, hence, all job to job transitions will generate upward jumps in the wage profile. As BHHN [**19**] point out, this is not verified in the NLSY79 data, where 32% of the job to job transitions result in wage decreases. We model this by allowing that with a certain probability, denoted by ρ, an employed worker will accept to change to a job that pays less than his actual wage. Table 2.2 presents simulations for $\rho = 0.5$ and the remaining parameters fixed at the same values as in the null hypothesis.

Sample	$\alpha = 10\%$			$\alpha = 5\%$			$\alpha = 1\%$		
	Rej.	No rej	Inc.	Rej.	No rej	Inc.	Rej.	No rej	Inc.
No MD	99.6%	0.4%	0%	99.3%	0.7%	0%	98.5%	1.5%	0%
With MD	38.9%	0%	61.1%	26.5%	0%	73.5%	11.2%	0%	88.8%

Table 2.3. Results of simulations under the second alternative hypothesis

When there is missing data, the behavior of the test under this alternative hypothesis resembles the one under the null hypothesis. Relatively few percentage of missing observations can result in an inconclusive test.

In our second alternative hypothesis we introduce unobserved heterogeneity in the workforce. We allow for workers to be of two types: stable and unstable, which will differ in their transition rates. Stable workers will have transition rates determined by the following parameters: $\lambda_0 = 0.03$, $\lambda_1 = 0.01$, $\delta = 0.0035$, whereas unstable workers will have transition rates determined by the following parameters: $\lambda_0 = 0.06$, $\lambda_1 = 0.02$, $\delta = 0.007$. As a consequence, unstable workers will transition more often between employment and unemployment and from job to job than stable workers. Table 2.3 presents simulations from this alternative hypothesis.

In this case, the hypothesis test has a significant rejection rate even under the presence of missing data.

2.5.2. Simulations from the Black-Scholes model

The Black-Scholes model is the cornerstone of the option pricing finance literature. Based on a set of simple assumptions, this model delivers a closed form formula for the price of an European call or put option. One key assumption in this model is that the return to the value of the underlying asset price behaves like a Brownian motion with a non-stochastic drift and a non-stochastic volatility process. In its most simplistic version, the drift and the volatility are assumed to be constant.

If we assume that the remaining assumptions of the Black-Scholes model hold, a test of whether the returns of the underlying behave like a Brownian motion would be a specification test for the Black-Scholes model. With this motivation as background, Cuesta-Albertos, del Barrio, Fraiman and Maltrán [**25**] present Monte Carlo simulations of this specification test. We produce our simulations using their design. The return paths are distributed according the following stochastic process,

$$R(t) = \big(s_1 + s_2 t^2 + s_3 \sin(2\pi t)\big) W(t) + \big(a_1 t + a_2 t^2 + a_3 \sin(2\pi t)\big)$$

where W is a standard Brownian motion and the s's and the a's are constants. Under the null hypothesis that returns are distributed according to a Brownian motion, a_1 represents the drift, s_1 represents the volatility and s_2, s_3, a_2 and a_3 are all equal to zero. By setting different values for the s's and the a's, we generate 18 different specifications, which are described in table 2.4. Notice that only specifications 1 and 3 satisfy the null hypothesis (Brownian motion without and with drift, respectively), whereas the rest of the models are example of the alternative hypothesis.

Our simulated data represents 100 randomly selected stock firm prices, which we intend to observe continuously over two years. If our data is composed of stock return paths, missing data naturally occurs when firms go out of business and stop being traded. We (randomly) choose 10% of the sample to be missing during the second year. The parameters of the test are as follows: $n = 100$, $m = 200$, $k = 4$, $T = 2$ (two years),

Model	s_2	s_3	a_1	a_2	a_3	Formula for the return process	Hypothesis type
1	0	0	0	0	0	$W(t)$	Null
2	0	0	0	1	0	$W(t)+t^2$	Alternative
3	0	0	1	0	0	$W(t)+t$	Null
4	0	0	1	1	0	$W(t)+t+t^2$	Alternative
5	1	0	0	0	0	$(1+t^2)\,W(t)$	Alternative
6	1	0	0	1	0	$(1+t^2)\,W(t)+t^2$	Alternative
7	1	0	1	0	0	$(1+t^2)\,W(t)+t$	Alternative
8	1	0	1	1	0	$(1+t^2)\,W(t)+(t+t^2)$	Alternative
9	0	0	0	0	1	$W(t)+\sin(2\pi t)$	Alternative
10	0	0	1	0	1	$W(t)+(t+\sin(2\pi t))$	Alternative
11	1	0	0	0	1	$(1+t^2)\,W(t)+\sin(2\pi t)$	Alternative
12	1	0	1	0	1	$(1+t^2)\,W(t)+(t+\sin(2\pi t))$	Alternative
13	0	1	0	0	0	$(1+\sin(2\pi t))\,W(t)$	Alternative
14	0	1	0	1	0	$(1+\sin(2\pi t))\,W(t)+t^2$	Alternative
15	0	1	1	0	0	$(1+\sin(2\pi t))\,W(t)+t$	Alternative
16	0	1	1	1	0	$(1+\sin(2\pi t))\,W(t)+(t+t^2)$	Alternative
17	0	1	0	0	1	$(1+\sin(2\pi t))\,W(t)+\sin(2\pi t)$	Alternative
18	0	1	1	0	1	$(1+\sin(2\pi t))\,W(t)+(t+\sin(2\pi t))$	Alternative

Table 2.4. Monte Carlo designs

$S = 200$ and $V = 200$. The true parameter vector $\theta_0 = (a_1, s_1)$ is estimated by sample analogue estimation using complete sample from the first period.

For the 1000 Monte Carlo replications, we implement the test procedure using the sharp and the alternative worst case scenario bounds. The results from the hypothesis test based on the sharp worst case scenario bounds are presented in table 2.5.

The results of these simulations show how our specification test can produce informative results even if we use the worst case scenario approach in the presence of missing functional data.

Under the null hypothesis (models 1 and 3) the test has rejection rates that are lower than the significance levels, as expected from theorem 32. Also, as expected, the percentage of tests that are not rejected (inconclusive) increase (decrease) as the significance

Model	$\alpha = 10\%$			$\alpha = 5\%$			$\alpha = 1\%$		
	Rej.	Not rej.	Inc.	Rej.	Not rej.	Inc.	Rej.	Not rej.	Inc.
1	0.8%	17.9%	81.3%	0%	42.7%	57.3%	0%	82.2%	17.8%
2	16.6%	1.9%	81.5%	7.5%	4.5%	88.0%	1.6%	26.5%	71.9%
3	0.4%	14.4%	85.2%	0%	38.7%	61.3%	0%	79.2%	20.8%
4	11.3%	0.4%	88.3%	4.3%	2.9%	92.8%	0.6%	29.6%	70.8%
5	33.2%	0.2%	66.6%	18.3%	2.0%	79.7%	4.6%	12/8%	82.6%
6	42.7%	0.1%	57.2%	25.4%	1.0%	73.6%	5.0%	11.0%	84.0%
7	36.6%	0%	63.4%	21.3%	0.9%	77.8%	5.1%	10.2%	84.7%
8	30.7%	0.3%	69%	15.2%	1.0%	83.8%	2.5%	9.4%	88.1%
9	99.7%	0%	0.3%	99.0%	0%	1.0%	96.1%	0%	3.9%
10	100%	0%	0%	100%	0%	0%	99.7%	0%	0.3%
11	99.7%	0%	0.3%	98.5%	0%	1.5%	93.3%	0%	6.7%
12	100%	0%	0%	99.9%	0%	0.1%	99.5%	0%	0.5%
13	87.7%	0.3%	12.3%	75.5%	0.7%	23.8%	43.7%	7.3%	49.0%
14	91.4%	0%	8.6%	81.8%	0.1%	18.1%	56.2%	3.3%	40.5%
15	87.7%	0.1%	12.2%	75.7%	0.5%	23.8%	41.3%	4.1%	54.6%
16	92.1%	0.1%	7.8%	79.8%	0.4%	19.8%	49.6%	3.1%	47.3%
17	99.8%	0%	0.2%	99.7%	0%	0.3%	98.7%	0%	1.3%
18	100%	0%	0%	100%	0%	0%	99.9%	0%	0.1%

Table 2.5. Results of simulations using the sharp worst case scenario bounds

level decreases. Under certain versions of the alternative hypothesis, the test presents relatively high rejection rates, especially when the stochastic process includes sinusoidal trends or volatilities.

Table 2.6 presents the results of performing the hypothesis test based on the alternative worst case scenario bounds for the same simulations.

As explained in corollary 31, using the alternative worst case scenario bounds leads to equal percentage of rejected simulations, lower percentage of non-rejected simulations and, consequently, higher percentage of inconclusive simulations than the ones obtained from using sharp worst case scenario bounds. In this particular model, the difference between the alternative and the sharp worst case scenario upper bounds is very small. Given that

Model	$\alpha = 10\%$			$\alpha = 5\%$			$\alpha = 1\%$		
	Rej.	Not rej.	Inc.	Rej.	Not rej.	Inc.	Rej.	Not rej.	Inc.
1	0.8%	17.8%	81.4%	0%	42.5%	57.5%	0%	82.2%	17.8%
2	16.6%	1.9%	81.5%	7.5%	4.4%	88.1%	1.6%	26.3%	72.1%
3	0.4%	14.4%	85.2%	0%	38.6%	61.4%	0%	79.2%	20.8%
4	11.3%	0.4%	88.3%	4.3%	2.8%	92.9%	0.6%	27.8%	71.6%
5	33.2%	0.2%	66.6%	18.3%	2.0%	79.7%	4.6%	12.8%	82.6%
6	42.7%	0.1%	57.2%	25.4%	1.0%	73.6%	5.0%	10.9%	84.1%
7	36.6%	0%	63.4%	21.3%	0.9%	77.8%	5.1%	10.2%	84.7%
8	30.7%	0.3%	69%	15.2%	1.0%	83.8%	2.5%	9.4%	88.1%
9	99.7%	0%	0.3%	99.0%	0%	1.0%	96.1%	0%	3.9%
10	100%	0%	0%	100%	0%	0%	99.7%	0%	0.3%
11	99.7%	0%	0.3%	98.5%	0%	1.5%	93.3%	0%	6.7%
12	100%	0%	0%	99.9%	0%	0.1%	99.5%	0%	0.5%
13	87.7%	0.3%	12.3%	75.5%	0.7%	23.8%	43.7%	7.3%	49.0%
14	91.4%	0%	8.6%	81.8%	0.1%	18.1%	56.2%	3.3%	40.5%
15	87.7%	0.1%	12.2%	75.7%	0.5%	23.8%	41.3%	4.0%	54.7%
16	92.1%	0.1%	7.8%	79.8%	0.4%	19.8%	49.6%	3.0%	47.4%
17	99.8%	0%	0.2%	99.7%	0%	0.3%	98.7%	0%	1.3%
18	100%	0%	0%	100%	0%	0%	99.9%	0%	0.1%

Table 2.6. Results of simulations using the alternative worst case scenario bounds

the alternative worst case scenario bounds are much easier and faster to compute, these results encourage the econometrician to use the alternative worst case scenario bounds.

2.6. Empirical Illustration

In this section, we use the test developed in this paper to test whether the observations of wage processes from the NLSY79 are distributed according to the Burdett-Mortensen model described in section 2.5.1.

2.6.1. Description of the data

Our data are composed of young individuals (ages 17 to 22, in our sample), first interviewed in 1979, who are re-interviewed in subsequent years. In each interview year, each individual is asked about their job spells that occurred since the last interview. The first job spell reported in an interview corresponds to the main job spell (called the current/most recent job spell) but the interview process allows up to 5 job spells between interviews. For each job spell, the individual reports the start week and the stop week of the job spell as well as its wage rate. With this information, we can construct the wage path for each individual from 1982[13] until 1991. We express all wages in terms of weekly remuneration and in terms of 1990 U.S. dollars using the Consumer Price Index[14].

The Burdett-Mortensen model assumes that workers in the economy are ex-ante homogeneous. Even though our sample contains very heterogeneous group of individuals, we hope that we can condition on observable characteristics to obtain an homogeneous sample. Following Bowlus, Kiefer and Neumann [**17**], we restriction attention to white males that are High school or GED graduates and who are not in the military sample. This constitutes a sample of 816 individuals. We eliminate from the sample individuals who, at any point in the survey, presented problems in their duration data[15] or reported having weekly wages of over a thousand 1990 U.S. dollars[16]. This reduces our representative sample to 589 individuals. Finally, in order to estimate the parameter consistently, we require the relevant sample to be completely observed over a certain period of time. Hence, we eliminate all observations that have any kind of missing data during 1982, which represents only 53 individuals or less than 10% of the sample. This produces the sample which we use for our hypothesis test, composed of 536 individuals.

---

[13] Even though we have data since 1979, we avoid using the first three years of the sample since, during those years, some of the individuals of the sample were less than 20 years of age and their job market opportunities could be different from their older counterparts.

[14] This information is publicly available in the U.S. Bureau of Labor Statistics website.

[15] These problems are either negative job spell duration or missing time information.

[16] This type of trimming is also utilized by Bowlus, Kiefer and Neumann [**17**].

2.6.2. Missing data

Our sample is mildly affected by missing wage information. Of a total of 536 individuals, 433 individuals (80.7%) have no missing wage information and 103 individuals (19.3%) have some episode of missing wage information. Moreover, only 6.07% of all the weeks in the sample are missing.

From the 103 individuals with some missing data, 58 of them (56.3%) suffer from attrition from the sample, that is, the individuals are lost at some point and remain unobserved for the rest of the sample. From the remaining 45 individuals, there are very few episodes that violate attrition. These figures indicate that sample attrition is a common explanation for missing observations in the NLSY79 survey.

Figure 2.1 presents the evolution of the percentage of individuals with missing data. The percentage of missing data is (almost) weakly increasing in time. Again, this is also indicative that most of the missing data is caused by sample attrition. For those individuals who have missing information, the average number of missing weeks is 166.4, which represents 31.6% of the weeks in our sample. The histogram of the number of missing weeks for this subset of individuals is shown in figure 2.2.

Even thought there are certain individuals with a huge number of missing information, most of the individuals in the sample have relatively few missing observations.

2.6.3. Test results

We now describe the result of testing whether the Burdett-Mortensen model is the right specification for the wage processes in the NLSY79 survey. After discarding individuals

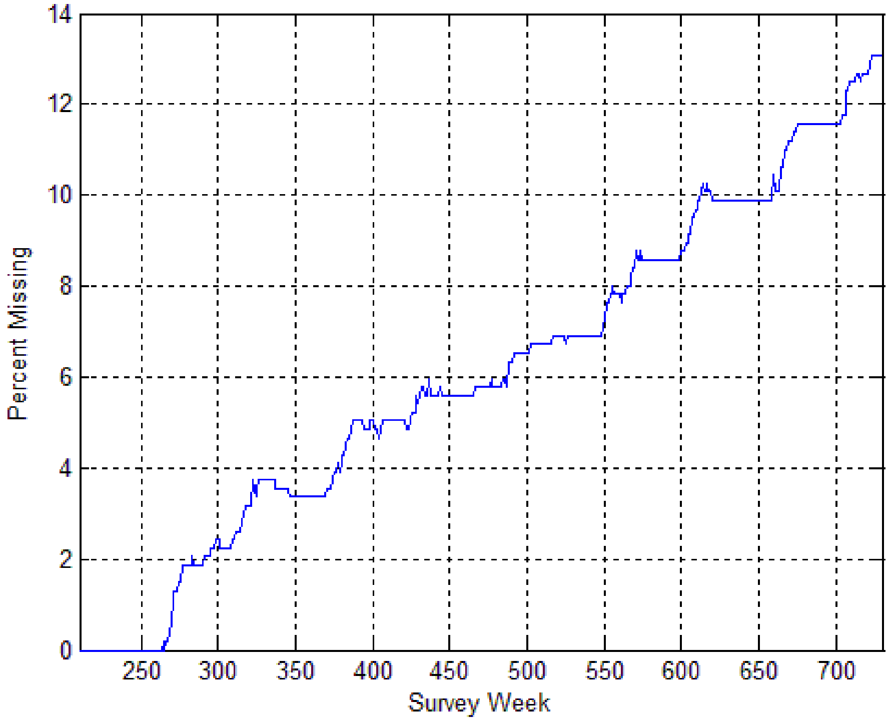

Figure 2.1. Percentage of individuals with missing data

with any missing data, BHHN [**19**] strongly reject the null hypothesis that the four sector Burdett-Mortensen model is the right specification for the data. We implement the specification test for a one, two, three, four and five sector Burdett-Mortensen model.

Table 2.7 presents the estimated sharp worst case scenario bounds for the test statistic, as well as the 90th, 95th and 99th quantiles of the statistic under the null hypothesis.

Our specification test strongly rejects each of the specifications of the Burdett-Mortensen model[17]. In other words, the information contained in the sample with missing data is

---

[17] The table presents the outcome of one run of the hypothesis test. Since the implementation of the test includes random components, the test was repeated several times. All of these repetitions resulted in the rejection of the null hypothesis.

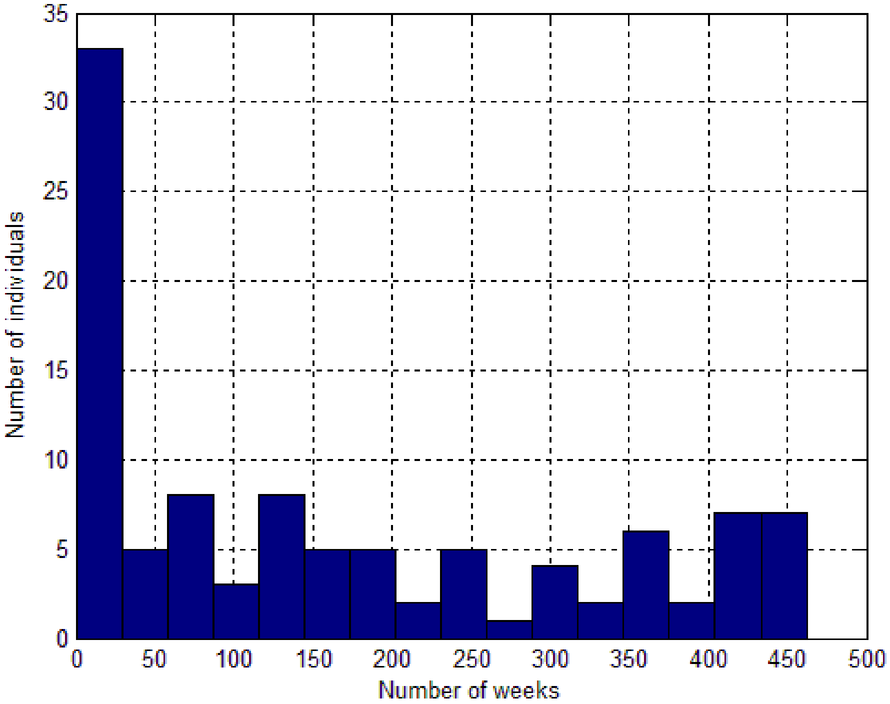

Figure 2.2. Number missing weeks for individuals with missing data

Model	Sharp WCSB		Quantiles under H_0		
	$n\hat{T}_L(X, Y_{\hat{\theta}_0})$	$n\hat{T}_H(X, Y_{\hat{\theta}_0})$	$t^*_{\hat{\theta}_0}(90\%)$	$t^*_{\hat{\theta}_0}(95\%)$	$t^*_{\hat{\theta}_0}(99\%)$
One sector	68.53	119.08	0.15	0.17	0.23
Two sectors	33.68	69.53	0.68	0.93	1.85
Three sectors	52.66	97.04	0.47	0.63	0.86
Four sectors	37.41	72.54	0.39	0.48	0.83
Five sectors	36.78	72.48	0.40	0.56	1.84

Table 2.7. Results of test on NLSY79 data using sharp worst case scenario bounds

sufficient to reject the model without making any assumptions about the nature of the missing observations.

For the sake of comparison, we also implement the specification test using the alternative worst case scenario bounds for the test statistic.

Model	Alternative WCSB		Quantiles under H_0		
	$n\hat{T}_L(X, Y_{\hat{\theta}_0})$	$n\hat{T}_H(X, Y_{\hat{\theta}_0})$	$t^*_{\hat{\theta}_0}(90\%)$	$t^*_{\hat{\theta}_0}(95\%)$	$t^*_{\hat{\theta}_0}(99\%)$
One sector	68.53	119.20	0.15	0.17	0.23
Two sectors	33.68	69.64	0.68	0.93	1.85
Three sectors	52.66	97.04	0.47	0.63	0.86
Four sectors	37.41	72.59	0.39	0.48	0.83
Five sectors	36.78	72.68	0.40	0.56	1.84

Table 2.8. Results of test on NLSY79 data using alternative worst case scenario bounds

As shown by theorem 30, the only difference between the two sets of bounds is that the alternative worst case scenario upper bounds are higher or equal than the sharp worst case scenario bounds. Table 2.8 reveals that the difference between the upper bounds is very small. As expected from 31, the specification test using the alternative worst case scenario bounds also rejects each of the specifications of the Burdett-Mortensen model.

2.7. Conclusion

This paper develops a specification test for functional data that allows for the presence of missing observations. In order to deal with the missing data problem, we adopt a worst case scenario approach which is agnostic about the distribution of the missing data. The specification test adapts the Cramér-von Mises specification test developed in Bugni, Hall, Horowitz and Neumann [**19**] to missing data.

In order to develop the specification test, we study the identification problem caused by missing observations. We show how missing data implies that the distribution of the functional data is partially identified and derive the sharp worst case scenario bounds for the distribution of the Cramér-von Mises statistic proposed by Bugni, Hall, Horowitz and Neumann [**19**]. We use the analogue principle to estimate these bounds and to implement a specification test. Our specification test can have thee outcomes: rejection of the null hypothesis, lack of rejection of the null hypothesis or inconclusive. The possibility of an inconclusive result is an undesired but unavoidable consequence of the existence of missing data and our unwillingness to impose assumptions regarding its distribution.

The theoretical properties of our specification test depend not only on whether the null hypothesis is true or false, but also on whether this can be learnt from the distribution of observed data. Under the null hypothesis, our specification test will reject the null hypothesis with a probability that, in the limit, does not exceed the significance level of the test. Unfortunately, the presence of missing data implies that the rejection rate may be conservative. Under the alternative hypothesis, the behavior of the test depends critically on whether this can be learnt from the distribution of the observed data. Whenever the distribution of the observed data contains enough information to reveal that the null hypothesis is false, our hypothesis test is consistent, that is, the power of the hypothesis test that converges to one.

The Monte Carlo evidence reveals that the behavior of the test depends strongly on the type of economic model and the type of hypothesis that is being considered. In certain situations, a small amount of missing data is enough to render our specification test completely uninformative but, in other situations, the test presents informative results.

As an empirical illustration, we test whether observations of the wage process in the NLSY79 are distributed according to the Burdett-Mortensen labor market model. In the 1982 - 1991 period, 19.3% of the individuals in the survey are affected by some form of missing data, typically caused by sample attrition. Even under the presence of missing data, our specification test strongly rejects that the Burdett-Mortensen model is the correct framework for the NLSY79 data. This illustration constitutes an ideal application of our specification test, since it delivers informative results even though we have missing data and we adopt a worst case scenario approach about the nature of the missing observations.

CHAPTER 3

Child Labor Legislation: Effective, Benign, Both or Neither?

3.1. Introduction

In 2002, the International Labor Organization's Statistical Information and Monitoring Program on Child Labor [**50**] estimated that 211 million children, or 18% of the children ages 5 to 14 in the world were economically active[1]. According to Edmonds and Pavcnik [**27**], the majority of these children lived in low income countries and only 2% lived in what we refer to as developed countries. These figures reveal that a working child in contemporary U.S. would also be extremely unusual. This has not always been the case. Until the end of the nineteenth century, child labor was both common and legal in developed economies. According to U.S. census data from 1880 (see Carter and Sutch [**22**]), 32% of boys and girls ages 10 to 15 declared having a gainful occupation. This rate fell significantly between 1880 and 1930: according to 1930 census data, the employment rate for children ages 10 to 15 was only 2%.

Among the reasons of such phenomenal decline, one can mention the growing opposition to child labor which ultimately materialized into a body of legislation restricting employers from hiring children[2]. According to Moehling [**45**], Moehling [**46**] and Basu [**7**], various degrees of resistance against child labor had always existed in the U.S., but

---

[1] A child is economically active if he or she works for wages (cash or in-kind), works in the family farm in the production and processing of primary products for the market, barter or own consumption, or is unemployed and looking for these types of work.

[2] For a description of the evolution of the legislation body against child labor, see Ogburn [**49**] and Moehling [**46**].

this opposition developed into a well-organized social movement in the 1880s and 1890s. Between 1880 and 1920, this movement was successful in enacting state-wide child labor legislation in many U.S. states. Typically, these laws took the form of state-wide prohibition for children of less than a certain age (typically, 14 years old) to be employed in the manufacturing sector. By 1910, child labor activists realized that employers had influence over certain state legislatures which limited the progress that could be made at the state level. Therefore, they decided to shift lobbying efforts from a state to a federal child labor legislation. After several unsuccessful attempts, the Fair Labor Standards Act was enacted in 1938. This is the federal law that currently prohibits employment of minors in occupations considered oppressive.

The objective of this paper is to characterize the effect caused by child labor legislation on child labor participation during the period 1880-1900. This issue is not of exclusive interest to economic historians: as the I.L.O. figures reveal, child labor is still a problem in certain parts of the world.

Existing literature has focused only on studying the effectiveness of the legislation, that is, whether the legislation managed to reduce child labor participation or not. This paper will revisit some of these results focusing on certain methodological criticisms. Moreover, we will also focus on what the labor market mechanisms by which the child labor legislation affected child labor are. By taking these into account, we may be able to establish if the legislation constituted a benign policy or not, that is, whether the legislation imposed constrains to the behavior of children (not benign) or whether it generated a change in the labor market equilibrium (benign). We argue that this novel analysis can help provide a new perspective on previous results.

It is not obvious that child labor legislation reduced child labor. Existing literature, most notably, Nardinelli [**48**] for the U.K. and Moehling [**46**] for the U.S., explain that the passing of such legislation could be followed by a reduction on child labor demand generated by external factors (e.g. change in technology or inflow of immigrants).

The effectiveness of the law in curtailing child labor during this period has been previously studied in the literature, most notably by Moehling [**45**] and Moehling [**46**]. In her dissertation, Moehling [**45**] uses a difference-in-differences estimation procedure to estimate the effect of child labor legislation using exclusively 1900 U.S. census data. She estimates a binary choice model and computes the difference in the labor market participation of younger and older 14-year-olds (group difference), between states that did and did not issue child labor legislation (spatial difference). Her estimation reveals that child labor laws imposed constraints on children participation in the labor market. Moehling [**46**] incorporates observations from the 1880 and 1910 U.S. census to study the same problem. The new dataset allows her to use a difference-in-difference-in-differences estimator to evaluate the effectiveness of the legislation. She computes the difference in labor market participation between 13 and 14-year-olds (group difference), between 1900 and 1880 (and also 1910 and 1900) (time difference) and between states that did and did not issue child labor legislation (spatial difference). Her conclusion is that child labor laws were ineffective in reducing child labor. Moehling states: “Although the predicted probabilities for the treatment group–13-year-old boys living in the states that enacted the age minima of 14– fell substantially between 1880 and 1900, so too did the predicted probabilities for the control groups”.

Even though Moehling [**45**] and Moehling [**46**] provide a very detailed study of this problem, we believe there are two problems with their differencing estimation procedure. The first issue is that in non-linear models, like the ones required to model binary explanatory variables such as (child) labor participation, differencing estimator procedures do not identify the object of interest. The second issue is that differencing estimators assume that there is only one labor market equilibrium at the end of the nineteenth century. In the presence of multiplicity of equilibria, such as the one described by Basu and Van [**8**], differencing estimators may underestimate the effect of the legislation.

Other papers in the literature have studied the determinants for child labor market participation and their relationship with child labor legislation. Sanderson [**59**] uses a cross section of data to compare employment rates between states with and without child labor legislation. This data will be affected by state fixed effects, which one can control for with panel data. Based on anecdotal evidence, Osterman [**51**] provides a detailed description of changes in the unskilled labor market (which includes child labor) at the end of the nineteenth century. Brown, Christiansen and Philips [**18**] study how changes in economic conditions and in the legislation impacted child labor in the U.S. fruit and vegetable canning industry. Goldin [**31**] studies the determinants of child labor using 1880 Philadelphia census data. Margo and Finegan [**43**] examine the effect of compulsory schooling laws and child labor laws on school attendance.

The rest of the paper proceeds as follows. In section 3.2, we discuss the inadequacy of differencing methods in identifying the effect of the legislation on child labor. Section 3.3 develops a simple but formal model to analyze the effect of the legislation. Section 3.4

defines the econometric procedure for estimation and inference and section 3.5 presents the results. Section 3.6 concludes.

3.2. Discussion

Our objective is to study the effect of the U.S. state-wide child labor legislation on the behavior of the working children at the end of the nineteenth century. By 1880, arguably none of the U.S. states had established any serious body of legislation, and by 1900, a significant subset of the U.S. states had already established state-wide prohibition for children to be employed in the manufacturing sector. If child labor legislation is considered an exogenous event, we can analyze this situation using the *natural experiment framework*[3]. In the jargon of this literature, the effect of the legislation on child labor is called *treatment effect*, children in states where the legislation was imposed are the *treatment group*, children in states where the legislation was not imposed are the *control group*, 1880 is a *pre-treatment* year and 1900 is a *post-treatment* year.

3.2.1. Differencing in non-linear models

Moehling [**45**] and Moehling [**46**] use differencing estimation techniques to estimate the treatment effect of the child labor legislation on child labor participation. In this section, we argue that this estimation method will not identify the treatment effect, precisely because the dependent variable of interest is non-linear.

Consider the following setup. There are two periods: period 1 and period 2. During period 1, no state had issued child labor legislation and, by period 2, some states had issued child labor legislation. We refer to those states that had such laws by period 2 as B

---

[3] For a rigorous treatment of these issues, see Meyer [**44**] or Woodridge [**64**].

states (treatment group) and we refer to the remaining states as A states (control group).

	Period 1	Period 2
A states	No C.L.L.	No C.L.L.
B states	No C.L.L.	C.L.L.

It is natural to allow for time fixed effects and state fixed effects to affect children employment. Time fixed effects are time-specific factors affecting all the states and state fixed effects are state-specific factors affecting each state in both periods. In order to identify the treatment effect of the legislation, we assume that the legislation is the only factor affecting exclusively B states in the second period.

The household's decision of sending a child to work is modeled with a binary response model. Denote by w the binary variable of interest that takes value of one if the child is employed and zero otherwise. Denote by $d2$ the binary variable that takes value of one if the observation corresponds to the second period and zero if it corresponds to the first period. Denote by dB the binary variable that takes the value of one if the observation corresponds to any of the B states and zero if it corresponds to any of the A states. Naturally, the interaction of these two variables is given by $d2dB$. Finally, denote by x the vector of the remaining variables that affect the decision. The structure of the binary response model is,

$$w = \begin{cases} 1 & \text{if} \quad \alpha_1 d2 + \alpha_2 dB + \alpha_3 d2dB + \beta x \geq \varepsilon \\ 0 & \text{if} \quad \alpha_1 d2 + \alpha_2 dB + \alpha_3 d2dB + \beta x < \varepsilon \end{cases}$$

where ε denotes an unobserved random term with a known continuous distribution, whose cumulative distribution function is denoted by F. From this model, we deduce the following equation,

$$P\left(w=1|d2,dB,d2dB,x\right)=\mathbb{E}\left(w|d2,dB,d2dB,x\right)=F\left(\alpha_1 d2+\alpha_2 dB+\alpha_3 d2dB+\beta x\right)$$

The object of interest, which we will refer as “treatment effect”, is the change in the probability of employment caused by issuing child labor legislation while keeping state effects, time effects and controls constant. By assuming that the child labor legislation is the only factor affecting exclusively B states in the second period, the effect of child labor legislation can be represented by going from $d2dB=0$ to $d2dB=1$, while keeping $d2$, dB and x constant. Formally, the treatment effect is given by,

$$TE\left(d\bar{2},d\bar{B},\bar{x}\right)=P\left(w=1|d\bar{2},d\bar{B},d2dB=1,\bar{x}\right)-P\left(w=1|d\bar{2},d\bar{B},d2dB=0,\bar{x}\right)$$

where $d\bar{2}, d\bar{B}$ and $\bar{x}$ are the relevant values that are used to evaluate the treatment effect.

When the model is linear, i.e., when F is the identity function, we deduce the following conclusions about the treatment effect,

(1) The treatment effect is constant and coincides with α_3, the coefficient of the interaction term,

$$TE\left(d\bar{2},d\bar{B},\bar{x}\right)=\mathbb{E}\left(w|d\bar{2},d\bar{B},d2dB=1,\bar{x}\right)-\mathbb{E}\left(w|d\bar{2},d\bar{B},d2dB=0,\bar{x}\right)=\alpha_3$$

(2) The treatment effect is equivalent to the difference-in-differences estimator,

$$DD\left(\bar{x}\right)=\left\{\begin{array}{c}\left[\mathbb{E}\left(w|d2=1,dB=1,\bar{x}\right)-\mathbb{E}\left(w|d2=0,dB=1,\bar{x}\right)\right]+\\-\left[\mathbb{E}\left(w|d2=1,dB=0,\bar{x}\right)-\mathbb{E}\left(w|d2=0,dB=0,\bar{x}\right)\right]\end{array}\right\}=\alpha_3$$

When the model is non-linear, the treatment effect is given by,

$$TE\left(d\bar{2},d\bar{B},\bar{x}\right)=F\left(\alpha_1 d\bar{2}+\alpha_2 d\bar{B}+\alpha_3+\beta\bar{x}\right)-F\left(\alpha_1 d\bar{2}+\alpha_2 d\bar{B}+\beta\bar{x}\right) \tag{3.1}$$

and the two previous conclusions are no longer valid because of the nonlinearity of the model. The treatment effect is neither a constant (i.e. it does not coincide with the coefficient of the interaction term, α_3)[4] nor does it coincide with the difference-in-differences estimator,

$$DD\left(\bar{x}\right)=\left[F\left(\alpha_1+\alpha_2+\alpha_3+\beta\bar{x}\right)-F\left(\alpha_1+\beta\bar{x}\right)\right]-\left[F\left(\alpha_2+\beta\bar{x}\right)-F\left(\beta\bar{x}\right)\right] \tag{3.2}$$

In fact, it is relatively straightforward to construct examples where difference-in-differences and the treatment effect have opposite signs[5].

Therefore, a difference-in-difference procedure will not identify the treatment effect in a non-linear model (such as the one we required in our analysis). The same conclusion applies to the difference-in-difference-in-differences estimator proposed by Moehling [**46**]. The treatment effect can be consistently estimated by plugging in the estimates for the parameter of the model into equation (3.1).

---

[4] Nevertheless, the treatment effect and the coefficient α_3 will share the sign.

[5] For example, consider $F\left(x\right)=\Phi\left(x\right)$ (probit model), and set $\bar{x}=d\bar{2}=d\bar{B}=0$, $\alpha_1=0.1$, $\alpha_2=-1.5$, $\alpha_3=0.1$. Since $\alpha_3>0$, then $TE\left(0,0,0\right)>0$, but calculations reveal that $DD\left(0,0,0\right)<0$.

One might also consider estimating the differences in treatment effects between the treatment and control groups. If we denote by $\left(d\bar{2}_Y, d\bar{B}_Y, \bar{x}_Y\right)$ the vector of covariate values for young children (treatment group) and by $\left(d\bar{2}_O, d\bar{B}_O, \bar{x}_O\right)$ the vector of covariate values for old children (control group), then the difference in treatment effects is given by,

$$DTE = TE\left(d\bar{2}_Y, d\bar{B}_Y, \bar{x}_Y\right) - TE\left(d\bar{2}_O, d\bar{B}_O, \bar{x}_O\right) \tag{3.3}$$

If our object of interest is the treatment effect for young children, the difference in treatment effect will identify it if and only if the treatment effect for old children is zero. In the next subsection, we will explain why multiplicity of equilibria can cause the treatment effect for old children to be different from zero.

3.2.2. Differencing with multiplicity of equilibria

In a seminal paper, Basu and Van [**8**], developed a reduced form model of child labor. The model is based on two main assumptions or axioms. The first one is the *luxury axiom*, by which a family will send the children to work only if the family's income from non-child labor sources is sufficiently low. Children have very important opportunity costs of working, such as not receiving education or not enjoying their leisure. As decision makers, the parents are forced to send their kids to work when the family income is so low that the work of every member of the household is necessary for survival. The second axiom is the *substitution axiom*, by which from a firm's point of view, adult and child labor are substitutes.

When these assumptions are incorporated into a household decision model where the main source of family income is labor, the model can present a downward sloping supply of aggregate labor. If the wage is very low, then families are forced to send their children to work, generating a high aggregate labor supply. If the wage is very high, then working parents can support their entire household by themselves and avoid sending their children to work, generating a low amount of labor supply. When combined with a downward sloping demand for labor, this model has the possibility of generating multiplicity of equilibria.

We can adopt the Basu and Van [**8**] model to analyze the equilibrium of the labor market in the U.S. at the end of the nineteenth century. Consider a situation where there are young and old children. The separation between young and old occurs at 14 years of age, which is precisely the cutoff imposed by the child labor legislation.

Suppose labor market conditions are such that there is only one equilibrium in which every household decides to send its children to work. When child labor legislation is imposed, young children are removed from the labor market and old children keep working. In this case, looking at the difference in treatment effects between young and old children correctly identifies the effect of the law. This is the reasoning followed by Moehling [**45**] and Moehling [**46**].

Instead, suppose that the labor market is such that there are two stable equilibria described by Basu and Van [**8**]. In this case, a ban on young child labor may generate a change from the equilibrium with high child labor to one with low child labor. Young child labor is directly reduced by the prohibition, but general equilibrium forces cause an increase in wage that justifies overall reduction in child labor. In this case, child labor legislation is extremely effective, in the sense that it reduces the labor participation of all children, and not only young children covered by the law. Moreover, the legislation is benign, because instead of constraining the behavior of economic agents, it causes a change to an equilibrium where agents voluntarily respect the law. In this case, the difference in treatment effects will fail to identify the effect of the law. Even though the law is extremely effective in reducing young child labor, it also reduces old child labor and hence, the effectiveness of the law is underestimated by the difference in treatment effects. It is possible that this could explain Moehling [**46**] finding: "Although the predicted probabilities for the treatment group–13-year-old boys living in the states that enacted the age minima of 14– fell substantially between 1880 and 1900, so too did the predicted probabilities for the control groups".

3.3. Economic model

In this section, we consider a structural model of the economy along the lines of Basu and Van [**8**].

3.3.1. Setup

Consider the following overlapping-generations model. Each household in the economy is composed of two individuals: a parent and a child. An agent in the economy lives four periods. He is a young child in the first period, an old child in the second period, a young adult in the third period and an old adult in the fourth period. In the first two periods of his life, the individual lives under the supervision of the adult, who makes all decisions in the household. At the end of the second period, the old adult dies, the old child becomes a young adult and gives birth to a young child. For the two remaining periods of his life, he will be the decision maker in his household.

The flow utility of the adult is given by,

$$u\left(c_{i,a}, c_{i,k}, l_{i,k}\right) = u\left(c_{i,a} + c_{i,k}, \left(\psi + \delta 1\left[i = y\right]\right)\left(1 - l_{i,k}\right)\right)$$

where $i \in \{o, y\}$ denotes the age of the household, $c_{i,a}$ refers to the adult's consumption, $c_{i,k}$ refers to the child's consumption and $l_{i,k} \in \{0, 1\}$ is a binary variable indicating whether the child works or not. The indicator variable $1\left[i = y\right]$ takes a value of one if we are referring to a young household (which includes a young child) and zero otherwise.

In this model, an adult is altruistic in two ways: he cares about his child's consumption and his child's leisure[6]. When an old child works, his parent suffers a disutility of $\psi > 0$. When a young child works, his parents suffers a disutility of $(\psi + \delta)$, where $\delta > 0$ represents the extra cost of forcing a young child to work. We assume that the utility function is weakly increasing in both coordinates. Moreover, we assume that the household has a subsistence consumption level, denoted by $\bar{C}$, such that if the household consumes less than this amount, the adult only cares about maximizing consumption and child leisure becomes irrelevant[7]. These features imply that adult's preferences satisfy the leisure axiom in Basu and Van [**8**].

A household's wealth is given by labor income. Households are subject to period budget constraints, and we assume, for simplicity, that there is no borrowing or lending,

$$c_{i,a} + c_{i,k} = w_a + w_k l_{i,k}$$

---

[6] Preference towards children's leisure could also be representing taste for kid's education.

[7] This feature is not necessary to get the main results of the model, but it helps to strengthen the intuition.

where w_a represents the equilibrium wage for the employed adult and w_k represents the equilibrium wage for the employed child[8].

In every period $2N$ simultaneous families coexist: N young families and N old families. At the end of each period, each young family becomes an old family and each old family becomes a young family (the old adult dies, the old child becomes a young adult and gives birth to a young child)[9].

We now model the production sector in this economy. There is a continuum of perfectly competitive firms, each producing the unique manufactured good according to the following production function,

$$f\left(L_a^d, L_k^d, K\right) = F\left(L_a^d + \frac{L_k^d}{\mu}, K\right)$$

where F is a strictly increasing, marginally decreasing and homogeneous of degree one function. Labor input is measured in adult labor equivalent units, $L_a^d + L_k^d/\mu$, where L_a^d and L_k^d are the amounts of adult and child labor demanded, respectively. Implicit in the equation is that adults and children are perfect substitutes in production: one working adult produces the same amount as $\mu\,(> 0)$ working children[10]. This introduces the substitution axiom by Basu and Van [**8**].

Capital for production is provided by wealthy capitalist families, who own certain amount of physical capital and offer it to the firms in a fixed supply. We denote this fixed supply by $\bar{K}$. In return, these families earn a rental rate of capital given by r.

---

[8] Implicit in the notation is the fact that firms discriminate between adults and children, but not between young and old adults and young and old children. This is mainly assumed for simplicity.

[9] Even though the age structure in this economy may be unrealistic, the objective is to keep a constant proportion of young and old children and adults in the economy.

[10] We assume that a child is less productive than a grown up and, therefore, $\mu > 1$.

3.3.2. Optimal Decisions

Profit maximization implies that equilibrium adult wage is given by,

$$w_a = F_1\left(\left(2 + \left(l_{y,k} + l_{o,k}\right)/\mu\right) N, \bar{K}\right)$$

where F_1 represents the derivative of the production function with respect to labor. The substitutability between child and adult labor implies the following relationship between wages,

$$w_k = w_a/\mu$$

The household head makes the child employment decision. If the household is old, the optimal decision is given by,

$$l_{o,k} = \begin{cases} 1 & \text{if} \quad u\left(w_a, \psi\right) \geq u\left(w_a + w_k, 0\right) \\ 0 & \text{if} \quad u\left(w_a, \psi\right) \leq u\left(w_a + w_k, 0\right) \end{cases}$$

and if the household is young and there is no child labor legislation, the optimal decision is given by,

$$l_{y,k} = \begin{cases} 1 & \text{if} \quad u\left(w_a, \psi + \delta\right) \geq u\left(w_a + w_k, 0\right) \\ 0 & \text{if} \quad u\left(w_a, \psi + \delta\right) \leq u\left(w_a + w_k, 0\right) \end{cases}$$

Notice that if young families decide to send their children to work, then old families will also decide to do so.

3.3.3. Equilibria in the model

This model can generate three different equilibria, each of them characterized by which are the economically active children. We assume the parameters of the model are such that all three equilibria exist[11], which are characterized in the following subsections.

3.3.3.1. The "all children work" equilibrium. In this equilibrium, all children in the economy work. Adult wage is given by $w_a = F_1\left((1+1/\mu)\,2N, \bar{K}\right)$ and child wage is given by $w_k = F_1\left((1+1/\mu)\,2N, \bar{K}\right)/\mu$. The necessary and sufficient condition for the existence of this equilibrium is,

$$u\left(F_1\left((1+1/\mu)\,2N, \bar{K}\right), \psi+\delta\right) \leq u\left(F_1\left((1+1/\mu)\,2N, \bar{K}\right)(1+1/\mu), 0\right)$$

This condition holds, for example, if the equilibrium wage for the adult is below the subsistence level but the combined wages of the adult and the children are over this level (that is, $w_a < \bar{C}$ and $w_k + w_a \geq \bar{C}$). Therefore, every adult forces his child to work, no matter his age, in order to guarantee the subsistence of the family.

3.3.3.2. The "no children work" equilibrium. In this equilibrium, none of the children in the economy work. Adult wage is given by $w_a = F_1\left(2N, \bar{K}\right)$ and child wage is given by $w_k = F_1\left(2N, \bar{K}\right)/\mu$. The necessary and sufficient condition for the existence of this equilibrium is,

$$u\left(F_1\left(2N, \bar{K}\right), \psi\right) \geq u\left(F_1\left(2N, \bar{K}\right), 0\right)$$

In order for this condition to hold, it is necessary that equilibrium adult wages are above the subsistence level (that is, $w_a > \bar{C}$).

---

[11] This is not necessarily true for every parameter value.

3.3.3.3. The “only old children work” equilibrium. In this equilibrium, old families send their old children to work and young families prefer not to send their young children to work. Adult wage is given by $w_a = F_1\left(\left(2+1/\mu\right)N,\bar{K}\right)$ and child wage is given by $w_k = F_1\left(\left(2+1/\mu\right)N,\bar{K}\right)/\mu$. The necessary and sufficient condition for the existence of this equilibrium is,

$$\begin{aligned}&u\left(F_1\left(\left(2+1/\mu\right)N,\bar{K}\right),\psi+\delta\right)\geq\\&\geq u\left(F_1\left(\left(2+1/\mu\right)N,\bar{K}\right)\left(1+1/\mu\right),0\right)\geq u\left(F_1\left(\left(2+1/\mu\right)N,\bar{K}\right),\psi\right)\end{aligned}$$

In order for this condition to hold, it is necessary that equilibrium adult wages are above the subsistence level (that is, $w_a > \bar{C}$).

3.3.4. The effect of child labor legislation

As mentioned earlier, we assume that all three equilibria exist as shown in figure 3.1. For low wages, all the households of the economy will decide to send their children to work. At intermediate wages, old parents will send their old children to work but young parents will decide not to. Finally, when wages are high enough, all families are sufficiently wealthy to avoid sending their children to work.

3.3.4.1. Pre-legislation equilibrium. In 1880, the U.S. labor market was characterized by high levels of child labor participation[12]. Based on this observation, we assume that the pre-legislation situation was an “all children work” equilibrium. Therefore, the

---

[12] In the 1880 U.S. census, 47% of boys ages 10 to 13 and 63% of boys ages 14 to 15 were reported working for wages. The corresponding figures for girls are 36% and 39%, respectively.

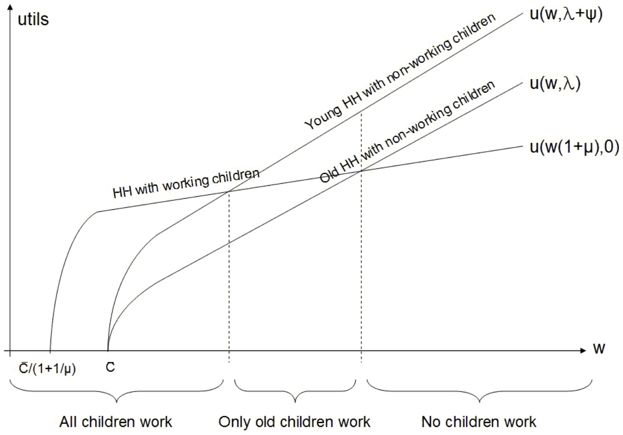

Figure 3.1. Multiple equilibria

pre-legislation equilibrium wages are given by,

$$F_1\left(2\left(1+1/\mu\right)N, \bar{K}\right) = w_a^{pre} = w_k^{pre}/\mu$$

The situation is depicted in figure 3.2, which represents the equilibrium for both young and old families.

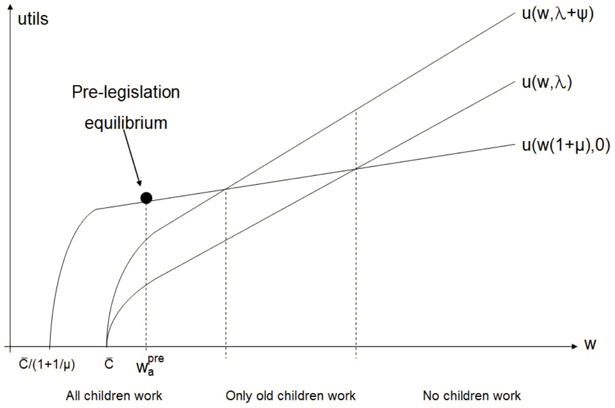

Figure 3.2. Pre-equilibrium situation

3.3.4.2. Post-legislation equilibrium. We now introduce child labor legislation to our model. Since the bulk of state-wide child labor legislation issued in the U.S. at the end of the nineteenth century was a ban on child labor for children of less than 14 years of age, we set the cutoff age between young and old children at 14 years old.

The effect of the child labor legislation depends on whether the legislation is properly enforced or not. In order to explore more interesting results, suppose that the legislation is properly enforced. In this case, the effect of the law depends on the fundamentals of the economy. As a consequence of the prohibition, young children are forced out of the labor market and thus, the pre-legislation "all children work" equilibrium is eliminated. The elimination of young child labor supply causes an increase in wages, which determines the general equilibrium effect of the legislation according tot the following cases.

Case 0: Ineffective legislation. In this case, legislation has no effect on child labor participation and the post-legislation situation is identical to the pre-legislation situation. This is shown in figure 3.3. This outcome is only possible if the legislation is not enforced.

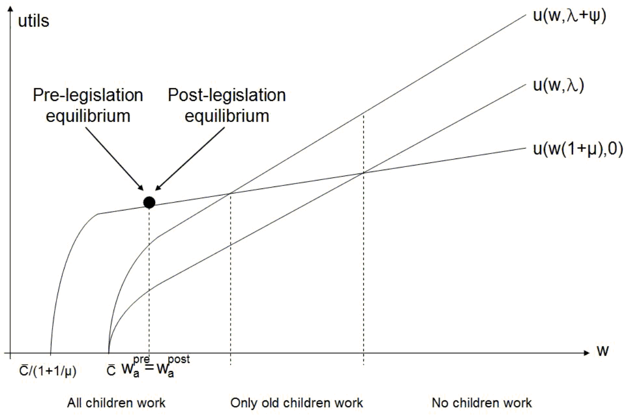

Figure 3.3. Case 0: no effect

Case 1: Distortive legislation. In this case, the prohibition of young child labor produces a mild increase in equilibrium wages, which is not sufficient to induce changes in households' decisions. Old households still decide to send their old children to work and, in absence of legislation, young households would do so too. The child labor legislation is not Pareto optimal and therefore, not benign[13].

This is shown in figure 3.4. The legislation is effective in reducing young child labor and ineffective in reducing old child labor.

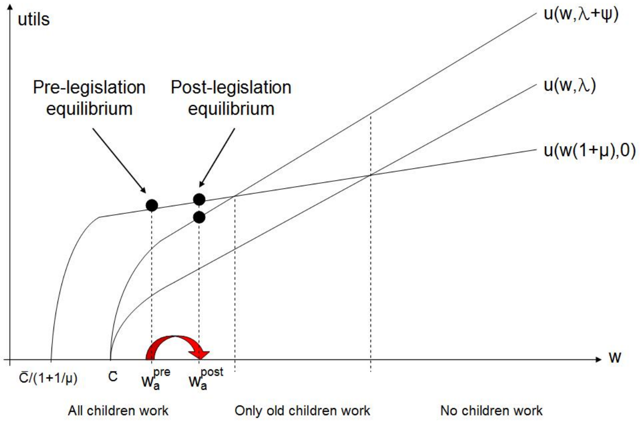

Figure 3.4. Case 1: small distortive effect

Case 2: Small benign legislation. In this case, the elimination of young child labor produces a greater increase in wages. The raise in family's income is large enough to induce young families to remove their children from the labor market, but not enough to convince old families to remove their children from the labor market.

The legislation causes a switch between two equilibria. In this case, the pre- and post-legislation equilibrium are both Pareto optimal situations. This is represented in figure

---

[13] This would also be the outcome if the "all children work" equilibrium were the only equilibrium in the economy.

3.5. As a consequence of the child labor laws, young child labor is reduced and old child labor remains high. Notice how this case is observationally equivalent to the previous one.

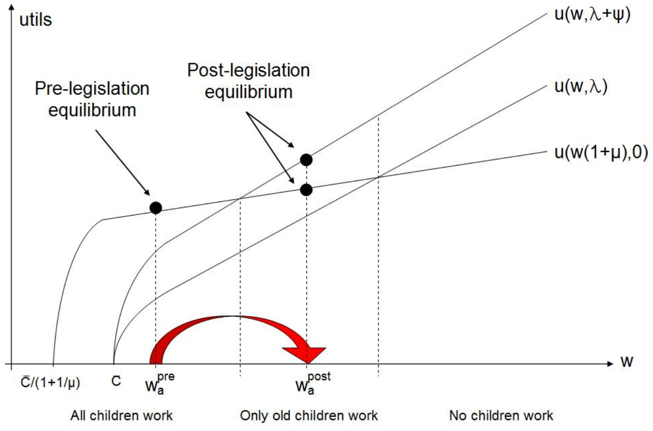

Figure 3.5. Case 2: small benign effect

Case 3: Large benign legislation. In this case, the removal of young child labor produces a big increase in wages, causing a significant increase in household income. This induces all families to remove their children from the labor force, regardless of their age. As in the previous case, the legislation causes a switch between equilibria and the post-legislation equilibrium is also Pareto optimal.

Figure 3.6 depicts the situation. The legislation is effective in reducing child labor levels across all ages, even though the legislation is only explicitly targeted to young children. Moreover, notice hat computing differences in treatment effects between treated and untreated children would severely underestimate the treatment effect on young child labor[14].

---

[14] In our simple theoretical illustration, the labor market participation for both children goes from 100% to 0% but the difference in the treatment effects is zero.

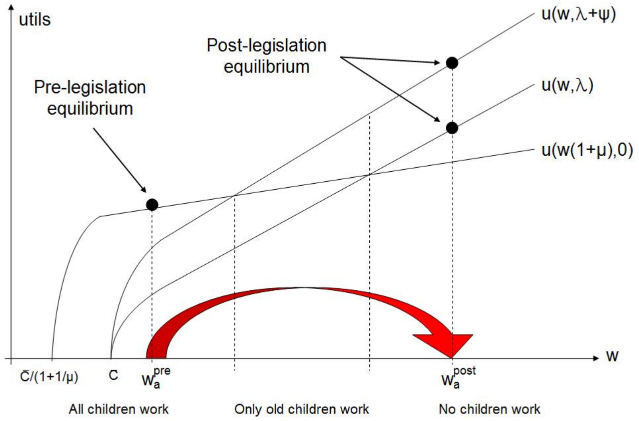

Figure 3.6. Case 3: large benign effect

3.3.4.3. Conclusions. The following table summarizes how young and old child labor participation can help us characterize whether child labor legislation was effective and/or benign.

	Equilibrium level of...		Legislation is...			Diff. in T.E. between young & old equals T.E. on young?
	Old child participation	Young child participation	Effective on young?	Effective on old?	Benign?	
Case 0	constant	constant	No	No	Yes	Yes
Case 1	constant	decreases	Yes	No	No	Yes
Case 2	constant	decrease	Yes	No	Yes	Yes
Case 3	decreases	decreases	Yes	Yes	Yes	No

The legislation is effective in reducing young (old) child labor if the treatment effect on young (old) children is negative. Treatment effects can be directly estimated by plugging in our estimators in equation 3.1.

The legislation is benign if its effect is not restricting the household's behavior but rather changing the pre-legislation equilibrium to a different one, where families with young children voluntarily choose to comply with the legislation.

From our previous analysis, case 1 is the only one where the legislation is not benign. Unfortunately, cases 1 and 2 are observationally equivalent and hence, we will only be able to determine that the legislation is benign in cases 0 and 3. In case 0, the legislation has no effect whatsoever, which makes it benign in a non-interesting way. In case 3, the legislation causes a reduction in old child labor, distinguishing it from the remaining cases.

3.4. Econometric methodology

In this section, we describe the econometric model and the data used for our inference.

3.4.1. Econometric model

Denote by w the binary variable of interest that takes value of one if the child is employed and zero otherwise. We assume that w is determined by the following binary response model,

$$w = \begin{cases} 1 & \text{if} \quad \alpha_1 d2 + \alpha_2 dB + \alpha_3 d2dB + \beta x + \varepsilon \geq 0 \\ 0 & \text{if} \quad \alpha_1 d2 + \alpha_2 dB + \alpha_3 d2dB + \beta x + \varepsilon < 0 \end{cases}$$

where $d2$ is the binary variable that takes value of one if the observation occurred in 1900 (period 2) and zero if it occurred in 1880 (period 1), dB is the binary variable that takes the value of one if the observation corresponds to a state that issued child labor legislation in 1990 (B states) and zero otherwise (A states), $d2dB$ is their interaction, x are remaining observable controls and ε denotes an unobserved random term. We assume

ε is independent and normally distributed, i.e. we adopt a probit specification[15]. If we denote the parameters of the model by $\pi = [\alpha_1, \alpha_2, \alpha_3, \beta]$ and we denote the observable covariates vector by $X = [d2, dB, d2dB, x]$, the probability of employment evaluated at X is given by,

$$P\left(w = 1 | d2, dB, d2dB, x\right) = F\left(X\pi\right)$$

The parameters of the model can be consistently and asymptotically efficiently estimated by maximum likelihood estimation[16]. We denote the estimated parameters $\hat{\pi} = \left[\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\beta}\right]'$. In this case, the estimate of the probability of employment evaluated at X is given by,

$$\hat{P}\left(w = 1 | d2, dB, d2dB, x\right) = F\left(X\hat{\pi}\right)$$

Under the usual regularity conditions, $\sqrt{n}\left(\hat{\pi} - \pi\right)$ is an asymptotically normally distributed vector with mean zero and variance covariance matrix given by the Outer Product matrix (or, equivalently, Hessian matrix), which we denote by $V\left(\pi\right)$, and whose consistent estimator is denoted $V\left(\hat{\pi}\right)$.

In the results section, we will be interested in testing whether the probability of children employment is zero or not. Using the Delta method, we deduce that,

$$\sqrt{n}\left(F\left(X\hat{\pi}\right) - F\left(X\pi\right)\right) \xrightarrow{d} N\left(0, f\left(X\pi\right) X'V\left(\pi\right) X f\left(X\pi\right)\right)$$

---

[15] The logit model produced similar results.

[16] For an excellent reference on maximum likelihood estimation and all other topics in this subsection, see Amemiya [**1**].

where f denotes the derivative of F. This result can be used to show that, under the null hypothesis that the probability of employment at X is zero ($H_0 : F(X\pi) = 0$), then,

$$\frac{n\left(F\left(X\hat{\pi}\right)\right)^2}{f\left(X\hat{\pi}\right)X'V\left(\hat{\pi}\right)Xf\left(X\hat{\pi}\right)} \xrightarrow{d} \chi_1$$

where χ_1 denotes the chi-squared distribution with one degree of freedom.

The treatment effect of the legislation corresponds to the change in the probability of employment caused exclusively by the child labor legislation, keeping the remaining covariates at a specific level of interest. Under our assumptions, the treatment effect can be identified by computing the difference in the probability from a situation with no child labor legislation ($d2dB = 0$) to a situation with child labor legislation ($d2dB = 1$), keeping the remaining covariates constant at a specific level of interest. If we denote $X_0 = \left[d\bar{2}, d\bar{B}, 0, \bar{x}\right]$ and $X_1 = \left[d\bar{2}, d\bar{B}, 1, \bar{x}\right]$, then the treatment effect evaluated at $\left(d\bar{2}, d\bar{B}, \bar{x}\right)$ is given by,

$$TE\left(d\bar{2}, d\bar{B}, \bar{x}\right) = F\left(X_1\pi\right) - F\left(X_0\pi\right)$$

and it can be consistently estimated by,

$$T\hat{E}\left(d\bar{2}, d\bar{B}, \bar{x}\right) = F\left(X_1\hat{\pi}\right) - F\left(X_0\hat{\pi}\right)$$

In the results section, we will be interested in testing whether the treatment effect is zero or not. Using the Delta method, under the null hypothesis that the treatment effect at $\left[d\bar{2}, d\bar{B}, \bar{x}\right]$ is zero ($H_0 : TE\left(d\bar{2}, d\bar{B}, \bar{x}\right) = 0$), then,

$$\frac{n\left(T\hat{E}\left(d\bar{2}, d\bar{B}, \bar{x}\right)\right)^2}{\left[\left(f\left(X_1\hat{\pi}\right)X_1 - f\left(X_0\hat{\pi}\right)X_0\right)V\left(\hat{\pi}\right)\left(X_1'f\left(X_1\hat{\pi}\right) - X_0'f\left(X_0\hat{\pi}\right)\right)\right]} \xrightarrow{d} \chi_1$$

3.4.2. Details of the empirical methodology

The data were constructed from a random sample of individual level records from the 1880 and 1900 U.S. federal censuses[17] which are part of the Integrated Public Use Microdata Series or IPUMS[18]. The 1880 dataset is a 1-in-100 sample containing data on over 107,000 households and 502,000 individuals and the 1900 dataset is a 1-in-760 sample containing data on over 89,000 households and 366,000 individuals.

Following Moehling [**46**], we restrict attention to children living in non-agricultural households with at least one parent[19]. We also restrict our analysis to white children, since white and non-white kids faced different labor market opportunities. To simplify the construction of variables, we restrict attention to households that contained only one family and to children who were sons or daughters of the household head[20].

In order to implement our inference, we need to adopt a working definition of a child. We define children to be individuals of ages 13 and 14, where 13-year-olds play the role of young children and 14-year-olds play the role of old children[21]. Since the labor market for

---

[17] Moehling [**46**] also uses information from the 1910 U.S. census. In the 1910 census, 81% of the children ages 10 to 15 have missing employment information. Since employment information is necessary to construct our dependent variable, we decided not to use this census year. For the 1880 and 1900 censuses, the percent of missing employment information is always below 18%.

[18] The IPUMS data and its description are publicly available at http://www.ipums.umn.edu.

[19] Children in agricultural households worked mainly in agriculture, which was not targeted by the child labor movement at the end of the century.

[20] We eliminate few observations of children between 12 and 15 years of age who were parents.

[21] We have also conducted the analysis defining children to be individuals of ages 12 to 15, where 12 to 13 are young children and 14 and 15 are the old children. This alternative definition is produces similar results, and therefore we consider that our conclusions are robust to how children are defined.

boys and girls were considered relatively different, we run separate estimations for each of these groups.

We now discuss the main variables in the study. The dependent variable of the study is a binary variable that indicates whether the child has a gainful occupation or not (possibly restricted to occupation in certain sector). Ideally, we would like to observe if an individual held any type of gainful occupation, whether regular and intermittent but, unfortunately, the census data only reports each individual's regular or usual form of employment. As a consequence, we will be limited to study the effect of child labor legislation on children that work *on a regular basis*.

Following Moehling [**46**], we run two separate estimations. In the first one, the dependent variable indicates if the individual works regularly in any sector[22]. In the second one, the dependent variable indicates if the individual works regularly in the manufacturing sector[23]. Observations with missing occupational information are ignored[24].

The typical state-wide child labor legislation imposed a variety of restrictions: minimum age limits for employment in the manufacturing sector, maximum work hour limits, minimum school enrollment and minimum grade completion requirements. Following Moehling [**46**], we focus on the minimum age for employment in the manufacturing sector since this is the one that imposed greater constraints on the employment of children. Specifically, we define child labor legislation to be the prohibition of children of less than 14 years of age to be employed in the manufacturing sector. Until 1880, almost none of

---

[22] A child will not be considered to have a regular gainful occupation if, according to the 1950-occupation classification, he is at school, keeps the house, helps his parents, is unemployed or without occupation, is invalid or disabled with no occupation reported or has any other non-occupation.

[23] A child will be considered to work in this sector if, according to the 1950-occupational classification, he is employed in a craftsmen or operatives category.

[24] There is no missing occupational information in the 1880 census and less than 18% in the 1900 census.

the U.S. states had passed child labor legislation and, according to Sanderson [**59**], these laws had little publicity and were poorly enforced. By 1900, already eleven states had issued child labor legislation.

We now proceed to explain the construction of the explanatory variables. The information regarding which states passed child labor legislation between 1880 and 1900 can be found in Ogburn [**49**], which is reproduced in Moehling [**46**]. We are also guided by Moehling [**46**] in the choice of our control variables. To control for the household's wealth we include the household head's age and squared age, his Duncan socioeconomic index and his occupational score. We also include variables indicating whether the head reported having no occupation, whether he had an occupation that required no skills, and whether he had a professional or technical occupation[25]. In addition to that, binary variables indicating whether the head could read and/or write are included as well. We also control for the months that the household head has been unemployed in the previous year. We include binary variables that indicate whether the mother and/or the father were absent, whether the child and/or the parents were foreign born, the number of older and younger sisters and brothers, and the presence in the household of children of less than 5 years of age. To capture the human capital stock of the child, we include binary variables that indicate whether the child could read and/or write. To capture the size of the labor market we introduce binary variables that indicate whether the household lived in an area with high population level (25,000 or more habitants), medium population level (between 2,500 and 24,999 habitants) or low population level (less than 2,500 habitants, omitted). We also introduce variables that indicate the metropolitan status of

---

[25] Excluded categories include occupations that require skill, clerical occupation, sales, managers, proprietors and officials.

the household, that is, whether the household was located outside a metropolitan area (Metro1), in a central city within a metropolitan area (Metro2) or outside a central city within a metropolitan area (Metro3, omitted).

Summary statistics for all the variables used in the regressions are provided in tables 3.1 and 3.2.

3.5. Results

This section reports the results of the estimation.

3.5.1. Regression estimates

Table 3.3 provides the estimates of the parameters of the likelihood of having an occupation in any sector, i.e. general employment. We indicate statistical significance of the coefficients with the usual star notation[26]. Under the assumptions of the model, the sign of the coefficient associated to the variable $d2dB$ is the sign of the treatment effect of the child labor legislation on the child labor. In all the samples, the child labor legislation reduced the probability of having an occupation in any sector.

Table 3.4 presents the estimates of the parameters for the likelihood of having an occupation in the manufacturing sector. Results indicate that child labor legislation reduced the probability of having an occupation in the manufacturing sector.

---

[26] One star means significant at 10% level, two stars mean significant at 5% level and three stars mean significant at 1% level.

	Boys			Girls		
Variable	Mean	Std. dev.	N.	Mean	Std. dev.	N.
Works in any sector	0.142	0.349	5737	0.054	0.225	5726
Works in manuf.	0.028	0.166	5737	0.009	0.097	5726
$d2dB$	0.225	0.418	5737	0.258	0.438	5726
dB	0.476	0.499	5737	0.493	0.5	5726
$d2$	0.466	0.499	5737	0.518	0.5	5726
Metro Area 1	0.323	0.468	5737	0.33	0.47	5726
Metro Area 2	0.096	0.295	5737	0.1	0.3	5726
U.S. born	0.919	0.272	5737	0.928	0.258	5726
Absent father	0.129	0.335	5737	0.128	0.334	5726
Absent mother	0.034	0.181	5737	0.033	0.178	5726
No. children under 5	0.605	0.826	5737	0.611	0.826	5726
School	1.837	0.369	5737	1.855	0.352	5726
Read	0.958	0.201	5737	0.973	0.162	5726
Write	0.944	0.229	5737	0.962	0.191	5726
No. older brother	0.731	0.945	5737	0.738	0.954	5726
No. younger brother	1.027	1.106	5737	1.038	1.123	5726
No. older sister	0.698	0.906	5737	0.67	0.899	5726
No. younger sister	1.013	1.1	5737	1.008	1.117	5726
Head's age	44.761	7.726	5737	44.584	7.813	5726
Head's age^2	2063.227	731.151	5737	2048.785	739.002	5726
Head reads	0.922	0.269	5737	0.926	0.262	5726
Head writes	0.905	0.293	5737	0.909	0.287	5726
Head S.E.I.	24.253	22.363	5737	24.776	22.759	5726
Head's occup. score	23.552	12.486	5737	23.875	12.744	5726
Head's unemp. months	0.911	2.205	5084	0.844	2.144	5093
Parents born in U.S.	0.505	0.5	5737	0.507	0.5	5726
Medium population	0.218	0.413	5737	0.229	0.42	5726
Big population	0.38	0.485	5737	0.393	0.488	5726
Head has no occup.	0.113	0.317	5695	0.11	0.313	5689
Head is unskilled	0.229	0.42	5695	0.235	0.424	5689
Head is professional	0.035	0.184	5695	0.041	0.199	5689

Table 3.1. Summary statistics for young children

3.5.2. Effectiveness analysis

Table 3.5 provides estimates of the probability of child employment in any sector with and without child labor legislation. By computing the difference between these two, we

	Boys			Girls		
Variable	Mean	Std. dev.	N.	Mean	Std. dev.	N.
Works in any sector	0.438	0.496	5333	0.18	0.384	5135
Works in manuf.	0.1	0.299	5333	0.038	0.19	5135
$d2dB$	0.236	0.424	5333	0.268	0.443	5135
dB	0.492	0.5	5333	0.513	0.5	5135
$d2$	0.47	0.499	5333	0.524	0.499	5135
Metro Area 1	0.352	0.478	5333	0.337	0.473	5135
Metro Area 2	0.097	0.295	5333	0.097	0.296	5135
U.S. born	0.895	0.307	5333	0.909	0.288	5135
Absent father	0.159	0.366	5333	0.142	0.349	5135
Absent mother	0.036	0.187	5333	0.037	0.189	5135
No. children under 5	0.462	0.748	5333	0.5	0.767	5135
School	1.554	0.497	5333	1.635	0.481	5135
Read	0.962	0.19	5333	0.975	0.155	5135
Write	0.949	0.219	5333	0.967	0.179	5135
No. older brother	0.680	0.875	5333	0.672	0.88	5135
No. younger brother	1.089	1.181	5333	1.096	1.176	5135
No. older sister	0.609	0.824	5333	0.59	0.825	5135
No. younger sister	1.061	1.147	5333	1.107	1.2	5135
Head's age	46.609	7.796	5333	46.47	7.63	5135
Head's age^2	2233.147	763.71	5333	2217.694	745.364	5135
Head reads	0.913	0.282	5333	0.927	0.26	5135
Head writes	0.892	0.311	5333	0.91	0.286	5135
Head S.E.I.	23.045	22.333	5333	25.41	23.429	5135
Head's occup. score	22.532	13.088	5333	23.647	13.424	5135
Head's unemp. months	0.87	2.168	4528	0.855	2.102	4456
Parents born in U.S.	0.485	0.5	5333	0.504	0.5	5135
Medium population	0.224	0.417	5333	0.224	0.417	5135
Big population	0.406	0.491	5333	0.396	0.489	5135
Head has no occup.	0.15	0.357	5286	0.129	0.336	5090
Head is unskilled	0.224	0.417	5286	0.208	0.406	5090
Head is professional	0.034	0.181	5286	0.044	0.205	5090

Table 3.2. Summary statistics for old children

also compute an estimate of the treatment effect of the child labor legislation on child labor. The remaining control variables are evaluated at five different values of interest: (a) the U.S. average on both periods, (b) the pre-treatment (1880) average on non-treated

	Young children		Old children	
Variable	Boys	Girls	Boys	Girls
$d2dB$	-0.542***	-0.292	-0.0897	-0.540***
$d2$	-0.342***	-0.235	-0.313***	0.342**
dB	-0.121	0.0175	-0.0886	0.530***
Metro area 1	0.192	0.490**	0.256	0.156
Metro area 2	-0.108	0.204	-0.0651	0.472***
U.S. born	-0.270*	-0.427***	-0.386***	-0.253*
Absent father	0.552***	0.549**	0.164	0.537**
Absent mother	0.240	-0.217	-0.0652	-0.222
No. children under 5	-0.209***	-0.208**	-0.0894	-0.0836
School	-1.990***	-1.660***	-2.073***	-1.908***
Reads	1.321***	0.0550	0.278	0.419
Writes	-1.003***	0.425	-0.0472	-0.421
No. older brothers	0.0513	-0.0322	0.0118	-0.0516
No. younger brothers	0.0903**	0.0878*	0.142***	0.0451
No. older sisters	-0.147***	0.0930	-0.0118	0.0879
No. younger sisters	0.116***	0.175***	0.0553	0.129***
Head's age	0.0757	-0.0636	-0.0143	-0.0206
Head's age^2	-0.000798	0.000658	0.000194	0.000273
Head reads	-0.310	-0.236	0.00439	-0.632
Head reads	-0.0114	0.0624	-0.376	0.453
Head writes	-0.00827*	-0.0187***	-0.008**	-0.0178***
Head's S.E.I	-0.00356	0.0127	0.0107	0.0169
Head's unemp. months	0.0260	0.0166	0.0469***	0.0486***
Parents born in U.S.	-0.0385	-0.193	0.0303	0.0407
Medium population	-0.142	-0.0370	-0.0270	0.0488
Big population	-0.369	-0.120	-0.163	0.457**
Head has no occup.	-0.611	dropped	-0.366	0.407
Head is unskilled	0.225**	0.209*	0.0366	0.169*
Head is professional	0.151	0.378	-0.360	-0.237
Constant	1.314	2.649**	3.473***	1.3465
Number of observations	2348	2388	2344	2375

Table 3.3. Probit estimates of the likelihood of having an occupation in any sector

states (A states), (c) the pre-treatment (1880) average on treated states (B states), (d) the post-treatment (1900) average on non-treated states (A states) and, finally, (e) the post-treatment (1900) average on treated states (B states).

	Young children		Old children	
Variable	Boys	Girls	Boys	Girls
$d2dB$	-0.360	-0.687*	-0.184	-1.265***
$d2$	-0.0206	0.818***	0.175	1.044***
dB	-0.451**	0.379	-0.117	0.595**
Metro area 1	0.231	0.166	-0.169	0.918*
Metro area 2	0.0547	-0.376	-0.506***	0.446
U.S. born	-0.313	-0.534**	-0.213*	-0.114
Absent father	0.704**	0.554	0.223	-0.498
Absent mother	0.224	-0.136	0.253	-0.425
No. children under 5	-0.183	0.111	0.0178	-0.0139
School	-1.411***	-1.72***	-1.165***	-1.606***
Reads	0.418	0.557	-0.384	1.181
Writes	-0.0771	-0.102	0.739	-0.974*
No. older brothers	0.0280	-0.206*	0.105**	-0.218**
No. younger brothers	0.0588	-0.0956	0.0672	0.0426
No. older sisters	-0.124	0.0583	-0.0829	-0.0731
No. younger sisters	0.169**	-0.0500	0.0191	-0.0361
Head's age	-0.127**	-0.0463	-0.0458	0.0821
Head's age^2	0.00136**	0.000742	0.000539	-0.000694
Head reads	-0.330	0.594	0.0379	-1.119
Head writes	0.153	-0.679	-0.331	1.423
Head's S.E.I.	-0.0515***	-0.0195	-0.0187***	0.00651
Head's occup. score	0.0969***	0.0317	0.0449***	-0.0236
Head's unemp. months	0.0607**	0.0559*	0.00568	0.0479
Parents U.S. born	-0.371**	-0.363	-0.450***	-0.385*
Medium population	-0.691***	0.0132	-0.270**	0.0970
Big population	-0.395	0.332	-0.0460	-0.190
Head has no occup.	0.907	dropped	1.051**	0.708
Head is unskilled	0.606***	0.0975	0.297***	0.170
Head is professional	dropped	dropped	-0.0989	0.156
Constant	2.085	0.00608	0.662	-2.955
Number of observations	2267	2278	2344	2375

Table 3.4. Probit estimates of the likelihood of having an occupation in manufacturing sector

Child labor legislation generated a significant decrease in the probability of employment for young boys and a (barely) insignificant decrease in probability of employment

Young children						
	Boys			Girls		
	With CLL	No CLL	T.E.	With CLL	No CLL	T.E.
1880-1900 US	5.29%***	14.11%***	-8.82%***	0.26%*	0.62%***	-0.36%
1880, A states	11.06%**	24.78%***	-13.71%***	0.23%	0.55%***	-0.32%**
1880, B states	6.94%***	17.41%***	-10.46%***	0.61%*	1.35%***	-0.73%
1900, A states	3.76%**	10.80%***	-7.03%***	0.09%	0.24%**	-0.15%
1900, B states	1.85%***	6.13%***	-4.28%**	0.29%***	0.69%**	-0.36%

Old children						
	Boys			Girls		
	With CLL	No CLL	T.E.	With CLL	No CLL	T.E.
1880-1900 US	20.55%***	23.20%***	-2.64%	1.18%**	4.24%***	-3.06%***
1880, A states	30.60%***	33.81%***	-3.21%	0.32%	1.38%***	-1.08%***
1880, B states	24.46%***	27.36%***	-2.90%	1.82%**	6.03%***	-4.21%***
1900, A states	16.15%***	18.45%***	-2.29%	0.58%*	2.36%***	-1.78%***
1900, B states	11.90%***	13.78%***	-1.88%	4.04%***	11.40%***	-7.36%**

Table 3.5. Likelihood of employment in any sector and treatment effects

for young girls. Also, the legislation generated an insignificant decrease in the probability of employment for older boys and significant decrease in the probability of employment of older girls.

Table 3.6 describes the effectiveness of the child labor legislation in reducing child labor in the manufacturing sector. The legislation produced insignificant decreases in probability of employment for all groups of children.

3.5.3. Benignity analysis

Child labor legislation decreased general employment levels of young and old boys, but the decrease is statistically significant only for young boys. In this case, we do not have sufficient evidence to decide if the legislation had a benign effect on general employment for boys. This type of outcome could be caused by a benign legislation (case 2) or

Young children						
	Boys			Girls		
	With CLL	No CLL	TE	With CLL	No CLL	TE
1880-1900 US	0.13%	0.41%**	-0.27%	0.00%	0.01%	-0.01%
1880, A states	0.38%	1.06%**	-0.67%	0.00%	0.00%	-0.00%
1880, B states	0.07%	0.23%	-0.16%	0.00%	0.01%	-0.01%
1900, A states	0.20%	0.60%**	-0.39%	0.00%	0.01%	-0.01%
1900, B states	0.03%	0.13%	-0.09%	0.01%	0.16%	-0.14%
Old children						
	Boys			Girls		
	With CLL	No CLL	TE	With CLL	No CLL	TE
1880-1900 US	2.27%**	3.47%***	-1.19%	0.00%	0.29%	-0.29%
1880, A states	2.69%*	4.06%***	-1.36%	0.00%	0.01%	-0.01%
1880, B states	1.91%**	2.95%***	-1.03%	0.00%	0.21%	-0.21%
1900, A states	2.71%**	4.08%***	-1.37%	0.00%	0.31%	-0.31%
1900, B states	1.82%***	2.82%***	-0.99%	0.11%	3.68%	-3.57%

Table 3.6. Likelihood of employment in manufacturing sector and treatment effects

by a distortive legislation (case 1). Girls present the opposite pattern. The legislation reduced general employment levels of young and old girls, but the reduction is statistically significant only for old girls. According to our analysis, this is evidence that the reduction of employment of girl labor caused by the imposition of child labor legislation was benign (case 3).

Since the child labor legislation was found to be ineffective in reducing manufacturing employment for all groups of children, we deduce that, in terms of manufacturing employment, it represented a trivially benign public policy (case 0).

3.6. Conclusions

Between 1880 and 1930, the employment rate of children ages 10 to 15 decreased by over 75% in the U.S. economy. During this period, several U.S. states dictated state-wide child labor legislation that imposed minimum age restrictions for employment in the manufacturing sector. This paper studies whether this child labor legislation contributed to the decline in child labor market participation.

In addition to evaluating whether the legislation was effective or not, we analyze the labor market mechanism by which this takes place. This analysis may allow us to establish if the legislation constituted a benign policy or not, that is, whether the legislation imposed constraints to the behavior of the children (not benign) or whether it generated a change in labor market equilibrium (benign).

The effectiveness of the child labor legislation in reducing child labor had already been addressed in the literature, mainly by Moehling [**45**] and Moehling [**46**]. In her work, Moehling estimates a non-linear model (probit or logit) to analyze the children's employment decision and applies differencing estimation methods to characterize the effect of the legislation. We show that differencing estimation methods are inadequate to study the effectiveness of child labor legislation. First, differencing estimators do not identify the treatment effect of interest in non-linear models, such as the one used to analyze labor market participation. Second, when the economy presents multiple equilibria, differencing estimators may severely underestimate the effect of the legislation.

In order to analyze the consequences of child labor legislation, we develop a model along the lines of Basu and Van [**8**], which takes into account the possible multiplicity of equilibria. This model allows us to derive observable consequences to identify whether the legislation was effective and/or benign.

We conduct separate estimates for young children (13-year-olds), who were legally prohibited to work in the manufacturing sector, and old children (14-year-olds), who were free to work. Our estimates indicate that the legislation was effective in reducing general employment for young boys, for old girls and, mildly, for young girls. Based on this information, we can deduce that the legislation was benign for general employment of girls. Unfortunately, our results do not allow us to decide if the legislation was benign for general employment of boys. When we conduct the estimation for labor participation in the manufacturing sector, we find that child labor legislation was ineffective in reducing child labor for both girls and boys and, hence, trivially benign.

References

[1] T. Amemiya. *Advanced Econometrics*. Harvard University Press, 1985.

[2] D.W.K. Andrews. Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, 68(2):399–405, March 2000.

[3] D.W.K. Andrews, S. Berry, and P. Jia. Confidence regions for parameters in discrete games with multiple equilibria with an application to discount chain store location. Mimeo: Yale, May 2004.

[4] D.W.K. Andrews and P. Guggenberger. Validity of subsampling and "plug-in asymptotic" inference for parameters defined by moment inequalities. Mimeo: Yale and UCLA, July 2007.

[5] D.W.K. Andrews and G. Soares. Inference for parameters defined by moment inequalities using generalized moment selection. Mimeo: Yale, May 2007.

[6] G.J. Babu and K. Singh. Edgeworth expansions for sampling without replacement from finite populations. *Journal of Multivariate Analysis*, 17(3):261–278, December 1985.

[7] K. Basu. Child labor: Cause, consequence and cure, with remarks on international labor standards. *Journal of Economic Literature*, 37(3):1083–1119, September 1999.

[8] K. Basu and P.H. Van. The economics of child labor: Reply. *American Economic Review*, 89(5):1386–1388, June 1999.

[9] R. Beran. Prepivoting to reduce level error of confidence sets. *Biometrica*, 74(3):457–468, September 1981.

[10] R. Beran. Prepivoting test statistics: A bootstrap view of asymptotic refinements. *Journal of the American Statistical Association*, 83(403):687–697, September 1988.

[11] A. Beresteanu and F. Molinari. Asymptotic properties for a class of partially identified models. Mimeo: Cornell Northwestern, June 2006.

[12] R.N. Bhattacharya and R.R. Rao. *Normal Approximation and Asymptotic Expansions*. John Wiley and Sons, Inc., 1976.

[13] P. Bickel and R. Freedman. Some asymptotic theory for the bootstrap. *Annals of Statistics*, 9(6):1196–1217, November 1981.

[14] P. Bickel, F. Götze, and W.R. van Zwet. Resampling fewer than n observations: gains, losses, and remedies for losses. *Statistica Sinica*, 7:1–31, 1997.

[15] P. Billingsley. *Probability and Measure*. John Wiley and Sons, Inc., 1995.

[16] R. Blundell, A. Gosling, H. Ichimura, and C. Meghir. Changes in the distribution of male and female wages accounting for employment composition using bounds. *Econometrica*, 75(2):323–364, March 2007.

[17] A. J. Bowlus, N. M. Kiefer, and G. R. Neumann. Equilibrium search models and the transition from school to work. *International Economic Review*, 42(2):317–43, May 2001.

[18] M. Brown, J. Christiansen, and P. Philips. The decline of child labor in the u.s. fruit and vegetable canning industry: Law or economics? *The Business History Review*, 66(4):723–770, Winter 1992.

[19] F. A. Bugni, P. Hall, J. L. Horowitz, and G. R. Neumann. Goodness-of-fit test for functional data. *The Econometrics Journal (Forthcoming)*, 2008. Mimeo: Northwestern University, Australian National University and University of Iowa.

[20] K. Burdett and D.T. Mortensen. Wage differentials, employer size, and unemployment. *International Economic Review*, 39(2):257–273, 1998.

[21] I. Canay. E. L. inference for partially identified models: large deviations optimality and bootstrap validity. Mimeo: University of Wisconsin - Madison, January 2007.

[22] S.B. Carter and R. Sutch. Fixing the facts: Editing of the 1880 u.s. census of occupations with implications fro long-term labor force trends and the sociology of official statistics. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 29:5–24, 1996.

[23] V. Chernozhukov, H. Hong, and E. Tamer. Parameter set inference in a class of econometric models. *Econometrica*, 75(5):1243–1284, September 2007.

[24] F. Ciliberto and E. Tamer. Market structure and multiple equilibria in airline markets. Mimeo: University of Virginia and Northwestern University, February 2006.

[25] J.A. Cuenta-Albertos, E. del Barrio, R. Fraiman, and C. Matrán. The random projection method in goodness of fit for functional data. *Computational Statistics and Data Analysis*, 51(10):4814–4831, June 2007.

[26] Yu.A. Davydov, M.A. Lifshits, and N.V. Smorodina. *Local Properties of Distributions of Stochastic Functionals*, volume 173 of *Translations of Mathematical Monographs*. American Mathematical Society, 1995.

[27] E.V. Edmonds and N. Pavcnik. Child labor in the global economy. *Journal of Economic Perspectives*, 19(1):199–220, Winter 2005.

[28] A. Galichon and M. Henry. Dilation bootstrap: A natural approach to inference in incomplete models. Mimeo: Harvard University and Columbia University, November 2007.

[29] A. Galichon and M. Henry. Inference in incomplete models. Mimeo: Harvard University and Columbia University, May 2007.

[30] E. Giné and J. Zinn. Bootstrapping general empirical measures. *Annals of Probability*, 18(2):851–869, April 1990.

[31] C. Goldin. Household and market production of families in a late nineteenth century american city. *Explorations in Economic History*, 16(1):111–131, April 1979.

[32] P. Hall. *The Bootstrap and Edgeworth Expansion*. Springer, New York, 1992.

[33] P. Hall and M.A. Martin. On bootstrap resampling and iteration. *Biometrika*, 75(4):661–671, December 1988.

[34] J. H*ü*sler, R. Liu, and K. Singh. A formula for the tail probability of a multivariate normal distribution and its applications. *Journal of Multivariate Analysis*, 82(2):422–430, February 2002.

[35] J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–61, 1979.

[36] J.L. Horowitz. *The Bootstrap*, volume 5 of *Handbook of Econometrics*, chapter 52, pages 3159–3228. Elsevier Science B.V., May 2002.

[37] J.L. Horowitz and C.F. Manski. Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association*, 95(449):77–88, March 2000.

[38] G. Imbens and C.F. Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857, November 2004.

[39] J. Kuelbs. Kolmogorov's law of iterated logarithm for banach space valued random variables. *Illinois Journal of Mathematics*, 21(7):784–800, December 1977.

[40] C.F. Manski. Anatomy of the selection problem. *The Journal of Human Resources*, 24(3):343–360, Summer 1989.

[41] C.F. Manski. *Partial Identification of Probability Distributions*. Springer-Verlag, 2003.

[42] C.F. Manski and E. Tamer. Inference of regressions with interval data on a regressor or outcome. *Econometrica*, 70(2):519–546, March 2002.

[43] R.A. Margo and T.A. Finegan. Compulsory schooling legislation and school attendance in turn-of-the century america: A 'natural experiment' approach. *Economic Letters*, 53:103–110, October 1996.

[44] B. Meyer. Natural and quasi-experiments in economics. *Journal of Business & Economic Statistics*, 13(2):151–61, April 1995.

[45] C. Moehling. Work and family: Intergenerational support in american families. Phd. Dissertation: Northwestern University, 1996.

[46] C. Moehling. State child labor laws and the decline of child labor. *Explorations in Economic History*, 36:72–106, 1999.

[47] D.T. Mortensen. *Wage Dispersion: Why are similar workers paid differently?* MIT press, 2003.

[48] C. Nardinelli. Child labor and the factory acts. *Journal of Economic History*, 36:72–106, December 1980.

[49] W. Ogburn. Progress and uniformity in child-labor legislation: A study in statistical measurement. Phd. Dissertation: Columbia University, 1912.

[50] International Labor Organization. Every child counts: New global estimates on child labor. *Geneva: ILO*, 2002.

[51] P. Osterman. Education and labor markets at the turn of the century. *Politics & Society*, 34(1):103–122, 1979.

[52] A. Pakes, J. Porter, K. Ho, and J. Ishii. Moment inequalities and their application. Mimeo: Harvard and Wisconsin, April 2006.

[53] D.N. Politis and J.P. Romano. Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, 22(4):2031–2050, December 1994.

[54] D.N. Politis, J.P. Romano, and M. Wolf. *Subsampling*. Springer, 1999.

[55] J.P. Romano and A.M. Shaikh. Inference for identifiable parameters in partially identified econometric models. Mimeo: Stanford University and University of Chicago, November 2005.

[56] J.P. Romano and A.M. Shaikh. Inference for the identified set in partially identified econometric models. Mimeo: Stanford University and University of Chicago, September 2006.

[57] A. Rosen. An asymptotic refinement for bootstrap confidence intervals for bounds. Mimeo: Northwestern University, October 2006.

[58] A. Rosen. Confidence sets for partially identified parameters that satisfy a finite number of moment inequalities. Mimeo: University College London, June 2006.

[59] A.R. Sanderson. Child-labor legislation and the labor force participation of children. *The Journal of Economic History*, 34(1):297–299, March 1974.

[60] G. Soares. Inference for partially identified models with inequality moment constraints. Mimeo: Yale University, June 2006.

[61] J. Stoye. More on confidence intervals for partially identified parameters. Mimeo: New York University, August 2007.

[62] E. Tamer. Incomplete simultaneous discrete response model with multiple equilibria. *Review of Economic Studies*, 70(1):147–165, January 2003.

[63] A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

[64] J.M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, 2002.

APPENDIX A

Bootstrap Inference on Partially Identified Models

A.1. Notation

- Throughout this appendix, "a.s." abbreviates almost surely, "w.p.a.1" abbreviates with probability approaching one, "WLLN" and "SLLN" refer to the weak and strong law of large numbers, respectively, "CLT" refers to the central limit theorem and "LIL" refers to the law of iterated logarithm.
- For any $\theta \in \Theta$, we denote $\mathbb{P}(m_\theta) \equiv \mathbb{E}(m(Z,\theta))$, $\mathbb{P}_n(m_\theta) \equiv \mathbb{E}_n(m(Z,\theta)) \equiv n^{-1}\sum_{i=1}^{n} m(Z_i,\theta)$ and $v(m_\theta) \equiv \sqrt{n}(\mathbb{P}_n - \mathbb{P})(m_\theta)$. For any $\theta \in \Theta$ and $j = 1,...,J$, $\mathbb{P}(m_{j,\theta}) \equiv \mathbb{E}(m_j(Z,\theta))$, $\mathbb{P}_n(m_{j,\theta}) \equiv n^{-1}\sum_{i=1}^{n} m_j(Z_i,\theta)$ and $v(m_{j,\theta}) \equiv \sqrt{n}(\mathbb{P}_n - \mathbb{P})(m_{j,\theta})$.
- We refer to the space of bounded that map Θ into $\mathbb{R}^J$ as $l_J^\infty(\Theta)$ and the space of bounded continuous functions that map Θ into $\mathbb{R}^J$ as $C_J(\Theta)$. For both spaces, we use the uniform metric, denoted $\|y\|_\infty$, i.e. $\forall y \in l_J^\infty(\Theta)$, $\|y\|_\infty \equiv \sup_{\theta\in\Theta}\|y(\theta)\|$. For matrix spaces, we use the Frobenius norm, i.e. $\forall M \in \mathbb{R}^{m\times n}$, $\|M\| \equiv \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} M_{[i,j]}^2}$.
- For any $s \in \mathbb{N}$, the space of Borel measurable convex sets in $\mathbb{R}^s$ is denoted by $\mathcal{C}_s$. For any function $H : A_1 \to A_2$, and any set $S \subset A_2$, $H^{-1}(S) \equiv \{x \in A_1 : H(x) \in S\}$. Also, for any $\varepsilon > 0$ and any set $S \subset \mathbb{R}^s$, define $S^\varepsilon \equiv$

$\{x \in \mathbb{R}^s : \exists x' \in S \cap \|x - x'\| \leq \varepsilon\}$. Finally, for any set $S \subset \mathbb{R}^s$, $Int(S)$ denotes interior of S and ∂S denotes boundary of S.

- For any square matrix $\Sigma \in \mathbb{R}^{r \times r}$ and any Borel measurable set A, $\Phi_\Sigma(A)$ denotes $P(Z \in A)$ where $Z \sim N(0, \Sigma)$. If Σ is non-singular, $\phi_\Sigma(x)$ denotes the density of Z, where $Z \sim N(0, \Sigma)$.

A.2. Confidence set for the identified set

A.2.1. Differences with the naive bootstrap

The bootstrap procedure we propose to cover the identified set differs qualitatively from replacing the subsampling scheme in CHT [**23**] with the traditional bootstrap. In the latter, we would construct the following criterion function,

$$Q_n^{*,AP}(\theta) = G\left(\left\{[\mathbb{E}_n^*(m_j(Z,\theta))]_+\right\}_{j=1}^J\right)$$

We denote this function by $Q_n^{*,AP}(\theta)$ since it represents the *analogy principle* criterion function for the bootstrap sample. Given that this procedure could be naively suggested based on the subsampling scheme from CHT [**23**], we refer to it as the *naive bootstrap*.

When we combine this analogy principle criterion function with subsampling, we produce consistent inference. When we combine this analogy principle criterion function with the bootstrap, we obtain two problems which, in general, will result in inconsistent inference in level. The first problem is caused by the estimation of the identified set. Recall from section 1.4.2.2 that we estimate the identified set by allowing the sample restrictions to be violated by a certain amount that converges to zero at a suitable rate. In other words, we estimate the identified set by artificially expanding the sample analogue by a certain amount. The introduction of this expansion will not be (asymptotically) problematic for inference based on subsampling but will generate inconsistent inference based on the bootstrap. We will refer to this problem as *the expansion problem.* The second problem is directly related to the well-known *inconsistency of the bootstrap in the boundary of the parameter space.* This problem was studied by Andrews [**2**] who suggests using subsampling as one possible solutions to this problem.

In order to understand the nature of these problems, we provide two examples. These examples show three things. First, replacing the subsampling procedure in CHT [**23**] with the bootstrap does not result in consistent inference in level. Second, our proposed bootstrap procedure corrects these inconsistencies. Finally, these problems are not present in the subsampling procedure.

A.2.1.1. Problem 1: the expansion problem. The objective is to construct a confidence set for the following identified set,

$$\Theta_I = \{\theta \in \Theta : \mathbb{E}\left(Y_1\right) \leq \theta \leq \mathbb{E}\left(Y_2\right)\}$$

where $\mathbb{E}\left(Y_1\right) = \mathbb{E}\left(Y_2\right) = 0$. Note that the identified set is composed of only one point, that is, the parameter is point identified. Suppose that the sample $\{Y_{1,i}, Y_{2,i}\}_{i=1}^{n}$ is i.i.d. such that for every $i = 1, 2, ..., n$, $(Y_{1,i}, Y_{2,i}) \sim N\left(0, \mathbf{I}_2\right)$, where $\mathbf{I}_k$ denotes the $k \times k$ identity matrix. Notice that all assumptions of the conditionally separable model are satisfied. Since $\Theta_I = \{0\}$, the distribution of interest is given by: $\Gamma_n = \sqrt{n} Q_n\left(0\right)$, which, by the central limit theorem, converges weakly to $G\left(\left[\zeta_1\right]_+, \left[\zeta_2\right]_+\right)$, where $\zeta \sim N\left(0, \mathbf{I}_2\right)$.

Now consider estimation of the identified set. The key feature of this setup is that even though the identified set is non-empty (because $\mathbb{E}\left(Y_1\right) \leq \mathbb{E}\left(Y_2\right)$), the sample analogue estimate of the identified set, given by: $\hat{\Theta}_I^{AP} = \{\theta \in \Theta : \mathbb{E}_n(Y_1) \leq \theta \leq \mathbb{E}_n(Y_2)\}$, is empty with positive probability (in this case, with probability 0.5). Hence, using the estimator $\hat{\Theta}_I^{AP}$ for inference will not produce consistent inference. This illustrates why we need to introduce the sequence $\{\tau_n/\sqrt{n}\}_{n=1}^{+\infty}$ to estimate the identified set. Our estimator for the identified set is given by,

$$\hat{\Theta}_I(\tau_n) = \left\{\theta \in \Theta : \mathbb{E}_n(Y_1) - \tau_n/\sqrt{n} \leq \theta \leq \mathbb{E}_n(Y_2) + \tau_n/\sqrt{n}\right\}$$

By the requirements on the sequence $\{\tau_n\}_{n=1}^{+\infty}$, this estimator will eventually be non-empty, almost surely.

Consider performing inference combining the bootstrap and the analogy principle criterion function, that is, the naive bootstrap. In this setting, we will obtain the following statistic,

$$\Gamma_n^{*,AP} = 1\left[\hat{\Theta}_I(\tau_n) = \varnothing\right] \max \left\{ \begin{array}{c} G\left(\begin{array}{c} \left[\sqrt{b_n}\left(\mathbb{E}_n^*(Y_1) - \mathbb{E}_n(Y_1)\right) + \tau_n\right]_+, \\ \left[\sqrt{b_n}\left(\mathbb{E}_n(Y_1) - \mathbb{E}_n^*(Y_2)\right) - \tau_n\right]_+ \end{array} \right), \\ G\left(\begin{array}{c} \left[\sqrt{b_n}\left(\mathbb{E}_n^*(Y_1) - \mathbb{E}_n(Y_2)\right) - \tau_n\right]_+, \\ \left[\sqrt{b_n}\left(\mathbb{E}_n(Y_2) - \mathbb{E}_n^*(Y_2)\right) + \tau_n\right]_+ \end{array} \right) \end{array} \right\}$$

We now show that for a set of probability measure one, the conditional distribution of the right hand side diverges to infinity. For any $\varepsilon > 0$, consider the following events,

$$\begin{aligned} A &= \left\{\left\{\sqrt{n}\left(\mathbb{E}_n^*(Y_1) - \mathbb{E}_n(Y_1), \mathbb{E}_n^*(Y_2) - \mathbb{E}_n(Y_2)\right) | \mathcal{X}_n\right\} \xrightarrow{d} N(0, \mathbf{I}_2)\right\} \\ B &= \liminf \left\{\left\{\hat{\Theta}_I(\tau_n) = \varnothing\right\} \cap \left\{\left|\sqrt{n}\left(\mathbb{E}_n(Y_1) - \mathbb{E}_n(Y_2)\right)\right| \leq \tau_n/2\right\}\right\} \end{aligned}$$

Let $\omega \in \{A \cap B\}$. Since $\omega \in B$, $\exists N \in \mathbb{N}$ such $\forall n \geq N$,

$$\Gamma_n^{*,AP} \geq \max \left\{ \begin{array}{c} G\left(\left[\sqrt{n}\left(\mathbb{E}_n^*\left(Y_1\right) - \mathbb{E}_n\left(Y_1\right)\right) + \tau_n\right]_+, \left[\sqrt{n}\left(\mathbb{E}_n\left(Y_2\right) - \mathbb{E}_n^*\left(Y_2\right)\right) - 1.5\tau_n\right]_+\right), \\ G\left(\left[\sqrt{n}\left(\mathbb{E}_n^*\left(Y_1\right) - \mathbb{E}_n\left(Y_1\right)\right) - 1.5\tau_n\right]_+, \left[\sqrt{n}\left(\mathbb{E}_n\left(Y_2\right) - \mathbb{E}_n^*\left(Y_2\right)\right) + \tau_n\right]_+\right) \end{array} \right\}$$

Since $\omega \in A$, the conditional distribution of the right hand side diverges to infinity, a.s.. By the LIL and the requirements on $\{\tau_n\}_{n=1}^{+\infty}$, $P(A) = 1$ (see proof of lemma 5). By theorem 2.1 in Bickel and Freedman [**13**], $P(B) = 1$. Hence, $P(A \cap B) = 1$ and the naive bootstrap is inconsistent in level, a.s. Hence, inference based on $\Gamma_n^{*,AP}$ will almost surely result in 100% coverage.

The intuition for this result is as follows. The estimation of the identified set requires the introduction of the sequence $\{\tau_n\}_{n=1}^{+\infty}$, which enters directly into the $[\cdot]_+$ term of the bootstrap version of the analogy principle criterion function. Since this term diverges to infinity, it is not surprising that the statistic also diverges to infinity. As we show next, our proposed criterion function corrects this problem by removing this term from the $[\cdot]_+$ term.

If we choose to use our proposed bootstrap method, we have the following statistic,

$$\begin{aligned} \Gamma_n^* &= 1\left[\hat{\Theta}_I\left(\tau_n\right) \neq \varnothing\right] * \\ &* \max_{\theta \in \hat{\Theta}_I} \left\{ G\left(\begin{array}{c} \left[\sqrt{n}\left(\mathbb{E}_n^*\left(Y_1\right) - \mathbb{E}_n\left(Y_1\right)\right)\right]_+ 1\left[\mathbb{E}_n\left(Y_1\right) - \theta \geq -\tau_n/\sqrt{n}\right], \\ \left[\sqrt{n}\left(\mathbb{E}_n\left(Y_2\right) - \mathbb{E}_n^*\left(Y_2\right)\right)\right]_+ 1\left[\theta - \mathbb{E}_n\left(Y_2\right) \geq -\tau_n/\sqrt{n}\right] \end{array} \right) \right\} \end{aligned}$$

Consider $\omega \in \{A \cap B'\}$ where B' is defined by,

$$B' = \liminf \left\{ \begin{array}{c} \left\{\{0\} \in \hat{\Theta}_I\left(\tau_n\right)\right\} \cap \left\{\sqrt{n}\mathbb{E}_n\left(Y_1\right) \geq -\tau_n\right\} \cap \\ \cap \left\{\sqrt{n}\mathbb{E}_n\left(Y_2\right) \leq -\tau_n\right\} \end{array} \right\}$$

Since $\omega \in B'$, $\exists N \in \mathbb{N}$, such that $\forall n \geq N$,

$$\Gamma_n^* = G\left(\left[\sqrt{n}\left(\mathbb{E}_n^*\left(Y_1\right) - \mathbb{E}_n\left(Y_1\right)\right)\right]_+, \left[\sqrt{n}\left(\mathbb{E}_n\left(Y_2\right) - \mathbb{E}_n^*\left(Y_2\right)\right)\right]_+\right)$$

Since $\omega \in A$, the conditional distribution of the right hand side converges weakly to $G\left(\left[\zeta_1\right]_+, \left[\zeta_2\right]_+\right)$ where $\zeta \sim N\left(0, \mathbf{I}_2\right)$, a.s.. By the same arguments as before, $P\left(A' \cap B\right) = 1$ and, thus, our proposed criterion function leads to consistent inference in level.

It is important to understand that the inconsistency of the naive bootstrap is a consequence of combining the analogy principle criterion function with the bootstrap rather than with subsampling. If the analogy principle criterion function is applied to subsampling samples (with subsampling size b_n) we obtain the following statistic,

$$\Gamma_{b_n,n}^{SS,AP} = 1\left[\hat{\Theta}_I\left(\tau_n\right) = \varnothing\right] \max \left\{ \begin{array}{c} G\left(\begin{array}{c} \left[\sqrt{b_n}\left(\mathbb{E}_{b_n,n}^{SS}\left(Y_1\right) - \mathbb{E}_n\left(Y_1\right)\right) + \frac{\sqrt{b_n}\tau_n}{\sqrt{n}}\right]_+, \\ \left[\sqrt{b_n}\left(\mathbb{E}_n\left(Y_1\right) - \mathbb{E}_{b_n,n}^{SS}\left(Y_2\right)\right) - \frac{\sqrt{b_n}\tau_n}{\sqrt{n}}\right]_+ \end{array} \right), \\ G\left(\begin{array}{c} \left[\sqrt{b_n}\left(\mathbb{E}_{b_n,n}^{SS}\left(Y_1\right) - \mathbb{E}_n\left(Y_2\right)\right) - \frac{\sqrt{b_n}\tau_n}{\sqrt{n}}\right]_+, \\ \left[\sqrt{b_n}\left(\mathbb{E}_n\left(Y_2\right) - \mathbb{E}_{b_n,n}^{SS}\left(Y_2\right)\right) + \frac{\sqrt{b_n}\tau_n}{\sqrt{n}}\right]_+ \end{array} \right) \end{array} \right\}$$

where for $i = 1, 2$, $\mathbb{E}_{b_n,n}^{SS}\left(Y_i\right)$ denotes the sample average of Y_i in the subsample.

For any $\varepsilon > 0$, consider the following events,

$$B'' = \liminf \left\{ \begin{array}{c} \left\{ \{0\} \in \hat{\Theta}_I(\tau_n) \right\} \cap \\ \cap \left\{ \left| \sqrt{b_n} \left(\mathbb{E}_n(Y_1) - \mathbb{E}_n(Y_2) \right) \right| \leq (1+\varepsilon) \sqrt{2(b_n \ln\ln n)/n} \right\} \end{array} \right\}$$

Consider $\omega \in \{A \cap B''\}$. If the sequence $\{\tau_n\}_{n=1}^{+\infty}$ is chosen such that $b_n^{1/2}\tau_n n^{-1/2} = o(1)$ a.s., then, conditionally on the sample, $\Gamma_{b_n,n}^{SS,AP}$ converges weakly to $G\left([\zeta_1]_+, [\zeta_2]_+\right)$ where $\zeta \sim N(0, \mathbf{I}_2)$. By usual arguments, $P(A'' \cap B) = 1$. Hence, subsampling results in consistent inference. Just like with the bootstrap, the estimation of the identified set introduces the sequence $\left\{\sqrt{b_n}\tau_n/\sqrt{n}\right\}_{n=1}^{+\infty}$ into the $[\cdot]_+$ term of the statistic. The key difference with the bootstrap is that this sequence converges to zero (instead of diverging to infinity), so it does not affect the asymptotic distribution. Notice that even though the sequence $\left\{\tau_n\sqrt{b_n}/\sqrt{n}\right\}_{n=1}^{+\infty}$ may converge to zero, it may represent a non-negligible number in small samples. As a consequence, it would not be surprising that subsampling might still exhibit some overcoverage.

A.2.1.2. Problem 2: boundary problem. In order to isolate this problem from the previous one, we consider an example where we can estimate the identified set without the need of introducing any expansion (that is, we can set $\tau_n = 0$ in $\hat{\Theta}_I(\tau_n)$).

Consider the following identified set,

$$\Theta_I = \{\theta \in \Theta : \max\{\mathbb{E}(Y_1), \mathbb{E}(Y_2)\} \leq \theta\}$$

where $\mathbb{E}(Y_1) = \mathbb{E}(Y_2) = 0$. In order to perform inference, we have an i.i.d. sample of $\{Y_{1,i}, Y_{2,i}\}_{i=1}^n$. For concreteness, assume that for every $i = 1, 2, ..., n$, $(Y_{1,i}, Y_{2,i}) \sim N(0, \mathbf{I}_2)$.

Notice that all assumptions of the conditionally separable model are satisfied. The distribution of interest is given by: $\Gamma_n = \sup_{\theta\in\Theta_I}\sqrt{n}Q_n(\theta)$, which, by the central limit theorem, converges weakly to $G\left([\zeta_1]_+, [\zeta_2]_+\right)$, where $\zeta \sim N(0, \mathbf{I}_2)$.

As opposed to the previous example, the identified set has non-empty interior and the sample analogue estimate will always be non-empty. Hence, we can estimate the identified set with the analogy principle estimate, $\hat{\Theta}_I(0) = \{\theta \in \Theta : \max\{\mathbb{E}_n(Y_1), \mathbb{E}_n(Y_2)\} \leq \theta\}$.

Now consider performing inference combining the bootstrap and the analogy principle criterion function, that is, the naive bootstrap. For any constant $c > 0$, consider the following events,

$$\begin{aligned} A &= \left\{\left\{\sqrt{n}\left(\mathbb{E}_n^*(Y_1) - \mathbb{E}_n(Y_1), \mathbb{E}_n^*(Y_2) - \mathbb{E}_n(Y_2)\right)|\mathcal{X}_n\right\} \xrightarrow{d} N(0, \mathbf{I}_2)\right\} \\ B &= \limsup\left\{\left\{\hat{\Theta}_I(0) \neq \varnothing\right\} \cap \left\{\sqrt{n}\left(\mathbb{E}_n(Y_1) - \mathbb{E}_n(Y_2)\right) < -c\right\}\right\} \end{aligned}$$

Suppose that $\omega \in \{A \cap B\}$. Since $\omega \in B$, there exists a subsequence $\{n_k\}_{k=1}^{+\infty}$ such that, along this subsequence, $\hat{\Theta}_I(0)$ is non-empty and $\sqrt{n_k}\left(\mathbb{E}_{n_k}(Y_1) - \mathbb{E}_{n_k}(Y_2)\right) < -c$. Along this subsequence,

$$\Gamma_{n_k}^{*,AP} \leq G\left(\left[\sqrt{n}\left(\mathbb{E}_{n_k}^*(Y_1) - \mathbb{E}_{n_k}(Y_1)\right) - c\right]_+, \left[\sqrt{n}\left(\mathbb{E}_{n_k}^*(Y_2) - \mathbb{E}_{n_k}(Y_2)\right)\right]_+\right)$$

and since $\omega \in A$, then the right hand side converges weakly to $G\left([\zeta_1 - c]_+, [\zeta_2]_+\right)$, where $\zeta \sim N(0, \mathbf{I}_2)$. By usual arguments, $P(A \cap B) = 1$. This implies that the naive bootstrap is not consistent in level. One may relate this result with the inconsistency of the bootstrap on the boundary of the parameter space (see, Andrews [**2**]). The boundaries of the unknown identified set are determined by the parameters $\mathbb{E}(Y_1)$ and $\mathbb{E}(Y_2)$, which happen to coincide. In any bootstrap sample, the boundaries of the sample identified set are determined by $\mathbb{E}_n(Y_1)$ and $\mathbb{E}_n(Y_2)$, which will almost never coincide. As a consequence, the structure of the boundaries in the population and in the sample almost surely differ, producing inconsistency of the naive bootstrap.

Now suppose that we choose to use our proposed bootstrap procedure. Assume that $\omega \in \{A \cap B'\}$ where B' is the following event,

$$B' = \liminf \left\{ \left\{ 0 \in \hat{\Theta}_I(0) \right\} \cap \left\{ \sqrt{n}\mathbb{E}_n(Y_1) \geq -\tau_n \right\} \cap \left\{ \sqrt{n}\mathbb{E}_n(Y_2) \geq -\tau_n \right\} \right\}$$

Since $\omega \in B'$, $\exists N \in \mathbb{N}$, such that $\forall n \geq N$, $0 \in \hat{\Theta}_I(0)$, $\sqrt{n}\,\mathbb{E}_n(Y_1) \geq -\tau_n$ and $\sqrt{n}\,\mathbb{E}_n(Y_2) \geq -\tau_n$. Thus, for any $\omega \in B'$, conditional on the sample and $\forall n \geq N$,

$$\Gamma_n^* = G\left(\left[\sqrt{n}\left(\mathbb{E}_n^*(Y_1) - \mathbb{E}_n(Y_1)\right)\right]_+, \left[\sqrt{n}\left(\mathbb{E}_n^*(Y_2) - \mathbb{E}_n(Y_2)\right)\right]_+ \right)$$

Since $\omega \in A$, the conditional distribution of the right hand side converges weakly to $G\left([\zeta_1]_+, [\zeta_2]_+\right)$, where $\zeta \sim N(0, \mathbf{I}_2)$. By usual arguments $P(A \cap B') = 1$ and so, our bootstrap is consistent in level.

Exactly as in the previous example, we note that the inconsistency of the naive bootstrap is a consequence of using bootstrap instead of subsampling. When we use subsampling (with sample size b_n) with the analogue principle criterion function, we obtain the following statistic,

$$\begin{aligned}\Gamma_{b_n,n}^{SS,AP} &= 1\left[\hat{\Theta}_I\left(0\right)\neq\varnothing\right]* \\ &*\left\{\begin{array}{c} G\left(\begin{array}{c}\left[\sqrt{b_n}\left(\mathbb{E}_{b_n,n}^{SS}\left(Y_1\right)-\mathbb{E}_n\left(Y_1\right)\right)\right]_+, \\ \left[\sqrt{b_n}\left(\mathbb{E}_{b_n,n}^{SS}\left(Y_2\right)-\mathbb{E}_n\left(Y_1\right)\right)\right]_+\end{array}\right)1\left[\mathbb{E}_n\left(Y_1\right)\geq\mathbb{E}_n\left(Y_2\right)\right]+ \\ +G\left(\begin{array}{c}\left[\sqrt{b_n}\left(\mathbb{E}_{b_n,n}^{SS}\left(Y_1\right)-\mathbb{E}_n\left(Y_2\right)\right)\right]_+, \\ \left[\sqrt{b_n}\left(\mathbb{E}_{b_n,n}^{SS}\left(Y_2\right)-\mathbb{E}_n\left(Y_2\right)\right)\right]_+\end{array}\right)1\left[\mathbb{E}_n\left(Y_1\right)<\mathbb{E}_n\left(Y_2\right)\right]\end{array}\right\}\end{aligned}$$

Let $\omega\in\{A\cap B''\}$, where B'' is given by,

$$B''=\liminf\left\{\left\{0\in\hat{\Theta}_I\left(0\right)\right\}\cap\left\{\left|\sqrt{b_n}\left(\mathbb{E}_n\left(Y_1\right)-\mathbb{E}_n\left(Y_2\right)\right)\right|\leq\left(1+\varepsilon\right)\sqrt{2\left(b_n\ln\ln n\right)/n}\right\}\right\}$$

Since $\left(b_n\ln\ln n\right)/n=o\left(1\right)$ and applying previous arguments, $\Gamma_{n,b_n}^{SS,AP}$ converges weakly to $G\left(\left[\zeta_1\right]_+,\left[\zeta_2\right]_+\right)$, where $\zeta\sim N\left(0,\mathbf{I}_2\right)$. Since $P\left(A\cap B''\right)=1$, we deduce subsampling generates consistent inference in level.

A.2.2. Preliminary results

Proof of lemma 4. This proof is elementary and is therefore omitted. □

Proof of lemma 5. Part 1. The definition of Θ_I implies the following sequence of inequalities,

$$\sup_{\theta\in\Theta_I}\max_{j=1:J}\mathbb{P}_n\left(m_{j,\theta}\right)\leq\left\{\begin{array}{c}\sup_{\theta\in\Theta_I}\max_{j=1:J}\left(\mathbb{P}_n-\mathbb{P}\right)\left(m_{j,\theta}\right)+ \\ +\sup_{\theta\in\Theta_I}\max_{j=1:J}\mathbb{P}\left(m_{j,\theta}\right)\end{array}\right\}\leq n^{-1/2}\sup_{\theta\in\Theta}\max_{j=1:J}v_n\left(m_{j,\theta}\right)$$

Therefore, the event $\{\sup_{\theta\in\Theta}\max_{j=1:J} v_n(m_{j,\theta}) \leq \tau_n\}$ implies the event $\left\{\Theta_I \subseteq \hat{\Theta}_I(\tau_n)\right\}$ and so,

$$P\left(\liminf\left\{\Theta_I \subseteq \hat{\Theta}_I(\tau_n)\right\}\right) \geq \sum_{j=1}^{J} P\left(\liminf\left\{\sup_{\theta\in\Theta}|v_n(m_{j,\theta})| \leq \tau_n\right\}\right) - J + 1$$

Since $\sqrt{\ln\ln n}/\tau_n = o(1)$, by the LIL for empirical processes (see, e.g., Kuelbs [**39**]), the right hand side expression is equal to one. The definition of $\hat{\Theta}_I(\tau_n)$ implies the following sequence of inequalities,

$$\begin{aligned}\sup_{\theta\in\hat{\Theta}_I(\tau_n)}\max_{j=1:J}\mathbb{P}(m_{j,\theta}) &\leq \sup_{\theta\in\hat{\Theta}_I(\tau_n)}\max_{j=1:J}(\mathbb{P}-\mathbb{P}_n)(m_{j,\theta}) + \sup_{\theta\in\hat{\Theta}_I(\tau_n)}\max_{j=1:J}\mathbb{P}_n(m_{j,\theta}) \\ &\leq n^{-1/2}\left(\tau_n - \inf_{\theta\in\Theta}\min_{j=1:J} v_n(m_{j,\theta})\right)\end{aligned}$$

Therefore, the event $\{\inf_{\theta\in\Theta}\min_{j=1:J} v_n(m_{j,\theta}) \geq -\tau_n\}$ and $2\tau_n/\varepsilon_n = o(1)$ implies the event $\left\{\hat{\Theta}_I(\tau_n) \subseteq \Theta_I(\varepsilon_n)\right\}$ and so,

$$P\left(\liminf\left\{\hat{\Theta}_I(\tau_n) \subseteq \Theta_I(\varepsilon_n)\right\}\right) \geq \sum_{j=1}^{J} P\left(\liminf\left\{\sup_{\theta\in\Theta}|v_n(m_{j,\theta})| \leq \tau_n\right\}\right) - J + 1$$

and for the same reasons as before, the right hand side expression is one. Elementary properties of lim inf operator complete the proof.

Part 2. Since the function $\mathbb{P}(m_\theta) : \Theta \to \mathbb{R}^J$ is lower-semi continuous and the set Θ is compact, $\max_{j=1:J}\mathbb{P}(m_{j,\theta})$ achieves a minimum on Θ. Since $\Theta_I = \varnothing$, such minimum value is a positive value $\varpi > 0$, and so,

$$\inf_{\theta\in\Theta}\max_{j=1:J}\mathbb{P}_n(m_{j,\theta}) \geq \left\{\begin{array}{c}\inf_{\theta\in\Theta}\min_{j=1:J}(\mathbb{P}_n-\mathbb{P})(m_{j,\theta}) + \\ +\inf_{\theta\in\Theta}\max_{j=1:J}\mathbb{P}(m_{j,\theta})\end{array}\right\} \geq n^{-1/2}\inf_{\theta\in\Theta}\min_{j=1:J} v_n(m_{j,\theta}) + \varpi$$

Therefore, the event $\{\inf_{\theta\in\Theta}\min_{j=1:J} v_n(m_{j,\theta}) \geq -\tau_n\}$ implies the event $\left\{\hat{\Theta}_I(\tau_n) = \varnothing\right\}$ and hence,

$$P\left(\liminf\left\{\hat{\Theta}_I(\tau_n) = \varnothing\right\}\right) \geq \sum_{j=1}^{J} P\left(\liminf\left\{\sup_{\theta\in\Theta}|v_n(m_{j,\theta})| \leq \tau_n\right\}\right) - J + 1$$

and for the same reasons as before, the right hand side expression is one. □

A.2.3. Representation results

Theorem 36. *Assume (A1)-(A4), (CF') and that* $\Theta_I \neq \varnothing$. *Then,* $\Gamma_n = H(v_n(m_\theta)) + \delta_n$, *where,*

(1) *for any* $\varepsilon > 0$, $\lim_{n\to\infty} P^*(|\delta_n| > \varepsilon) = 0$,

(2) $v_n(m_\theta) : \Omega_n \to l_J^\infty(\Theta)$ *is an empirical process that converges weakly to a tight zero-mean Gaussian process, denoted* ζ, *with covariance function,*

$$\Sigma(\theta_1, \theta_2) = \mathbb{E}\left[(m(Z,\theta_1) - \mathbb{E}(m(Z,\theta_1)))(m(Z,\theta_2) - \mathbb{E}(m(Z,\theta_2)))'\right]$$

for each $\{\theta_1, \theta_2\} \in \Theta$,

(3) $H : l_J^\infty(\Theta) \to \mathbb{R}$ *is continuous, non-negative, weakly convex and* $H(y) = 0$ *implies that for some* $\theta_0 \in \Theta$ *and for some* $j = 1, 2, ..., J$, $y_j(\theta_0) \leq 0$.

If we assume (B1)-(B4), (CF) and that $\Theta_I \neq \varnothing$, *then, for some* $r \in \mathbb{N} \cap [1, J \times K]$, $\Gamma_n = \tilde{H}\left(\sqrt{n}\bar{Z}\right) + \tilde{\delta}_n$, *where,*

(1) *for any* $\varepsilon_n = O\left(n^{-1/2}\right)$, $P\left(\left|\tilde{\delta}_n\right| > \varepsilon_n\right) = o\left(n^{-1/2}\right)$,

(2) $\bar{Z} : \Omega_n \to \mathbb{R}^r$ *is a zero mean sample average of* n *i.i.d. observations from a distribution with variance-covariance matrix is* $\mathbf{I}_r$*. If we add (B5), this distribution has finite third absolute moments,*

(3) $\tilde{H} : \mathbb{R}^r \to \mathbb{R}$ *is continuous, non-negative and weakly convex. For all* $\mu > 0$*, any* $|h| \geq \mu > 0$ *and any sequence* $\varepsilon_n = o(1)$, $\tilde{H}^{-1}((h - \varepsilon_n, h + \varepsilon_n]) \subseteq \tilde{H}^{-1}(\{h\})^{\eta_n}$ *where* $\eta_n = O(\varepsilon_n)$*. Finally,* $\tilde{H}(y) = 0$ *implies that for some non-zero vector* $b \in \mathbb{R}^r$, $b'y \leq 0$.

If, instead, we assume (A1)-(A4), (CF') and that $\Theta_I = \varnothing$*, then* $\Gamma_n = 0$*.*

Proof. Part 1. Let δ_n be defined as,

$$\delta_n = \sup_{\theta \in \Theta_I} G\left(\left\{\left[\sqrt{n}\mathbb{P}_n(m_{j,\theta})\right]_+\right\}_{j=1}^J\right) - \sup_{\theta \in \Theta_I} G\left(\left\{\left[v_{j,n}(m_\theta)\right]_+ 1\left[\mathbb{P}(m_{j,\theta}) = 0\right]\right\}_{j=1}^J\right)$$

and set $H(y) = \sup_{\theta \in \Theta_I} G\left(\left\{[y_j]_+ 1\left[\mathbb{P}(m_{j,\theta}) = 0\right]\right\}_{j=1}^J\right)$.

Point 1. Restrict attention to $\theta \in \Theta_I$ and fix $\varepsilon > 0$ arbitrarily. By definition of $v_n(m_\theta)$, $\delta_n \geq 0$ and so it suffices to show $P^*(\delta_n > \varepsilon) = o(1)$. For any positive sequence $\{\varepsilon_n\}_{n=1}^{+\infty}$ such that $\sqrt{\ln \ln n}/\varepsilon_n = o(1)$ and $\varepsilon_n/\sqrt{n} = o(1)$, denote $A_n = \left\{\sup_{\theta \in \Theta_I} \|v_n(m_\theta)\| \leq \varepsilon_n\right\}$. By the LIL for empirical processes (see, e.g. Kuelbs [**39**]), $P(A_n^c) = o(1)$ and so, it suffices to show $P^*(\delta_n > \varepsilon \cap A_n) = o(1)$.

Denote: $G_{n,1}(\theta) = G\left(\left\{[\sqrt{n}\mathbb{P}_n(m_{j,\theta})]_+\right\}_{j=1}^J\right)$, $\bar{G}_{n,1} = \sup_{\theta \in \Theta_I} G_{n,1}(\theta)$, $G_{n,2}(\theta) = G\left(\left\{[v_n(m_{j,\theta})]_+ 1\left[\mathbb{P}(m_{j,\theta}) = 0\right]\right\}_{j=1}^J\right)$ and $\bar{G}_{n,2} = \sup_{\theta \in \Theta_I} G_{n,2}(\theta)$.

By definition of supremum, for every $\varepsilon > 0$, implies that $\exists \theta \in \Theta_I$ so that $G_{n,1}(\theta) + \varepsilon/2 \geq \bar{G}_{n,1}$ and so, the event $\{\delta_n > \varepsilon \cap A_n\}$ is equivalent to,

$$\left\{\{\delta_n > \varepsilon\} \cap \left\{\exists \theta \in \Theta_I : \left\{G_{n,1}(\theta) + \varepsilon/2 \geq \bar{G}_{n,1}\right\} \cap \left\{G_{n,1}(\theta) - G_{n,2}(\theta) \geq \varepsilon/2\right\}\right\} \cap A_n\right\}$$

The event $\{\exists\theta \in \Theta_I : \{G_{n,1}(\theta) - G_{n,2}(\theta) \geq \varepsilon/2\}\}$ implies $\exists j \in \{1, 2, ..., J\}$ such that $\mathbb{P}_n(m_{j,\theta}) \geq 0$ and $\mathbb{P}(m_{j,\theta}) < 0$. Let $\left\{\mathcal{P}^{\{1,2,..,J\}}/\varnothing\right\}$ denote the set of all non-empty subsets of $\{1, 2, .., J\}$ and for any $S \in \left\{\mathcal{P}^{\{1,2,..,J\}}/\varnothing\right\}$, let $\bar{S}$ denote $\left\{\mathcal{P}^{\{1,2,..,J\}}/\varnothing\right\}/S$. For any $S \in \left\{\mathcal{P}^{\{1,2,..,J\}}/\varnothing\right\}$, consider the random set $D_n(S)$ be given by,

$$D_n(S) = \left\{\Theta_I \cap \bigcap_{j\in S} \{\mathbb{P}_n(m_{j,\theta}) \geq 0\}\right\}$$

Therefore, $\{\exists\theta \in \Theta_I : \{G_{n,1}(\theta) - G_{n,2}(\theta) \geq \varepsilon/2\}\}$ implies $\bigcup_{S\in\left\{\mathcal{P}^{\{1,2,..,J\}}/\varnothing\right\}} \{\exists\theta \in D_n(S)\}$.

For any $S \in \left\{\mathcal{P}^{\{1,2,..,J\}}/\varnothing\right\}$ and $n \in \mathbb{N}$, define the following two non-random sets,

$$\tilde{D}_n(S) = \left\{\Theta_I \cap \bigcap_{j\in S} \left\{\mathbb{P}(m_{j,\theta}) \in \left[-\varepsilon_n/\sqrt{n}, 0\right]\right\}\right\} \qquad D(S) = \left\{\Theta_I \cap \bigcap_{j\in S} \{\mathbb{P}(m_{j,\theta}) = 0\}\right\}$$

For any $S \in \left\{\mathcal{P}^{\{1,2,..,J\}}/\varnothing\right\}$, $\{\{\exists\theta \in D_n(S)\} \cap A_n\}$ implies $\left\{\left\{\exists\theta \in \tilde{D}_n(S)\right\} \cap A_n\right\}$. Also, $\forall S \in \left\{\mathcal{P}^{\{1,2,..,J\}}/\varnothing\right\}$, $\lim \tilde{D}_n(S) = \cap_{n\in\mathbb{N}}\tilde{D}_n(S) = D(S)$, which implies that $\forall\eta > 0$, $\exists N \in \mathbb{N}$ such that $\forall n \geq N$, $\left\{\exists\theta \in \tilde{D}_n(S)\right\}$ implies $\{\exists\theta' \in D(S) : \|\theta - \theta'\| < \eta\}$. It follows that $\forall\eta > 0$, $\exists N \in \mathbb{N}$ such that $\forall n \geq N$, $\{\delta_n > \varepsilon \cap A_n\}$ is equivalent to the event,

$$\bigcup_{S\in\left\{\mathcal{P}^{\{1,2,..,J\}}/\varnothing\right\}} \left\{\{\delta_n > \varepsilon \cap A_n\} \cap \left\{\begin{array}{c} \exists(\theta, \theta') \in \{D_n(S) \times D(S)\} : \\ \|\theta - \theta'\| \leq \eta \cap \left\{G_{n,1}(\theta) + \varepsilon/2 \geq \bar{G}_{n,1}\right\} \end{array}\right\}\right\}$$

Now, $\forall\eta > 0$, $\forall S \in \left\{\mathcal{P}^{\{1,2,..,J\}}/\varnothing\right\}$, the event,

$$\{\delta_n > \varepsilon\} \cap \left\{\exists(\theta, \theta') \in \{D_n(S) \times D(S)\} : \|\theta - \theta'\| \leq \eta \cap \left\{G_{n,1}(\theta) + \varepsilon/2 \geq \bar{G}_{n,1}\right\}\right\}$$

leads to the following derivation,

$$
\begin{aligned}
& G\left(\left\{[v_n(m_{j,\theta})]_+\right\}_{j\in S},\{0\}_{j\in\bar{S}}\right)+\frac{\varepsilon}{2} \\
& \overset{(1)}{\geq} G\left(\left\{\left[\sqrt{n}\mathbb{P}_n(m_{j,\theta})\right]_+\right\}_{j\in\left\{\mathbb{P}_n\left(m_{j,\theta}\right)\geq 0\right\}},\{0\}_{j\in\left\{\mathbb{P}_n\left(m_{j,\theta}\right)<0\right\}}\right)+\frac{\varepsilon}{2} \\
& \overset{(2)}{\geq} \sup_{\tilde{\theta}\in\Theta_I} G\left(\left\{\left[\sqrt{n}\mathbb{P}_n\left(m_{j,\tilde{\theta}}\right)\right]_+\right\}_{j=1}^{J}\right) \overset{(3)}{\geq} \sup_{\tilde{\theta}\in\Theta_I} G\left(\left\{\left[v_n\left(m_{j,\tilde{\theta}}\right)\right]_+ 1\left[\mathbb{P}\left(m_{j,\tilde{\theta}}\right)=0\right]\right\}_{j=1}^{J}\right)+\varepsilon \\
& \overset{(4)}{\geq} G\left(\left\{\left[v_n\left(m_{j,\theta'}\right)\right]_+\right\}_{j\in S},\{0\}_{j\in\bar{S}}\right)+\varepsilon
\end{aligned}
$$

where $\overset{(1)}{\geq}$ holds because $\theta\in D_n(S)\subseteq\Theta_I$, $\overset{(2)}{\geq}$ holds because $\left\{G_{n,1}(\theta)+\varepsilon/2\geq\bar{G}_{n,1}\right\}$, $\overset{(3)}{\geq}$ holds because $\delta_n>\varepsilon$ and $\overset{(4)}{\geq}$ holds because $\theta'\in D(S)$. As a consequence,

$$
\left\{\sup_{\theta\in\Theta_I}\sup_{\|\theta'-\theta\|\leq\eta}\left|G\left(\left\{[v_n(m_{j,\theta})]_+\right\}_{j\in S},\{0\}_{j\in\bar{S}}\right)-G\left(\left\{\left[v_n\left(m_{j,\theta'}\right)\right]_+\right\}_{j\in S},\{0\}_{j\in\bar{S}}\right)\right|\right\}>\frac{\varepsilon}{2}
$$

By continuity, $\forall\eta>0$, $\exists N\in\mathbb{N}$ such that $\forall n\geq N$, the event $\{\delta_n>\varepsilon\cap A_n\}$ implies the event $\sup_{\theta\in\Theta_I}\sup_{\|\theta'-\theta\|\leq\eta}\|v_n(m_\theta)-v_n(m_{\theta'})\|>\gamma$. As a consequence,

$$
\limsup_{n\to\infty}P^*(\delta_n>\varepsilon\cap A_n)\leq\limsup_{n\to\infty}P^*\left(\sup_{\theta\in\Theta}\sup_{\|\theta'-\theta\|\leq\eta}\|v_n(m_\theta)-v_n(m_{\theta'})\|>\gamma\right)
$$

Taking $\eta\downarrow 0$ and by stochastic equicontinuity, this part is completed.

Point 2. The empirical process $v_n(m_\theta):\Omega_n\to l_J^\infty(\Theta_I)$ is asymptotically equicontinuous in probability (with respect to the Euclidean metric) and Θ_I is totally bounded (for this metric). By multivariate CLT, for every finite collection of elements of Θ_I, $\{v_n(m_{\theta m})\}_{m=1}^{M}$ converges to a Gaussian random vector with a variance covariance matrix whose (m_1,m_2) element is given by $\Sigma(\theta_{m_1},\theta_{m_2})$. Arguments in van der Vaart and Wellner [**63**] (theorem 1.5.7. and example 1.5.10.), complete this proof.

Point 3. The function $H(y) = \sup_{\theta\in\Theta_I} G\left(\left\{[y_j]_+ 1\left[\mathbb{P}(m_{j,\theta}) = 0\right]\right\}_{j=1}^{J}\right)$ is trivially continuous and non-negative. Weak convexity can be verified by definition. We now show that $\exists(\theta_0, j) \in \{\Theta_I, (1, 2, ..., J)\}$ such that $\mathbb{P}(m_{j,\theta_0}) = 0$. Since that $\mathbb{E}(m(Z,\cdot))$ is a lower semi-continuous function, Θ_I is closed or equivalently, $\Theta \cap \Theta_I^c$ is open. Now proceed by contradiction. That is, suppose that $\forall\theta \in \Theta_I$, $\max_{j=1:J} \mathbb{P}(m_{j,\theta}) < 0$, which implies that Θ_I is open. By assumption, $\exists\theta' \in \Theta \cap \Theta_I^c$. By the case under consideration, $\Theta_I \neq \varnothing$, and so $\exists\theta'' \in \Theta_I$. Consider the set $S = \{\theta \in \Theta : \theta''\pi + \theta'(1-\pi), \pi \in [0,1]\}$. By definition, it is a convex set (hence, connected). Moreover, S can be expressed as the union of two non-empty open sets by intersecting it with Θ_I and Θ_I^c. This is a contradiction. As a corollary, $H(y) = 0$ implies that $y_j(\theta_0) \leq 0$.

<u>Part 2.</u> $\tilde{\delta}_n$ is given by,

$$\tilde{\delta}_n = \sup_{\theta\in\Theta_I} G\left(\left\{\left[\begin{array}{c}\sqrt{n}\hat{p}_k\left(\mathbb{E}_n(Y_j|x_k) - \mathbb{E}(Y_j|x_k)\right) + \\ +\sqrt{n}\hat{p}_k\left(\mathbb{E}(Y_j|x_k) - M_{j,k}(\theta)\right)\end{array}\right]_+\right\}_{(j,k)=1}^{J\times K}\right) +$$

$$-\sup_{\theta\in\Theta_I} G\left(\left\{\begin{array}{c}\left[\sqrt{n}\hat{p}_k\left(\mathbb{E}_n(Y_j|x_k) - \mathbb{E}(Y_j|x_k)\right)\right]_+ * \\ *1\left[p_k\left(M_{j,k}(\theta) - \mathbb{E}(Y_j|x_k)\right) = 0\right]\end{array}\right\}_{(j,k)=1}^{J\times K}\right)$$

Point 1. Define $y_n = \sqrt{n}\hat{p}_k\left(\mathbb{E}_n(Y_j|x_k) - \mathbb{E}(Y_j|x_k)\right)$, $\hat{p} = \{\hat{p}_k\}_{k=1}^{K}$ and the functions $R_n(\pi, y, \theta)$ and $R(y, \theta)$ as follows,

$$\begin{aligned} R_n(\pi, y, \theta) &= G\left(\left\{\left\{\left[y_{j,k} + \sqrt{n}\pi_k\left(\mathbb{E}(Y_j|x_k) - M_{j,k}(\theta)\right)\right]_+\right\}_{j=1}^{J}\right\}_{k=1}^{K}\right) \\ R(y, \theta) &= G\left(\left\{\left\{[y_{j,k}]_+ 1\left[p_k\left(M_{j,k}(\theta) - \mathbb{E}(Y_j|x_k)\right) = 0\right]\right\}_{j=1}^{J}\right\}_{k=1}^{K}\right) \end{aligned}$$

Then, $\tilde{\delta}_n = \sup_{\theta\in\Theta_I} R_n\left(p_n, y_n, \theta\right) - \sup_{\theta\in\Theta_I} R\left(y_n, \theta\right).$

Denote $p_L = \min\left\{p_k\right\}_{k=1}^{K}$ and $\Delta = \left\{\pi : \sum_{k=1}^{K}\pi_k = 1,\ \pi_k \geq p_L/2\right\}$. For any sequence $\left\{\varepsilon_n\right\}_{n=1}^{+\infty}$ with $\varepsilon_n = o\left(1\right)$, consider the following derivation,

$$\sqrt{n}P\left(\left|\tilde{\delta}_n\right| > \varepsilon_n\right) =$$

$$= \left\{\begin{array}{c} \sqrt{n}P\left(\begin{array}{c} \left|\sup_{\theta\in\Theta_I} R_n\left(\hat{p}, y_n, \theta\right) - \sup_{\theta\in\Theta_I} R\left(y_n, \theta\right)\right| > \varepsilon_n \cap \\ \cap\left\{\hat{p}\in\Delta\cap\left\|y_n\right\|\leq n^{1/4}\right\} \end{array}\right) + \\ +\sqrt{n}P\left(\begin{array}{c} \left|\sup_{\theta\in\Theta_I} R_n\left(\hat{p}, y_n, \theta\right) - \sup_{\theta\in\Theta_I} R\left(y_n, \theta\right)\right| > \varepsilon_n \cap \\ \cap\left\{\hat{p}\notin\Delta\cup\left\|y_n\right\| > n^{1/4}\right\} \end{array}\right) \end{array}\right\}$$

$$\leq \left\{\begin{array}{c} \sqrt{n}1\left[\sup_{\pi\in\Delta}\sup_{\|y\|\leq n^{1/4}}\left|\sup_{\theta\in\Theta_I} R_n\left(\pi, y, \theta\right) - \sup_{\theta\in\Theta_I} R\left(y, \theta\right)\right| > \varepsilon_n\right] + \\ +\sqrt{n}P\left(\left\|y_n\right\| > n^{1/4}\right) + \sum_{k=1}^{K}\sqrt{n}P\left(\hat{p}_k \leq p_L/2\right) \end{array}\right\}$$

By Chebyshev's Inequality, $\sqrt{n}P\left(\left\|y_n\right\| > n^{1/4}\right) = o\left(1\right)$ and, therefore, it follows that $\forall k = 1, 2, ..., K$, $\sqrt{n}P\left(\hat{p}_k \leq p_L/2\right) = o\left(1\right)$. To conclude this point, we show that $\exists N \in \mathbb{N}$ such that $\forall n \geq N$, $\forall \pi \in \Delta$, $\forall y : \left\|y\right\| < n^{1/4}$, $\sup_{\theta\in\Theta_I} R_n\left(\pi, y, \theta\right) = \sup_{\theta\in\Theta_I} R\left(y, \theta\right)$. By assumption Θ_I is non-empty and compact, the functions $R_n\left(\pi, y, \cdot\right)$ and $R\left(y, \cdot\right)$ are upper semi-continuous, which implies that both achieve a maximum. The rest of the argument works by explicit calculation of the maxima in a case by case fashion. This is tedious but elementary.

Points 2 and 3. We first show that $\left\{\hat{p}_k\left(\mathbb{E}_n\left(Y_j|x_k\right) - \mathbb{E}\left(Y_j|x_k\right)\right)\right\}_{(j,k)=1}^{J\times K} = B\bar{Z}$ where $\bar{Z}$ is an average of i.i.d. vectors with zero mean, $V\left(Z_i\right) = \mathbf{I}_r$ and, under (B5), Z_i finite third moments.

Notice that $\{\hat{p}_k (\mathbb{E}_n (Y_j|x_k) - \mathbb{E}(Y_j|x_k))\}_{(j,k)=1}^{J\times K}$ is the average of i.i.d. vectors with zero mean, variance-covariance matrix given by the block diagonal matrix Σ, whose k^{th} block is given by $p_k V_k$ and V_k denotes the variance of $\{Y|X = x_k\}$. For every $k = 1, 2, ..., K$, let r_k be the rank of V_k and let B_k be the $J \times r_k$ dimensional matrix (with rank r_k) such that $B_k B_k' = p_k V_k$. Then, $\forall i = 1, 2, ..., n$, $\exists Z_{k,i} \in \mathbb{R}^{r_k}$ such that: $(Y_i - \mathbb{E}(Y|x_k)) 1[X_i = x_k] = B_k Z_{k,i}$. Let $B = [B_1, B_2, ..., B_k]'$ so that, by construction, $\Sigma = BB'$ and let $Z_i = [Z_{1,i}, Z_{2,i}, ..., Z_{K,i}]'$. Hence, $\forall i = 1, 2, ..., n$, $\exists Z_i \in \mathbb{R}^r$ so that: $\{(Y_{i,j} - \mathbb{E}(Y_j|x_k)) 1[X_i = x_k]\}_{(j,k)=1}^{J\times K} = BZ_i$. Since the variance of BZ_i equals BB' and B has full rank, $V(Z_i) = \mathbf{I}_r$. Finally, if $\{Y_i|X_i = x_k\}_{k=1}^{K}$ is assumed to have finite third moments, then Z_i will also have finite third moments. By averaging these observations, we notice that $\{\hat{p}_k (\mathbb{E}_n (Y_j|x_k) - \mathbb{E}(Y_j|x_k))\}_{(j,k)=1}^{J\times K} = B\bar{Z}$.

Next, consider the function $\tilde{H}(y) : \mathbb{R}^r \to \mathbb{R}$,

$$\tilde{H}(y) = \sup_{\theta\in\Theta_I} \left\{ G\left(\left\{ \left[B_{(j,k)} y\right]_+ 1\left[p_k \left(M_{j,k}(\theta) - \mathbb{E}(Y_j|x_k)\right) = 0\right] \right\}_{(j,k)=1}^{J\times K} \right) \right\}$$

We show that the function has the desired properties. This function is continuous and non-negative by the same arguments as in the previous part. Weak convexity can be verified by definition. By arguments in part 1, $\exists$ $(\theta_0, (j,k)) \in \{\Theta_I, \{1, 2, ..., J\} \times \{1, 2, ..., K\}\}$ such that $\mathbb{E}(Y_j|x_k) = M_{j,k}(\theta_0)$ and hence, $\tilde{H}(y) = 0$ implies that for $b = B_{(j,k)} \neq \vec{0}$, $b'y \leq 0$.

The remaining property is the one that requires the special functional forms imposed by assumption (CF). Consider $y_A \in \tilde{H}^{-1}((h_B - \varepsilon_n, h_B + \varepsilon_n])$, that is, $\exists h_A : \|h_A - h_B\| < \varepsilon_n$ such that $\tilde{H}(y_A) = h_A$. We need to show that $\exists y_B \in \mathbb{R}^r$ such that $\|y_A - y_B\| \leq O(\varepsilon_n)$ and $\tilde{H}(y_B) = h_B$. We consider first the case when $G(x) = \sum_{j=1}^{J} w_j x_j$ for $w > 0$. For any $z \in \mathbb{R}^r$, Let $g(z,\theta) = G\left(\left\{\left[B_{(j,k)}z\right]_+ 1\left[p_k\left(M_{j,k}(\theta) - \mathbb{E}(Y_j|x_k)\right) = 0\right]\right\}_{(j,k)=1}^{J\times K}\right)$. Since $g(z,\theta)$ depends on θ through indicator functions, then we classify Θ_I into finitely many subsets, according to the behavior they induce on the $J \times K$ indicator functions. From each group, we can extract one representative. Let $\{\theta_1, \theta_2, ..., \theta_\pi\}$ denote the group of such representatives. By definition, $\forall z \in \mathbb{R}^r$, $\max_{\theta\in\Theta_I} g(z,\theta) = \max_{\theta\in\{\theta_1,...,\theta_\pi\}} g(z,\theta)$. For any $(z,\theta) \in \{\mathbb{R}^r, \Theta_I\}$, let $\Lambda_+(z,\theta)$ denote the subset of $\{1, 2, ..., J\times K\}$ such that $M_{j,k}(\theta) = \mathbb{E}(Y_j|x_k)$ and $B_{(j,k)}z > 0$ and let $\Lambda_0(z,\theta)$ denote the subset of $\{1, 2, ..., J\times K\}$ such that $M_{j,k}(\theta) = \mathbb{E}(Y_j|x_k)$ and $B_{(j,k)}z = 0$.

Let $\{\theta_1, ..., \theta_m\}$ denote the subset of the representatives such that maximize $g(y_A, \theta)$. Consider any arbitrary $\theta' \in \{\theta_1, ..., \theta_m\}$. By definition $y_A \in \mathbb{R}^r$ satisfies the following equations: $\forall (j,k) \in \Lambda_0(y_A, \theta') : B_{(j,k)}x = 0$ and $\forall (j,k) \in \Lambda_+(y_A,\theta') : B_{(j,k)}x = h_{A,(j,k)} > 0$. By summing the equations for $(j,k) \in \Lambda_+(y_A,\theta')$, we get $\sum_{(j,k)\in\Lambda_+(y_A,\theta')} h_{A,(j,k)} = h_A$. Thus, $y_A \in \mathbb{R}^r$ satisfies the following system,

$$\begin{bmatrix} \sum_{(j,k)\in\Lambda_+(y_A,\theta')} B_{(j,k)} \\ \left[B_{(j,k)}\right]_{(j,k)\in\Lambda_0(y_A,\theta')} \end{bmatrix} x = \begin{bmatrix} h_A \\ \left[\vec{0}\right]_{(j,k)\in\Lambda_0(y_A,\theta')} \end{bmatrix}$$

We can repeat this process for the rest of the maximizers, i.e., $\forall\theta'' \in \{\theta_2, ..., \theta_m\} \backslash \theta'$. Instead of expressing the information contained in $\Lambda_0(y_A, \theta'')$ as $\sum_{(j,k)\in\Lambda_+(y_A,\theta'')} B_{(j,k)} = h_A$ we reexpress it as, $\sum_{(j,k)\in\Lambda_+(y_A,\theta'')} B_{(j,k)} - \sum_{(j,k)\in\Lambda_+(y_A,\theta')} B_{(j,k)} = 0$, which gives the following new equations,

$$\begin{bmatrix} \sum_{(j,k)\in\Lambda_+(y_A,\theta'')} B_{(j,k)} - \sum_{(j,k)\in\Lambda_+(y_A,\theta')} B_{(j,k)} \\ \left[B_{(j,k)}\right]_{(j,k)\in\Lambda_0(y_A,\theta'')} \end{bmatrix} x = \begin{bmatrix} 0 \\ \left[\vec{0}\right]_{(j,k)\in\Lambda_0(y_A,\theta'')} \end{bmatrix}$$

If we put together all the equations from $\theta \in \{\theta_1, \theta_2, ..., \theta_m\}$ in this manner, we will produce a system of linear equations, that can be expressed as $[C_1, C_2]' \, x = [h_A, \vec{0}]'$ where the matrix $[C_1, C_2]'$ does not depend on h_A. Consider the homogenous system $C_2 x = \vec{0}$. The matrix C_2 may or may not have full rank, but can always be reduced to a system $C_3 x = \vec{0}$, where C_3 has full rank. Since $h_A > 0$, $[C_1, C_3]'$ has full rank. If this rank is r, then $y_A = \left[[C_1, C_3]'\right]^{-1} [h_A, \vec{0}]'$. If the rank is less than r, we can any add additional (equality) restrictions that are satisfied by y_A, of the form $C_4 x = c$, until $[C_1, C_3, C_4]'$ has rank r. It is easy to see that C_4 constructed this way will not depend on y_A. Then, $y_A = \left[[C_1, C_3, C_4]'\right]^{-1} [h_A, \vec{0}, c]'$.

Consider $y_B = \left[[C_1, C_3, C_4]'\right]^{-1} [h_B, \vec{0}, c]'$. By construction $\|y_A - y_B\| = O\left(\varepsilon_n\right)$, (where $[C_1, C_3, C_4]$ does not depend on y_A). By construction and continuity, $\forall \theta \in \{\theta_1, ...\theta_m\}$ $\Lambda_+\left(y_A, \theta\right) = \Lambda_+\left(y_B, \theta\right)$ and if $\sum_{(j,k) \in \Lambda_+(y_A, \theta)} B_{(j,k)} y_A = h_A$, then $\sum_{(j,k) \in \Lambda_+(y_B, \theta)} B_{(j,k)} y_B = h_B$. Also by construction, $\forall \theta \in \{\theta_1, ...\theta_m\}$ then $\Lambda_0\left(y_A, \theta\right) = \Lambda_0\left(y_B, \theta\right)$. By continuity, $\forall \left(j, k\right) \in \{1, 2, ..., J \times K\}$ such that $M_{j,k}\left(\theta\right) = \mathbb{E}\left(Y_j | x_k\right)$ and $B_{(j,k)} y_A < 0$, then $B_{(j,k)} y_B < 0$. As a consequence, $\forall \theta \in \{\theta_1, ...\theta_m\}$, $g\left(y_B, \theta\right) = h_B$. By continuity, $\forall \theta \in \{\theta_1, ..., \theta_\pi\} \setminus \{\theta_1, ...\theta_m\}$, $g\left(y_B, \theta\right) < h_B$. Thus, by construction, $\tilde{H}\left(y_B\right) = h_B$.

The arguments for $G\left(x\right) = \max_{i=1,...,J \times K}\left\{w_i x_i\right\}$ for positive weights $\{w_i\}_{i=1}^{J \times K}$ are similar and, therefore, omitted.

Part 3. If $\Theta_I = \varnothing$, then, by definition, $\Gamma_n = 0$. $\square$

Theorem 37. *If we assume (B1)-(B4), (CF') and that* $\Theta_I \neq \varnothing$, *then, for some* $r \in \mathbb{N} \cap [1, J \times K]$, $\Gamma_n = \tilde{H}\left(\sqrt{n}\bar{Z}\right) + \tilde{\delta}_n$, *where,*

(1) *for any* $\varepsilon_n = O\left(n^{-1/2}\right)$, $P\left(\left|\tilde{\delta}_n\right| > \varepsilon_n\right) = o\left(n^{-1/2}\right)$,

(2) $\bar{Z} : \Omega_n \to \mathbb{R}^r$ *is a zero mean sample average of* n *i.i.d. observations from a distribution with variance-covariance matrix* $\mathbf{I}_r$. *If we add (B5), this distribution has finite third absolute moments,*

(3) $\tilde{H} : \mathbb{R}^r \to \mathbb{R}$ *is continuous, non-negative, weakly convex and homogeneous of degree* β*. For any* $\mu > 0$*, any* $|h| \geq \mu > 0$ *and any sequence* $\varepsilon_n = o(1)$, $\left\{\tilde{H}^{-1}\left(\{h\}^{\varepsilon_n}\right) \cap \|y\| \leq O\left(\sqrt{g_n}\right)\right\} \subseteq \left\{\tilde{H}^{-1}\left(\{h\}\right)\right\}^{\delta_n}$ *where* $\delta_n = O\left(\sqrt{g_n}\varepsilon_n\right)$*. Finally,* $\tilde{H}(y) = 0$ *implies for some non-zero vector that for a non-zero vector* $b \in \mathbb{R}^r$, $b'y \leq 0$.

Proof. The definitions of $\bar{Z}$ and $\tilde{H}$ are exactly the same as in theorem 36. Homogeneity of degree β can be verified by definition. To conclude, we need to show that $\forall h \geq \mu$, $\left\{\tilde{H}^{-1}\left(\{h\}^{\varepsilon_n}\right) \cap \|y\| \leq O\left(\sqrt{g_n}\right)\right\} \subseteq \left\{\tilde{H}^{-1}\left(\{h\}\right)\right\}^{\delta_n}$ *where* $\delta_n = O\left(\sqrt{g_n}\varepsilon_n\right)$. Consider $y' \in \tilde{H}^{-1}\left(\{h\}^{\varepsilon_n}\right)$ such that $\|y'\| \leq O\left(\sqrt{g_n}\right)$. Define $y = y'\left(h/h'\right)^{1/\beta}$. By homogeneity of degree β, $\tilde{H}(y) = h$. By definition:

$$\|y' - y\| \leq \|y'\| \left|1 - \left(h'/h\right)^{-1/\beta}\right| \leq O\left(\sqrt{g_n}\right) \max\left\{1 - \left(h'/h\right)^{-1/\beta}, \left(h'/h\right)^{-1/\beta} - 1\right\}$$

where $|h' - h| \leq \varepsilon_n$. For fixed h, $|h| \geq \mu > 0$, and $h' \in (h - \varepsilon_n, h + \varepsilon_n]$, a first order Taylor series argument implies that $\max\left\{1 - \left(h'/h\right)^{-1/\beta}, \left(h'/h\right)^{-1/\beta} - 1\right\} \leq O\left(|h' - h|\right) = O\left(\varepsilon_n\right)$. As a consequence, $\|y' - y\| \leq O\left(\sqrt{g_n}\varepsilon_n\right)$, completing the proof. □

Theorem 38. *Assume (A1)-(A4), (CF') and that* $\Theta_I \neq \varnothing$*. Then,* $\Gamma_n^* = H\left(v_n^*\left(m_\theta\right)\right) + \delta_n^*$*, where,*

(1) *for any* $\varepsilon > 0$, $\lim_{n\to\infty} P^*\left(|\delta_n^*| > \varepsilon | \mathcal{X}_n\right) = 0$, *a.s.,*

(2) $\{v_n^*(m_\theta)|\mathcal{X}_n\}:\Omega_n \to l_J^\infty(\Theta)$ *is an empirical process that converges weakly to the same Gaussian process as in theorem 36, a.s.,*

(3) $H : l_J^\infty(\Theta) \to \mathbb{R}$ *is the same function as in theorem 36.*

If we assume (B1)-(B4), (CF), that $\Theta_I \neq \varnothing$ *and we choose the bootstrap procedure to be the one specialized for the conditionally separable model, then, for* r *as in theorem 36,* $\Gamma_n^* = \tilde{H}\left(\sqrt{n}\bar{Z}^*\right) + \tilde{\delta}_n^*$*, where,*

(1) *for any* $\varepsilon_n = O\left(n^{-1/2}\right)$, $P\left(\left|\tilde{\delta}_n^*\right| > \varepsilon_n|\mathcal{X}_n\right) = o\left(n^{-1/2}\right)$*, a.s.,*

(2) $\left\{\bar{Z}^*|\mathcal{X}_n\right\}:\Omega_n \to \mathbb{R}^r$ *is a zero mean sample average of* n *independent observations from a distribution with variance covariance matrix* $\hat{V}$*. If we also assume (B5), this distribution has finite third moments, a.s., and* $\left\|\hat{V} - \mathbf{I}_r\right\| \leq O_p\left(n^{-1/2}\right)$*,*

(3) $\tilde{H} : \mathbb{R}^r \to \mathbb{R}$ *is the same function as in theorem 36.*

If, instead, we assume (A1)-(A4), (CF') and that $\Theta_I = \varnothing$ *then,* $\liminf\{\Gamma_n^* = 0\}$*, a.s..*

Proof. Part 1. By the CLT for bootstrapped empirical processes applied to Donsker classes (see, e.g., Giné and Zinn [**30**]), $\{v_n^*|\mathcal{X}_n\} \xrightarrow{d} \zeta$ in $l_J^\infty(\Theta)$, a.s., where ζ is the same Gaussian process as in theorem 36. The function $H(y) : l_J^\infty(\Theta) \to \mathbb{R}$ given by

$$H(y) = \sup_{\theta \in \Theta_I} G\left(\left\{[y_j(\theta)]_+ 1\left[\mathbb{P}(m_{j,\theta}) = 0\right]\right\}_{j=1}^J\right)$$

is continuous and so, by the continuous mapping theorem (see, e.g., van der Vaart and Wellner [**63**]), $\{H(v_n^*)|\mathcal{X}_n\} \xrightarrow{d} H(\zeta)$, a.s..

Conditional on the sample, let $H_n : l_J^\infty(\Theta) \to \mathbb{R}$ be the following function,

$$H_n(y) \;=\; \sup_{\theta \in \hat{\Theta}_I(\tau_n)} G\left(\left\{[y_j(\theta)]_+ 1\left[\mathbb{P}_n(m_{j,\theta}) \geq -\tilde{\tau}_n/\sqrt{n}\right]\right\}_{j=1}^J\right)$$

Let δ_n^* be defined as: $\delta_n^* = H\left(v_n^*\left(m_\theta\right)\right) - H_n\left(v_n^*\left(m_\theta\right)\right)$. If $\forall \varepsilon > 0$, $P\left(\left|\delta_n^*\right| > \varepsilon | \mathcal{X}_n\right) = o\left(1\right)$, a.s., then $\left\{H_n\left(v_n^*\right)|\mathcal{X}_n\right\} \overset{d}{\to} H\left(Z\right)$, a.s.. In fact, this concludes the proof since the function H has all the required properties.

Step 1: Show that $\lim_{n\to\infty} P\left(\delta_n^* \geq 0 | \mathcal{X}_n\right) = 1$, a.s.. Denote by A_n the following event,

$$A_n = \left\{\left\{\Theta_I \subseteq \hat{\Theta}_I\left(\tau_n\right)\right\} \bigcap \left\{\bigcap_{\theta\in\Theta} \bigcap_{j=1:J} \left\{\mathbb{P}\left(m_{j,\theta}\right) \geq 0 \Longrightarrow \mathbb{P}_n\left(m_{j,\theta}\right) \geq -\tilde{\tau}_n/\sqrt{n}\right\}\right\}\right\}$$

By definition, A_n implies $\left\{\delta_n^* \geq 0\right\}$. Conditional on the sample, A_n is non-random, and so it suffices to show $P\left(\liminf\left\{A_n\right\}\right) = 1$, which follows from the LIL.

Step 2: Show that $\forall \varepsilon > 0$, $P\left(\delta_n^* > \varepsilon | \mathcal{X}_n\right) = o\left(1\right)$, a.s. For any $\epsilon > 0$, let $\Theta_I\left(\epsilon\right)$ $= \left\{\theta \in \Theta : \mathbb{P}\left(m_\theta\right) \leq \epsilon\right\}$ and let $H^\epsilon\left(y\right) : l_J^\infty\left(\Theta\right) \to \mathbb{R}$ denote the function $H^\epsilon\left(y\right) =$ $\sup_{\theta\in\Theta_I(\epsilon)} G\left(\left\{\left[y_j\left(\theta\right)\right]_+ 1\left[\mathbb{P}\left(m_{j,\theta}\right) \geq -\epsilon\right]\right\}_{j=1}^J\right)$. Consider a positive sequence $\left\{\varepsilon_n\right\}_{n=1}^{+\infty}$ such that $\varepsilon_n = o\left(1\right)$, $\left(\tau_n/\sqrt{n}\right)\varepsilon_n^{-1} = o\left(1\right)$ and $\left(\tilde{\tau}_n/\sqrt{n}\right)\varepsilon_n^{-1} = o\left(1\right)$, a.s..

Step 2.1: We now show $\lim_{n\to\infty} P\left(H_n\left(v_n^*\left(m_\theta\right)\right) \leq H^{\varepsilon_n}\left(v_n^*\left(m_\theta\right)\right)|\mathcal{X}_n\right) = 1$, a.s.. Let A_n' denote the following event,

$$A_n' = \left\{\left\{\hat{\Theta}_I\left(\tau_n\right) \subseteq \Theta_I\left(\varepsilon_n\right)\right\} \bigcap \left\{\bigcap_{\theta\in\Theta} \bigcap_{j=1:J} \left\{\mathbb{P}_n\left(m_{j,\theta}\right) \geq -\tilde{\tau}_n/\sqrt{n} \Longrightarrow \mathbb{P}\left(m_{j,\theta}\right) \geq -\varepsilon_n\right\}\right\}\right\}$$

The event A_n' implies $\left\{H_n\left(v_n^*\left(m_\theta\right)\right) \leq H^{\varepsilon_n}\left(v_n^*\left(m_\theta\right)\right)\right\}$ and therefore, it is sufficient to show that $P\left(\liminf\left\{A_n'\right\}\right) = 1$, which follows from the LIL.

Step 2.2: Let $\eta_{n,1}^*$ be defined by,

$$\eta_{n,1}^* = \left\{\begin{array}{c} \sup_{\theta\in\Theta_I(\varepsilon_n)} G\left(\left\{\left[v_n^*\left(m_{j,\theta}\right)\right]_+ 1\left[\mathbb{P}\left(m_{j,\theta}\right) \geq -\varepsilon_n\right]\right\}_{j=1}^J\right) + \\ -\sup_{\theta\in\Theta_I} G\left(\left\{\left[v_n^*\left(m_{j,\theta}\right)\right]_+ 1\left[\mathbb{P}\left(m_{j,\theta}\right) \geq -\varepsilon_n\right]\right\}_{j=1}^J\right) \end{array}\right\}$$

then, we show that $\forall \varepsilon > 0$, $\limsup_{n\to\infty} P^* \left(\eta^*_{n,1} > \varepsilon | \mathcal{X}_n\right) = 0$ a.s..

Fix $\varepsilon > 0$. For any $\theta \in \Theta$, define the following functions,

$$\begin{aligned} G_{n,1}(\theta) &= G\left(\left\{\left[v^*_n(m_{j,\theta})\right]_+ 1\left[\mathbb{P}(m_{j,\theta}) \geq -\varepsilon_n\right]\right\}_{j=1}^{J}\right) \\ G_{n,2}(\theta) &= G\left(\left\{\left[v^*_n(m_{j,\theta})\right]_+ 1\left[\mathbb{P}(m_{j,\theta}) \geq -\varepsilon_n\right]\right\}_{j=1}^{J}\right) \end{aligned}$$

and set $\bar{G}_{n,1} = \sup_{\theta\in\Theta_I(\varepsilon_n)} G_{n,1}(\theta)$ and $\bar{G}_{n,2} = \sup_{\theta\in\Theta_I} G_{n,2}(\theta)$. With these definitions, $\eta^*_{n,1} = \bar{G}_{n,1} - \bar{G}_{n,2}$ and so, it follows that the event $\left\{\eta^*_{n,1} > \varepsilon | \mathcal{X}_n\right\}$ implies the event $\left\{\exists \theta \in \{\Theta_I(\varepsilon_n) \cap \Theta_I^c\} : G_{n,1}(\theta) + \varepsilon/2 \geq \bar{G}_{n,1} | \mathcal{X}_n\right\}$.

Let $\left\{\mathcal{P}^{\{1,2,..,J\}}/\varnothing\right\}$ denote the set of all non-empty subsets of the set $\{1,2,..,J\}$ and $\forall S \in \left\{\mathcal{P}^{\{1,2,..,J\}}/\varnothing\right\}$, let $\bar{S}$ denote $\{1,2,...,J\}/S$. Then,

$$\{\Theta_I(\varepsilon_n) \cap \Theta_I^c\} = \bigcup_{\substack{S_0 \in \left\{\mathcal{P}^{\{1,2,..,J\}}/\varnothing\right\} \\ S_1 \cap S_2 = \bar{S}_0}} \left\{\Theta \cap \left\{\begin{array}{c} \left\{\bigcap_{j\in S_0}\{\mathbb{P}(m_{j,\theta}) \in (0,\varepsilon_n]\}\right\} \\ \bigcap\left\{\bigcap_{j\in S_1}\{\mathbb{P}(m_{j,\theta}) \in [-\varepsilon_n, 0]\}\right\} \\ \bigcap\left\{\bigcap_{j\in S_2}\{\mathbb{P}(m_{j,\theta}) < -\varepsilon_n\}\right\} \end{array}\right\}\right\}$$

For any $S \in \left\{\mathcal{P}^{\{1,2,..,J\}}/\varnothing\right\}$, consider the sets $D_n(S)$ and $D(S)$ given by,

$$\begin{aligned} D_n(S) &= \left\{\Theta \cap \left\{\left\{\bigcap_{j\in S}\{\mathbb{P}(m_{j,\theta}) \in [-\varepsilon_n, \varepsilon_n]\}\right\} \bigcap \left\{\bigcap_{j\in\bar{S}}\{\mathbb{P}(m_{j,\theta}) < -\varepsilon_n\}\right\}\right\}\right\} \\ D_n'(S) &= \left\{\Theta \cap \left\{\bigcap_{j\in S}\{\mathbb{P}(m_{j,\theta}) = 0\}\right\} \bigcap \left\{\bigcap_{j\in\bar{S}}\{\mathbb{P}(m_{j,\theta}) < -\varepsilon_n\}\right\}\right\} \end{aligned}$$

According to these definitions, the event $\left\{\eta^*_{n,1} > \varepsilon | \mathcal{X}_n\right\}$ implies the following event,

$$\bigcup_{S \in \left\{\mathcal{P}^{\{1,2,..,J\}}/\varnothing\right\}} \left\{\exists \theta \in D_n(S) : G_{n,1}(\theta) + \varepsilon/2 \geq \bar{G}_{n,1} | \mathcal{X}_n\right\}$$

For every $S \in \left\{\mathcal{P}^{\{1,2,..,J\}}/\varnothing\right\}$ and $\forall\eta > 0$, $\exists N \in \mathbb{N} : \forall n \geq N$, $\{\theta \in D_n(S)\}$ implies $\{\exists\theta' \in D_n'(S) : \|\theta - \theta'\| < \eta\}$. Thus, $\forall S \in \left\{\mathcal{P}^{\{1,2,..,J\}}/\varnothing\right\}$ and $\forall\eta > 0$, $\exists N \in \mathbb{N}$ such that $\forall n \geq N$, the event $\left\{\exists\theta \in D_n(S) : G_{n,1}(\theta) + \varepsilon/2 \geq \bar{G}_{n,1}|\mathcal{X}_n\right\}$ is equivalent to the event,

$$\left\{\exists(\theta,\theta') \in \{D_n(S) \times D_n'(S)\} : \left\{\|\theta - \theta'\| \leq \eta \cap G_{n,1}(\theta) + \varepsilon/2 \geq \bar{G}_{n,1}\right\}|\mathcal{X}_n\right\}$$

Therefore, $\forall\eta > 0$, $\exists N \in \mathbb{N}$ such that $\forall n \geq N$, the event $\left\{\eta_{n,1}^* > \varepsilon|\mathcal{X}_n\right\}$ is equivalent to the event,

$$\bigcup_{S\in\left\{\mathcal{P}^{\{1,2,..,J\}}/\varnothing\right\}} \left\{ \left\{\eta_{n,1}^* > \varepsilon\right\} \cap \left\{ \begin{array}{c} \exists(\theta,\theta') \in \{D_n(S) \times D_n'(S)\} : \\ \left\{\|\theta - \theta'\| \leq \eta \cap G_{n,1}(\theta) + \varepsilon/2 \geq \bar{G}_{n,1}\right\} \end{array} \right\} \middle| \mathcal{X}_n \right\}$$

Now, $\forall\eta > 0$, $\forall S \in \left\{\mathcal{P}^{\{1,2,..,J\}}/\varnothing\right\}$, the event,

$$\left\{ \left\{\eta_{n,1}^* > \varepsilon\right\} \cap \left\{ \begin{array}{c} \exists(\theta,\theta') \in \{D_n(S) \times D_n'(S)\} : \\ \left\{\|\theta - \theta'\| \leq \eta \cap G_{n,1}(\theta) + \varepsilon/2 \geq \bar{G}_{n,1}\right\} \end{array} \right\} \middle| \mathcal{X}_n \right\}$$

leads to the following derivation,

$$\begin{aligned}
& G\left(\left\{[v_n^*(m_{j,\theta})]_+\right\}_{j\in S}, \{0\}_{j\in\bar{S}}\right) + \frac{\varepsilon}{2} \\
& \overset{(1)}{\geq} \sup_{\tilde{\theta}\in\Theta_I(\varepsilon_n)} G\left(\left\{\left[v_n^*\left(m_{j,\tilde{\theta}}\right)\right]_+ 1\left[\mathbb{P}\left(m_{j,\tilde{\theta}}\right) \geq -\varepsilon_n\right]\right\}_{j=1}^J\right) \\
& \overset{(2)}{\geq} \sup_{\tilde{\theta}\in\Theta_I} G\left(\left\{\left[v_n^*\left(m_{j,\tilde{\theta}}\right)\right]_+ 1\left[\mathbb{P}\left(m_{j,\tilde{\theta}}\right) \geq -\varepsilon_n\right]\right\}_{j=1}^J\right) + \varepsilon \\
& \overset{(3)}{\geq} G\left(\left\{\left[v_n^*\left(m_{j,\theta'}\right)\right]_+\right\}_{j\in S}, \{0\}_{j\in\bar{S}}\right) + \varepsilon
\end{aligned}$$

where $\overset{(1)}{\geq}$ holds by $\left\{G_{n,1}(\theta)+\varepsilon/2 \geq \bar{G}_{n,1}\right\}$, $\overset{(2)}{\geq}$ holds because $\left\{\eta_{n,1}^{*}>\varepsilon\right\}$ and $\overset{(3)}{\geq}$ holds because $\left\{\theta' \in D_n'(S)\right\}$. By the same arguments as in the proof of theorem 36 (part 1), $\forall \eta>0$, $\exists \gamma>0$ such that,

$$\limsup_{n\to\infty} P^{*}\left(\eta_{n,1}^{*}>\varepsilon|\mathcal{X}_n\right) \leq \limsup_{n\to\infty} P^{*}\left(\sup_{\theta\in\Theta_I}\sup_{\|\theta'-\theta\|\leq\eta}\left\|v_n^{*}(m_\theta)-v_n^{*}(m_{\theta'})\right\|>\gamma \middle| \mathcal{X}_n\right)$$

If we take $\eta \downarrow 0$, the right hand side can be shown to be equal to zero a.s., using arguments of theorem 2.4 in Giné and Zinn [**30**] (equation 2.16).

Step 2.3: Let $\eta_{n,2}^{*}$ be defined by,

$$\eta_{n,2}^{*}=\left\{\begin{array}{c}\sup_{\theta\in\Theta_I} G\left(\left\{\left[v_n^{*}(m_{j,\theta})\right]_{+} 1\left[\mathbb{P}(m_{j,\theta})\geq-\varepsilon_n\right]\right\}_{j=1}^{J}\right)+ \\ -\sup_{\theta\in\Theta_I} G\left(\left\{\left[v_n^{*}(m_{j,\theta})\right]_{+} 1\left[\mathbb{P}(m_{j,\theta})\geq 0\right]\right\}_{j=1}^{J}\right)\end{array}\right\}$$

We show that $\forall\varepsilon>0$, $\limsup_{n\to\infty} P^{*}\left(\eta_{n,2}^{*}<\varepsilon|\mathcal{X}\right)=0$ a.s.. The structure of the arguments are similar to the ones used in the previous step and, therefore, omitted.

Step 3: Combine steps 1 and 2 to complete the proof.

<u>Part 2.</u> *Step 1:* Show that $\{\tilde{p}_k^{*}(\mathbb{E}_n^{*}(Y|x_k)-\mathbb{E}_n(Y|x_k))\}_{k=1}^{K}$ is the average of n (conditionally) independent observations from the sample distribution, with variance $\hat{\Sigma}$ and finite third moments, a.s.. For every $k=1,2,...,K$, let $\tilde{p}_k^{*}\equiv\bar{p}_k$ in the fixed design case and $\tilde{p}_k^{*}\equiv\hat{p}_k^{*}$ in the random design case.

In the first case, suppose that the design is fixed. Let $\{n_1,n_2,...,n_K\}$ denote the number of observations of each covariate value, which implies that $\sum_{k=1}^{K} n_k=n$. For each $k=1,2,...,K$, we sample n_k observations by sampling randomly with replacement from the sample $\{Y_i|X_i=x_k\}$. Denote this sample by $\left\{Y_{i,k}^{*}\right\}_{i=1}^{n_k}$. The (Y,X) pairs produced in this fashion constitutes our bootstrap sample, denoted $\mathcal{X}_n^{*}$. We construct a sample of size

n, such that the first n_1 observations are given by $\left\{Y^*_{i,1} - \bar{Y}_1, 0_{1\times J}, ..., 0_{1\times J}\right\}_{i=1}^{n_1}$, the second n_2 observations are given by $\left\{0_{1\times J}, Y^*_{i,2} - \bar{Y}_2, 0_{1\times J}, ..., 0_{1\times J}\right\}_{i=1}^{n_2}$, and so on. This results in n observations of $J \times K$ dimensional vectors, $\left\{\left\{\left(Y^*_{i,k} - \mathbb{E}_n\left(Y|x_k\right)\right) 1\left[X_i = x_k\right]\right\}_{k=1}^{K}\right\}_{n=1}^{n}$, and whose average is $\left\{\bar{p}_k\left(\mathbb{E}^*_n\left(Y|x_k\right) - \mathbb{E}_n\left(Y|x_k\right)\right)\right\}_{k=1}^{K}$. Conditional on the sample (and the design), these observations are independent, with variance $\hat{\Sigma}$ and finite third moments, a.s.. In the second case, suppose that the design is random. This is the usual bootstrap of pairs (X, Y) and the same result follows from simpler arguments.

Step 2: The next step is to show that: $\left\{\tilde{p}^*_k\left(\mathbb{E}^*_n\left(Y|x_k\right) - \mathbb{E}_n\left(Y|x_k\right)\right)\right\}_{k=1}^{K} = B\bar{Z}^*$, where $BB' = \Sigma$ and $\bar{Z}^*$ is the average of a sample of an independent sample with (conditional) mean zero, variance covariance $\hat{V}$ such that $\left\|\hat{V} - \mathbf{I}_r\right\| = O_p\left(n^{-1/2}\right)$ and finite third moments, a.s.. We cover proof for the random design case (the fixed design case only requires change of notation).

As in the proof of theorem 36 (part 2), $\forall k = 1, 2, ..., K$, let r_k be the rank of V_k and let B_k be the $J \times r_k$ dimensional matrix (with rank r_k) such that $B_k B_k' = p_k V_k$. We show that $\forall i = 1, 2, ..., n$, there exists $Z^*_{k,i} \in \mathbb{R}^{r_k}$ such that: $Y_i^* 1\left[X_i^* = x_k\right] = B_k Z^*_{k,i}$.

If $\{Y_i^*, X_i^*\}$ is such that $X_i^* \neq x_k$ and since B_k has full rank, $Z^*_{k,i} = 0$. If $\{Y_i^*, X_i^*\}$ is such that $X_i^* = x_k$ and since we resample our observations, then for some $Z_{k,i} \in \mathbb{R}^{r_k}$, $\left(Y_i^* - \mathbb{E}_n\left(Y|x_k\right)\right) = B_k\left(Z_{k,i} - \mathbb{E}_n\left(Z|x_k\right)\right)$. Since B_k has full rank, the value of $\left(Y_i^* - \mathbb{E}_n\left(Y|x_k\right)\right)$ determines a unique value for $\left(Z_{k,i} - \mathbb{E}_n\left(Z|x_k\right)\right)$. Hence, we choose, $Z^*_{k,i} = \left(Z_{k,i} - \mathbb{E}_n\left(Z|x_k\right)\right) 1\left[X_i^* = x_k\right]$. By repeating $\forall k = 1, ..., K$, we construct $Z_i^* = \left[Z^*_{1,i}, ..., Z^*_{K,i}\right]$ such that $\left\{\left(Y_i^* - \mathbb{E}_n\left(Y|x_k\right)\right) 1\left[X_i^* = x_k\right]\right\}_{k=1}^{K} = BZ_i^*$.

By construction, $\{Z_i^*\}_{i=1}^{n}$ is a sample of random vectors sampled from a distribution with (conditional) mean zero. Now add assumption (B5). By construction, $V\left(BZ_i^*|\mathcal{X}_n\right) =$

$\hat{\Sigma}$, and therefore: $\left\|\hat{\Sigma}-\Sigma\right\| = \left\|B\left(\hat{V}-\mathbf{I}_r\right)B'\right\|$ where $\hat{V} = V\left(Z_i^*|\mathcal{X}_n\right)$. By definition, $\hat{V}$ is the sample variance of $\{Z_i\}_{i=1}^{n}$. Since $B \in \mathbb{R}^{(J\times K)\times r}$ has rank r and by the CLT, it follows that $\left\|\hat{V}-\mathbf{I}_R\right\| \leq O\left(n^{-1/2}\right)$. Finally, since $\{(Y_i^* - \mathbb{E}_n\left(Y|x_k\right))1\left[X_i^* = x_k\right]\}_{k=1}^{K}$ has finite third moments, a.s., Z_i^* also has finite third moments, a.s..

Step 3: We now show that $\Gamma_n^* = \tilde{H}\left(\sqrt{n}\bar{Z}^*\right) + \tilde{\delta}_n^*$ where $\tilde{H}$ is the same function as in theorem 36 and for any sequence $\{\varepsilon_n\}_{n=1}^{+\infty}$ such that $\varepsilon_n = O\left(n^{-1/2}\right)$, $P\left(|\delta_n^*| > \varepsilon_n|\mathcal{X}\right) = o\left(n^{-1/2}\right)$, a.s.. Let $\mathcal{P}^{\{(1:J)\times(1:K)\}}$ denote the set of all subsets of $\{(1,...,J)\times(1,...,K)\}$ and for every $S\in\mathcal{P}^{\{(1:J)\times(1:K)\}}$, let $\bar{S} = \{(1,...,J)\times(1,...,K)\}\setminus S$.

Step 3.1: Let $S\in\mathcal{P}^{\{(1:J)\times(1:K)\}}$ and suppose that $\exists\theta_0\in\Theta_I$ such that for $(j,k)\in S$, $p_k\left(\mathbb{E}\left(Y_j|x_k\right) - M_j\left(\theta_0,x_k\right)\right) = 0$. We show that $\exists N\in\mathbb{N}$ such that $\forall n\geq N$, $\exists\theta\in\hat{\Theta}_I\left(\tau_n\right)$ such that $\hat{p}_k\left(\mathbb{E}_n\left(Y_j|x_k\right) - M_j\left(\theta,x_k\right)\right) \geq -\tilde{\tau}_n/\sqrt{n}$ for $(j,k)\in S$, a.s.. In particular, we show this for $\theta=\theta_0$, i.e., we show that,

$$P\left(\liminf\left\{\left\{\theta_0\in\hat{\Theta}_I\left(\tau_n\right)\right\}\cap\left\{\hat{p}_k\left(\mathbb{E}_n\left(Y_j|x_k\right) - M_j\left(\theta_0,x_k\right)\right)\geq-\tilde{\tau}_n/\sqrt{n}\right\}_{(j,k)\in S}\right\}\right) = 1$$

This result follows from the LIL for random vectors (see, e.g., Billingsley [**15**]).

Step 3.2: For $S\in\mathcal{P}^{\{(1:J)\times(1:K)\}}\backslash\varnothing$, suppose that $\nexists\theta\in\Theta_I$ such that for $(j,k)\in S$, $p_k\left(\mathbb{E}\left(Y_j|x_k\right) - M_j\left(\theta,x_k\right)\right) = 0$. We show that $\exists N\in\mathbb{N}$ such that $\forall n\geq N$, $\nexists\theta\in\hat{\Theta}_I\left(\tau_n\right)$ such that: $\hat{p}_k\left(\mathbb{E}_n\left(Y_j|x_k\right) - M_j\left(\theta_n,x_k\right)\right)\geq-\tilde{\tau}_n/\sqrt{n}$ for $(j,k)\in S.$, a.s.. For $S\in\mathcal{P}^{\{(1:J)\times(1:K)\}}\backslash\varnothing$, let $\Psi_n\left(S\right)$ to be given by:

$$\Psi_n\left(S\right) = \left\{\theta\in\Theta : \left\{\bigcap_{(j,k)\in S}\left\{\hat{p}_k\left(\mathbb{E}_n\left(Y_j|x_k\right) - M_j\left(\theta,x_k\right)\right)\geq-\tilde{\tau}_n/\sqrt{n}\right\}\right\}\right\}$$

It suffices to show that: $\liminf\left\{\nexists\theta\in\left\{\hat{\Theta}_I\left(\tau_n\right)\cap\Psi_n\left(S\right)\right\}\right\}$, a.s..

Step 3.2.1: We show that if for $S \in \mathcal{P}^{\{(1:J)\times(1:K)\}}\backslash\varnothing$, $\nexists\theta \in \Theta_I$ such that for $(j,k) \in S$, $p_k\left(\mathbb{E}\left(Y_j|x_k\right) - M_j\left(\theta, x_k\right)\right) = 0$, then $\exists\varpi > 0$ such that,

$$\begin{aligned}&\left\{\theta \in \Theta : \left\{\begin{array}{c}\left\{\max\limits_{(j,k)\in S}\left|p_k\left(M_j\left(\theta,x_k\right) - \mathbb{E}\left(Y_j|x_k\right)\right)\right| \le \varpi\right\}\bigcap \\ \bigcap\left\{\bigcap\limits_{(j,k)\in\bar{S}}\left\{p_k\left(\mathbb{E}\left(Y_j|x_k\right) - M_j\left(\theta,x_k\right)\right) \le 0\right\}\right\}\end{array}\right\}\right\} \subseteq \\ &\subseteq \left\{\Theta_I^c \cap \Theta\right\}\end{aligned}$$

To show this, notice that the problem $\inf_{\theta\in\Theta_I}\left\{\max_{(j,k)\in S}\left|p_k\left(M_j\left(\theta,x_k\right) - \mathbb{E}\left(Y_j|x_k\right)\right)\right|\right\}$ achieves a minimum and, by hypothesis, the minimum cannot be zero. Assign this minimum to $\varpi > 0$.

Step 3.2.2: For any $\varphi \ge 0$, define the set $R\left(\varphi\right)$ as,

$$R\left(\varphi\right) = \left\{\theta \in \Theta : \left\{\left\{p_k\left(\mathbb{E}\left(Y_j|x_k\right) - M_j\left(\theta,x_k\right)\right) \le \varphi\right\}_{(j,k)=1}^{J\times K}\right\}\right\}$$

Notice that $\lim_{\varphi\to0} R\left(\varphi\right) = \Theta_I$. By continuity of $\left\{M_j\left(\cdot,x_k\right)\right\}_{(j,k)=1}^{J\times K}$ and step 3.2.1, $\exists\eta > 0$, such that,

$$R\left(\eta\right) \subseteq \left\{\theta \in \Theta : \left\{\begin{array}{c}\left\{\max\limits_{(j,k)\in S}\left|p_k\left(M_j\left(\theta,x_k\right) - \mathbb{E}\left(Y_j|x_k\right)\right)\right| > \varpi\right\}\bigcup \\ \bigcup\left\{\bigcup\limits_{(j,k)\in\bar{S}}\left\{p_k\left(\mathbb{E}\left(Y_j|x_k\right) - M_j\left(\theta,x_k\right)\right) > \eta\right\}\right\}\end{array}\right\}\right\}$$

Step 3.2.3: By elementary properties, $\liminf\left\{\nexists\theta \in \left\{\hat{\Theta}_I\left(\tau_n\right) \cap \Psi_n\left(S\right)\right\}\right\}$, a.s. holds if we show that,

$$P\left(\limsup\left\{\exists\theta \in \left\{\hat{\Theta}_I\left(\tau_n\right) \cap \Psi_n\left(S\right) \cap R\left(\eta\right)\right\}\right\}\right) = 0 \tag{A.1}$$

$$P\left(\limsup\left\{\exists\theta\in\left\{\hat{\Theta}_I\left(\tau_n\right)\cap R\left(\eta\right)^c\right\}\right\}\right)=0 \tag{A.2}$$

To show (A.1), use step 3.2.2 to deduce,

$$\begin{gathered}\left\{\hat{\Theta}_I\left(\tau_n\right)\cap\Psi_n\left(S\right)\cap R\left(\eta\right)\right\}\subseteq\\ \subseteq\left\{\begin{array}{c}\left\{\hat{\Theta}_I\left(\tau_n\right)\cap\Psi_n\left(S\right)\cap\max\limits_{(j,k)\in S}\left|p_k\left(M_j\left(\theta,x_k\right)-\mathbb{E}\left(Y_j|x_k\right)\right)\right|>\varpi\right\}\cup\\ \cup\left\{\bigcup\limits_{(j,k)\in\bar{S}}\left\{\hat{\Theta}_I\left(\tau_n\right)\cap p_k\left(\mathbb{E}\left(Y_j|x_k\right)-M_j\left(\theta,x_k\right)\right)>\eta\right\}\right\}\end{array}\right\}\end{gathered}$$

Thus, it suffices to show that,

$$P\left(\limsup\left\{\exists\theta\in\left\{\hat{\Theta}_I\left(\tau_n\right)\cap\Psi_n\left(S\right)\cap\max\limits_{(j,k)\in S}\left|p_k\left(M_j\left(\theta,x_k\right)-\mathbb{E}\left(Y_j|x_k\right)\right)\right|>\varpi\right\}\right\}\right)=0 \tag{A.3}$$

and $\forall\left(j,k\right)\in\bar{S}$,

$$P\left(\limsup\left\{\exists\theta\in\left\{\hat{\Theta}_I\left(\tau_n\right)\cap p_k\left(\mathbb{E}\left(Y_j|x_k\right)-M_j\left(\theta,x_k\right)\right)>\eta\right\}\right\}\right)=0 \tag{A.4}$$

To show (A.3), notice that,

$$\begin{gathered}\left\{\hat{\Theta}_I\left(\tau_n\right)\cap\Psi_n\left(S\right)\cap\max\limits_{(j,k)\in S}\left|p_k\left(M_j\left(\theta,x_k\right)-\mathbb{E}\left(Y_j|x_k\right)\right)\right|>\varpi\right\}\subseteq\\ \subseteq\bigcup\limits_{(j,k)\in S}\left\{\begin{array}{c}\left\{\hat{p}_kM_j\left(\theta,x_k\right)+\frac{\tau_n}{\sqrt{n}}\geq\hat{p}_k\mathbb{E}_n\left(Y_j|x_k\right)\geq\hat{p}_kM_j\left(\theta,x_k\right)-\frac{\tilde{\tau}_n}{\sqrt{n}}\right\}\cap\\ \left\{p_k\left(M_j\left(\theta,x_k\right)-\mathbb{E}\left(Y_j|x_k\right)\right)>\varpi\cup p_k\left(M_j\left(\theta,x_k\right)-\mathbb{E}\left(Y_j|x_k\right)\right)<-\varpi\right\}\end{array}\right\}\\ \subseteq\bigcup\limits_{(j,k)\in S}\left\{\begin{array}{c}\left\{\left|\hat{p}_k\left(\mathbb{E}_n\left(Y_j|x_k\right)-\mathbb{E}\left(Y_j|x_k\right)\right)\right|\geq\frac{\varpi}{2}-\frac{\max\{\tau_n,\tilde{\tau}_n\}}{\sqrt{n}}\right\}\\ \cup\left\{\left|p_k-\hat{p}_k\right|\max_{\theta\in\Theta}\left(M_j\left(\theta,x_k\right)-\mathbb{E}\left(Y_j|x_k\right)\right)>\frac{\varpi}{2}\right\}\end{array}\right\}\end{gathered}$$

and so, the result follows from the SLLN. To show (A.4), notice that, for $(j,k) \in \bar{S}$,

$$\begin{aligned}
&\left\{\hat{\Theta}_I\left(\tau_n\right) \cap p_k\left(\mathbb{E}\left(Y_j|x_k\right) - M_j\left(\theta, x_k\right)\right) > \eta\right\} \subseteq \\
&\subseteq \left\{\left\{p_k\left(\mathbb{E}\left(Y_j|x_k\right) - M_j\left(\theta, x_k\right)\right) > \eta\right\} \cap \left\{\hat{p}_k\left(\mathbb{E}_n\left(Y_j|x_k\right) - M_j\left(\theta, x_k\right)\right) < \tau_n/\sqrt{n}\right\}\right\} \\
&\subseteq \left\{\begin{array}{c} \left\{\hat{p}_k\left(\mathbb{E}\left(Y_j|x_k\right) - \mathbb{E}_n\left(Y_j|x_k\right)\right) \geq -\tau_n/\sqrt{n} + \eta/2\right\} \cup \\ \cup \left\{\left(p_k - \hat{p}_k\right)\left(\mathbb{E}\left(Y_j|x_k\right) - M_j\left(\theta, x_k\right)\right) > \eta/2\right\} \end{array}\right\}
\end{aligned}$$

and so, again, the result follows from the SLLN.

Now consider (A.2). Notice that,

$$\begin{aligned}
&P\left(\limsup\left\{\exists\theta \in \left\{\hat{\Theta}_I\left(\tau_n\right) \cap R\left(\eta\right)^c\right\}\right\}\right) = \\
&= P\left(\limsup\left\{\exists\theta \in \left\{\Theta : \left\{\begin{array}{c} \bigcap_{(j,k)=1}^{J,K} \left\{\hat{p}_k\left(\mathbb{E}_n\left(Y_j|x_k\right) - M_j\left(\theta, x_k\right)\right) \leq \tau_n/\sqrt{n}\right\} \\ \cap \bigcup_{(j,k)=1}^{J,K} \left\{p_k\left(\mathbb{E}\left(Y_j|x_k\right) - M_j\left(\theta, x_k\right)\right) > \eta\right\} \end{array}\right\}\right\}\right\}\right) \\
&\leq \sum_{(j,k)=1}^{J,K} \left\{\begin{array}{c} P\left(\limsup\left\{\hat{p}_k\left(\mathbb{E}\left(Y_j|x_k\right) - \mathbb{E}_n\left(Y_j|x_k\right)\right) > \frac{\eta}{4}\right\}\right) + \\ +P\left(\limsup\left\{\left|p_k - \hat{p}_k\right| \left|\sup_{\theta\in\Theta}\left(\mathbb{E}\left(Y_j|x_k\right) - M_j\left(\theta, x_k\right)\right)\right| > \frac{\eta}{2}\right\}\right) \end{array}\right\}
\end{aligned}$$

and the right hand side is zero by SLLN.

Step 3.3: Let $\tilde{\Gamma}_n^*$ be defined as follows,

$$\tilde{\Gamma}_n^* = \tilde{H}\left(\sqrt{n}\bar{Z}^*\right) = \sup_{\theta\in\Theta_I}\left\{G\left(\left\{\left[B_{(j,k)}\bar{Z}^*\right]_+ 1\left[p_k\left(\mathbb{E}\left(Y_j|x_k\right) - M_j\left(x_k, \theta\right)\right) = 0\right]\right\}_{(j,k)=1}^{J\times K}\right)\right\}$$

and then $\tilde{\delta}_n^* = \Gamma_n^* - \tilde{\Gamma}_n^*$. From steps 3.1, it follows that $\liminf\left\{\tilde{\delta}_n^* \geq 0\right\}$, a.s. and from and 3.2, if follows that $\liminf\left\{\tilde{\delta}_n^* \leq 0\right\}$, a.s.. Combining both results, we deduce that $\liminf \left\{ \tilde{\delta}_n^* = 0 \right\}$, a.s. and thus, for any non-negative sequence $\varepsilon_n = O\left(n^{-1/2}\right)$, $\sqrt{n} P\left(\left| \tilde{\delta}_n^* \right| > \varepsilon_n | \mathcal{X}_n \right) = o\left(1\right)$, a.s..

Part 3. By definition $\left\{ \hat{\Theta}_I\left(\tau_n\right) = \varnothing \right\}$ implies $\left\{ \Gamma_n^* = 0 \right\}$. By lemma 5, if $\Theta_I = \varnothing$, $\liminf \left\{ \hat{\Theta}_I\left(\tau_n\right) = \varnothing \right\}$, a.s.. □

Theorem 39. *If we assume (B1)-(B4), (CF') and that* $\Theta_I \neq \varnothing$, *then,* $\Gamma_n^* = \tilde{H}\left(\sqrt{n}\bar{Z}^*\right) + \tilde{\delta}_n^*$, *where,*

(1) *for any* $\varepsilon_n = O\left(n^{-1/2}\right)$, $P\left(\left| \tilde{\delta}_n^* \right| > \varepsilon_n | \mathcal{X}_n \right) = o\left(n^{-1/2}\right)$, *a.s.,*

(2) $\left\{ \bar{Z}^* | \mathcal{X}_n \right\} : \Omega_n \to \mathbb{R}^r$ *is a zero mean sample average of* n *independent observations from a distribution with variance covariance matrix* $\hat{V}$. *If we also assume (B5), this distribution has finite third moments, a.s., and* $\left\| \hat{V} - \mathbf{I}_r \right\| \leq O_p\left(n^{-1/2}\right)$,

(3) $\tilde{H} : \mathbb{R}^r \to \mathbb{R}$ *is the same function as in theorem 37.*

Proof. Trivial from the proof of theorem 38 and theorem 37. □

A.2.4. Consistency results

Proof of lemma 6. Part 1. Follows from theorem 36 and the CLT for empirical processes.

Part 2. If $\Theta_I = \varnothing$, then $\lim_{n\to\infty} P\left(\Gamma_n = h\right) = 1\left[h = 0\right]$, which is continuous at h if and only if $h \neq 0$. If $\Theta_I \neq \varnothing$, then by the first part, $\lim_{n\to\infty} P\left(\Gamma_n = h\right) = P\left(H\left(\zeta\right) = h\right)$. Since $H \geq 0$, we only need to consider $h > 0$. By theorem 36, H is weakly convex and lower semicontinuous and so, the result follows from theorem 11.1 in Davydov, Lifshits and Smorodina [**26**] (part (i)). □

Proof of theorem 7. The proof follows directly from theorem 38 by using CLT for empirical processes. $\square$

Proof of theorem 9. Part 1. Consider the case $\Theta_I \neq \varnothing$. Fix $\mu > 0$ arbitrarily. By lemma 7, for every $|h| \geq \mu$, $\lim_{n\to\infty} |P(\Gamma_n^* \leq h|\mathcal{X}_n) - P(H(\zeta) \leq h)| = 0$, a.s.. By lemma 6, $P(H(\zeta) \leq h)$ is continuous for every $|h| \geq \mu$. The combination of the two implies the result.

Part 2. Now consider $\Theta_I = \varnothing$. By theorem 38, $\liminf \{P(\Gamma_n^* \leq h|\mathcal{X}_n) = 1[h \geq 0]\}$, a.s., which implies uniform convergence, a.s.. $\square$

Corollary 40. *Suppose that* $\Theta_I \neq \varnothing$. *For any* $\alpha \in [0, 0.5)$, *define* $q_n^B(1-\alpha) \equiv P\left(\Gamma_n^* \leq \hat{c}_n^B(1-\alpha)|\mathcal{X}_n\right)$. *Then,* $\left|q_n^B(1-\alpha) - (1-\alpha)\right| = o_p(1)$.

Proof. By lemma 6, $\lim_{n\to\infty} P(\Gamma_n \leq h) = P(H(\zeta) \leq h)$ with $\zeta \sim N(0, \mathbf{I}_r)$. By theorem 36, $H(\zeta) \leq 0$ implies that $\exists j \in \{1, ..., J\}$ and $\exists \theta_0 \in \Theta_I$ such that $\zeta(\theta_0) \leq 0$. Since $\zeta(\theta_0) \sim N(0, Var(m_j(Z, \theta_0)))$ with $Var(m_j(Z, \theta_0)) > 0$, then: $P(H(\zeta) \leq 0) \leq P(\zeta(\theta_0) \leq 0) \leq 0.5 < 1 - \alpha$.

Let $c_\infty(1-\alpha)$ denote the $(1-\alpha)$ quantile of the limiting distribution. By lemma 6, for any $h > 0$, $P(H(\zeta) \leq h)$ is continuous, and so $\forall \alpha \in [0, 0.5)$, $P(H(\zeta) \leq c_\infty(1-\alpha)) = 1 - \alpha$, which implies $c_\infty(1-\alpha) > 0$.

By theorem 9, $\sup_{|h|\geq\mu} |P(\Gamma_n^* \leq h|\mathcal{X}_n) - P(H(\zeta) \leq h)| \leq \varepsilon/2$ w.p.a.1. For any $\varepsilon/2 > 0$, choose $\mu > 0$ so that $\{c_\infty(1-\alpha+\varepsilon/2) \geq \mu\}$. By the continuity of $P(H(\zeta) \leq h)$ this implies that,

$$\lim_{n\to\infty} P\left((1-\alpha) \leq P(\Gamma_n^* \leq c_\infty(1-\alpha+\varepsilon/2)|\mathcal{X}_n) \leq (1-\alpha)+\varepsilon\right) = 1$$

By definition,

$$\{(1-\alpha) \leq P(\Gamma_n^* \leq c_\infty(1-\alpha+\varepsilon/2)|\mathcal{X}_n)\} \subseteq \left\{\hat{c}_n^B(1-\alpha) \leq c_\infty(1-\alpha+\varepsilon/2)\right\}$$

$$\subseteq \left\{(1-\alpha) \leq P\left(\Gamma_n^* \leq \hat{c}_n^B(1-\alpha)|\mathcal{X}_n\right) \leq P(\Gamma_n^* \leq c_\infty(1-\alpha+\varepsilon/2)|\mathcal{X}_n)\right\}$$

Therefore,

$$\lim_{n\to\infty} P\left(\left|P\left(\Gamma_n^* \leq \hat{c}_n^B(1-\alpha)|\mathcal{X}_n\right) - (1-\alpha)\right| \leq \varepsilon\right) \geq$$

$$\geq \lim_{n\to\infty} P\left(\begin{array}{c} (1-\alpha) \leq P\left(\Gamma_n^* \leq \hat{c}_n^B(1-\alpha)|\mathcal{X}_n\right) \leq \\ \leq P(\Gamma_n^* \leq c_\infty(1-\alpha+\varepsilon/2)|\mathcal{X}_n) \leq (1-\alpha)+\varepsilon \end{array}\right) = 1$$

completing the proof. □

Proof of corollary 10. Fix $\alpha \in [0, 0.5)$. It suffices to show that $\forall\varepsilon > 0$, $\exists N \in \mathbb{N}$, such that $\forall n \geq N$, $\left|P\left(\Theta_I \subseteq \hat{C}_n^B(1-\alpha)\right) - (1-\alpha)\right| < \varepsilon$. Fix $\varepsilon > 0$ and consider the following derivation,

$$\left|P\left(\Theta_I \subseteq \hat{C}_n^B(1-\alpha)\right) - (1-\alpha)\right| \leq \tag{A.5}$$

$$\leq \left\{\begin{array}{c} +\left|P\left(\Gamma_n \leq \hat{c}_n^B(1-\alpha)\right) - P\left(H(\zeta) \leq \hat{c}_n^B(1-\alpha)\right)\right| + \\ +\left|P\left(H(\zeta) \leq \hat{c}_n^B(1-\alpha)\right) - P\left(\Gamma_n^* \leq \hat{c}_n^B(1-\alpha)|\mathcal{X}_n\right)\right| + \\ +\left|P\left(\Gamma_n^* \leq \hat{c}_n^B(1-\alpha)|\mathcal{X}_n\right) - (1-\alpha)\right| \end{array}\right\}$$

The right hand side of equation (A.5) is the sum of three terms. For any $\mu > 0$, the first term satisfies,

$$\left|P\left(\Gamma_n \leq \hat{c}_n^B(1-\alpha)\right) - P\left(H(\zeta) \leq \hat{c}_n^B(1-\alpha)\right)\right| \leq \tag{A.6}$$

$$\leq \sup_{|h|\geq\mu} \left|P(\Gamma_n \leq h) - P(H(\zeta) \leq h)\right| + 2 * 1\left[\left|\hat{c}_n^B(1-\alpha)\right| < \mu\right]$$

By theorem 36 and the arguments used in the proof of theorem 9, $\forall\mu > 0$, the first term in the right hand side of equation (A.6) is $o(1)$. Next, we show that the second term of the right hand side of equation (A.6) is $o_p(1)$. It suffices to show that $\exists\mu > 0$ such that $\hat{c}_n^B(1-\alpha) \geq \mu$, w.p.a.1. By the arguments in corollary 40, $\forall\alpha \in [0, 0.5)$, $c_\infty(1-\alpha) > 0$. By lemma 6, the limiting distribution attains $(1-\alpha)$ level at $c_\infty(1-\alpha)$ and is continuous on $[0, c_\infty(1-\alpha)]$. Then, by intermediate value theorem, $\exists\eta \in (0, c_\infty(1-\alpha))$ such that $P(H(\zeta) \leq \eta) = ((1-\alpha) - 0.5)/2 + 0.5$, and so, pick $\mu = \eta$. Hence, by theorem 9,

$$\left|P(\Gamma_n^* \leq \eta|\mathcal{X}_n) - ((1-\alpha) - 0.5)/2 + 0.5\right| = \left|P(\Gamma_n^* \leq \eta|\mathcal{X}_n) - \lim_{n\to\infty} P(\Gamma_n \leq \eta)\right| = o_p(1)$$

and hence, $P(\Gamma_n^* \leq \eta|\mathcal{X}_n) < (1-\alpha)$, w.p.a.1. By definition of quantile, $(1-\alpha) \leq P\left(\Gamma_n^* \leq \hat{c}_n^B(1-\alpha)|\mathcal{X}_n\right)$, and so, by the monotonicity of the CDF, $\hat{c}_n^B(1-\alpha) \geq \eta = \mu$, w.p.a.1.

The second term on hand side of equation (A.5) is $o_p(1)$ by theorem 9. Finally, the third term is $o_p(1)$ by corollary 40. Combining the three terms, the left hand side of equation (A.5) is $o_p(1)$ and since it is non-stochastic, it has to be $o(1)$. $\square$

A.2.5. Rates of convergence results

Lemma 41. *Let $\tilde{H}$ be the function in theorem 36, let $\xi \sim N(0, \Upsilon)$ with non-singular $\Upsilon \in \mathbb{R}^{r \times r}$ and let $\{\varepsilon_n\}_{n=1}^{+\infty}$ be a positive sequence with $\varepsilon_n = o(1)$. Then, $\forall \mu > 0$,*

$$\sup_{|h| \geq \mu} \left| P\left(\tilde{H}(\xi) \in (h - \varepsilon_n, h + \varepsilon_n]\right) \right| \leq O(\varepsilon_n)$$

Proof. First, consider $h \leq -\mu$. Since $\tilde{H}(\xi) \geq 0$ and $\varepsilon_n = o(1)$ then, eventually, $h + \varepsilon_n < 0$ and so $P\left(\tilde{H}(\xi) \leq h + \varepsilon_n\right) = 0$.

Now consider $h \geq \mu$. Since $\varepsilon_n = o(1)$, $h - \varepsilon_n > 0$, eventually. Since $h > 0$ and $\varepsilon_n = o(1)$ then by theorem 36, $\tilde{H}^{-1}((h - \varepsilon_n, h + \varepsilon_n]) \subseteq \left\{\tilde{H}^{-1}(\{h\})\right\}^{\gamma_n}$ for $\gamma_n = O(\varepsilon_n)$. The submultiplicative property of the matrix norm implies that $\Upsilon^{-1/2}\left\{\tilde{H}^{-1}(\{h\})\right\}^{\gamma_n} \subseteq \left\{\Upsilon^{-1/2}\tilde{H}^{-1}(\{h\})\right\}^{\eta_n}$ for some $\eta_n = O(\varepsilon_n)$.

By theorem 36, $\tilde{H}$ is continuous and weakly quasiconvex, and since $h \neq 0$, $\tilde{H}^{-1}(\{h\}) = \partial \tilde{H}^{-1}((-\infty, h])$, where $\tilde{H}^{-1}((-\infty, h]) \in \mathcal{C}_r$. By the submultiplicative property of the matrix norm, $\Upsilon^{-1/2}\partial\tilde{H}^{-1}((-\infty, h]) = \partial\Upsilon^{-1/2}\tilde{H}^{-1}((-\infty, h])$ with $\Upsilon^{-1/2}\tilde{H}^{-1}((-\infty, h]) \in \mathcal{C}_r$. Combining all these steps, we deduce that,

$$P\left(\tilde{H}(\xi) \in (h - \varepsilon_n, h + \varepsilon_n]\right) \leq P\left(\vartheta \in \left\{\partial\Upsilon^{-1/2}\tilde{H}^{-1}((-\infty, h])\right\}^{\delta'_n}\right)$$

where $\vartheta \sim N(0, \mathbf{I}_r)$ and $\Upsilon^{-1/2}\tilde{H}^{-1}((-\infty, h]) \in \mathcal{C}_r$. Corollary 3.2. in Bhattacharya and Rao [**12**] (with $s = 0$) completes the proof. □

Proof of theorem 11. Fix $\mu > 0$ arbitrarily and let $\{g_n\}_{n=1}^{+\infty}$ be any positive sequence such that $g_n = O\left(n^{-1/2}\right)$.

Part 1. Consider the case when $\Theta_I \neq \varnothing$.

Step 1: show that $\sup_{|h|\geq\mu}\left|P\left(\Gamma_n\leq h\right)-P\left(\tilde{H}\left(\vartheta\right)\leq h\right)\right|\leq O\left(n^{-1/2}\right)$ where $\vartheta\sim N\left(0,\mathbf{I}_r\right).$ For $h\leq-\mu$, the statement holds since both Γ_n and $\tilde{H}\left(\vartheta\right)$ are non-negative. For $h\geq\mu$, from theorem 36, $P\left(\Gamma_n\leq h\right)\leq P\left(\left|\delta_n\right|>g_n\right)+P\left(\tilde{H}\left(\sqrt{n}\bar{Z}\right)\leq h+g_n\right)$, which implies that,

$$\sup_{|h|\geq\mu}\left\{\begin{array}{c}P\left(\Gamma_n\leq h\right)+\\-P\left(\tilde{H}\left(\vartheta\right)\leq h\right)\end{array}\right\}\leq$$

$$\leq\sup_{|h|\geq\mu}\left\{\begin{array}{c}P\left(\tilde{H}\left(\sqrt{n}\bar{Z}\right)\leq h+g_n\right)-P\left(\tilde{H}\left(\vartheta\right)\leq h+g_n\right)+\\+P\left(\tilde{H}\left(\vartheta\right)\in\left(h-g_n,h+g_n\right]\right)+P\left(\left|\tilde{\delta}_n\right|>g_n\right)\end{array}\right\}$$

$$\leq\left\{\begin{array}{c}\sup_{|h|\geq\mu}\left|P\left(\sqrt{n}\bar{Z}\in\tilde{H}^{-1}\left(\left(-\infty,h+g_n\right]\right)\right)-\Phi_{\mathbf{I}_r}\left(\tilde{H}^{-1}\left(\left(-\infty,h+g_n\right]\right)\right)\right|+\\+\sup_{|h|\geq\mu}P\left(\tilde{H}\left(\vartheta\right)\in\left(h-g_n,h+g_n\right]\right)+P\left(\left|\tilde{\delta}_n\right|>g_n\right)\end{array}\right\}$$

From theorem 36, $\forall\left|h\right|\geq\mu$, $\forall g_n$, $\tilde{H}^{-1}\left(\left(-\infty,h+g_n\right]\right)\in\mathcal{C}_r$, which implies that,

$$\sup_{|h|\geq\mu}\left\{\begin{array}{c}P\left(\Gamma_n\leq h\right)+\\-P\left(\tilde{H}\left(\vartheta\right)\leq h\right)\end{array}\right\}\leq$$

$$\leq\left\{\begin{array}{c}\sup_{A\in\mathcal{C}_r}\left|P\left(\sqrt{n}\bar{Z}\in A\right)-\Phi_{\mathbf{I}_r}\left(A\right)\right|+\\+\sup_{|h|\geq\mu}P\left(\tilde{H}\left(\vartheta\right)\in\left(h-g_n,h+g_n\right]\right)+P\left(\left|\tilde{\delta}_n\right|>g_n\right)\end{array}\right\}$$

The right hand side is a sum of three terms. By the Berry-Esseén theorem, the first term is $O\left(n^{-1/2}\right)$, by lemma 41, the second term is $O\left(g_n\right)=O\left(n^{-1/2}\right)$ and by theorem 36, the last term is $o\left(n^{-1/2}\right)$. If we combine this result with the analogous argument for $P\left(\Gamma_n>h\right)$ (instead of $P\left(\Gamma_n\leq h\right)$) we complete this step.

Step 2: show that $\sup_{|h|\geq\mu}\left|P\left(\Gamma_n^*\leq h|\mathcal{X}_n\right)-P\left(\tilde{H}\left(\hat{\vartheta}\right)\leq h|\mathcal{X}_n\right)\right|\leq O_p\left(n^{-1/2}\right)$ where $\hat{\vartheta}\sim N\left(0,\hat{V}\right)$ and $\hat{V}$ is the sample variance of $\{Z_i\}_{i=1}^n$. For $h\leq-\mu$, the statement holds since both Γ_n^* and $\tilde{H}\left(\hat{\vartheta}\right)$ are non-negative. For $h\geq\mu$, theorem 36 implies, that conditional on the sample, $\Gamma_n^*=\tilde{H}\left(\sqrt{n}\bar{Z}^*\right)+\tilde{\delta}_n^*$, where $P\left(\left|\tilde{\delta}_n^*\right|>g_n|\mathcal{X}_n\right)=o\left(n^{-1/2}\right)$, a.s.. By the same argument as in the previous step, it follows that,

$$\sup_{|h|\geq\mu}\left\{P\left(\Gamma_n^*\leq h|\mathcal{X}_n\right)-P\left(\tilde{H}\left(\hat{\vartheta}\right)\leq h|\mathcal{X}_n\right)\right\}\leq$$

$$\leq\left\{\begin{array}{c}\sup_{A\in\mathcal{C}_r}\left|P\left(\sqrt{n}\bar{Z}^*\in A|\mathcal{X}_n\right)-\Phi_{\hat{V}}\left(A\right)\right|+\\+\sup_{|h|\geq\mu}P\left(\tilde{H}\left(\hat{\vartheta}\right)\in\left(h-g_n,h+g_n\right]|\mathcal{X}_n\right)+P\left(\left|\tilde{\delta}_n^*\right|>g_n|\mathcal{X}_n\right)\end{array}\right\}$$

The right hand side is a sum of three terms. Conditional on the sample (and on the design), $\bar{Z}^*$ is the average of independent observations with mean zero, variance-covariance $\hat{V}$ and finite third moments, w.p.a.1. Thus, the Berry-Esseén theorem implies that the first term is $O_p\left(n^{-1/2}\right)$. By the CLT, $\left\|\hat{V}-\mathbf{I}_r\right\|\leq O_p\left(n^{-1/2}\right)$ and so, $\hat{V}$ is non-singular, w.p.a.1. Thus, by lemma 41, the second term is $O_p\left(n^{-1/2}\right)$. By theorem 38, the last term is $o_p\left(n^{-1/2}\right)$. We combine this with the same argument for $P\left(\Gamma_n^*>h|\mathcal{X}_n\right)$ (instead of $P\left(\Gamma_n^*\leq h|\mathcal{X}_n\right)$).

Step 3: show that $\sup_{|h|\geq\mu}\left|P\left(\tilde{H}\left(\vartheta\right)\leq h\right)-P\left(\tilde{H}\left(\hat{\vartheta}\right)\leq h|\mathcal{X}_n\right)\right|=O_p\left(n^{-1/2}\right)$ for $\vartheta\sim N\left(0,\mathbf{I}_r\right)$ and $\hat{\vartheta}\sim N\left(0,\hat{V}\right)$, with $\left\|\hat{V}-\mathbf{I}_r\right\|\leq O_p\left(n^{-1/2}\right)$. It suffices to show that $\int_{\mathbb{R}^r}\left|\phi_{\hat{V}}\left(x\right)-\phi_{\mathbf{I}_r}\left(x\right)\right|dx=O_p\left(n^{-1/2}\right)$, which follows from simple arguments.

Step 4: Combine steps 1, 2, and 3 to conclude the proof for the case $\Theta_I\neq\varnothing$.

Part 2. Now suppose that $\Theta_I = \varnothing$. By lemma 5, $\liminf \left\{ \hat{\Theta}_I (\tau_n) = \varnothing \right\}$, a.s.. Since the event $\left\{ \hat{\Theta}_I (\tau_n) = \varnothing \right\}$ implies the event $\{\forall h \in \mathbb{R} : P (\Gamma_n^* \leq h | \mathcal{X}_n) = 1 [h \geq 0]\}$, this completes the proof. □

Corollary 42. *Suppose that* $\Theta_I \neq \varnothing$. *For any* $\alpha \in [0, 0.5)$, *define* $q_n^B (1 - \alpha) \equiv P \left(\Gamma_n^* \leq \hat{c}_n^B (1 - \alpha) | \mathcal{X}_n \right)$. *Then,* $\left| q_n^B (1 - \alpha) - (1 - \alpha) \right| \leq O_p \left(n^{-1/2} \right)$.

Proof. Let $c_\infty (1 - \alpha)$ denote the $(1 - \alpha)$ quantile of the limiting distribution. By arguments in corollary 40, $c_\infty (1 - \alpha) > 0$. By the proof of theorem 11, $\forall \mu > 0$ and $\forall \gamma > 0$, $\exists K_\gamma > 0$ such that $\forall n \in \mathbb{N}$,

$$P \left(\sup_{|h| \geq \mu} |P (\Gamma_n^* \leq h | \mathcal{X}_n) - P (H (\vartheta) \leq h)| \leq K_\gamma n^{-1/2} \right) \geq 1 - \gamma$$

where $\vartheta \sim N (0, \mathbf{I}_r)$. Take $\varepsilon_n = K_\gamma n^{-1/2}$ and choose $\mu > 0$ so that $\exists N \in \mathbb{N}$ so that $\forall n \geq N$, $\{c_\infty (1 - \alpha + \varepsilon_n) > \mu\}$. As a consequence, $\forall n \geq N$,

$$P \left(\left| \begin{array}{c} P \left(\Gamma_n^* \leq c_\infty \left(1 - \alpha + K_\gamma n^{-1/2} \right) | \mathcal{X}_n \right) + \\ -P \left(H (\vartheta) \leq c_\infty \left(1 - \alpha + K_\gamma n^{-1/2} \right) \right) \end{array} \right| \leq K_\gamma n^{-1/2} \right) \geq 1 - \gamma$$

By the continuity of $P (H (\vartheta) \leq h)$, $P \left(H (\vartheta) \leq c_\infty \left(1 - \alpha + K_\gamma n^{-1/2} \right) \right) = 1 - \alpha + K_\gamma n^{-1/2}$, so that $\forall n \geq N$,

$$P \left((1 - \alpha) \leq P \left(\Gamma_n^* \leq c_\infty \left(1 - \alpha + K_\gamma n^{-1/2} \right) | \mathcal{X}_n \right) \leq (1 - \alpha) + 2 K_\gamma n^{-1/2} \right) \geq 1 - \gamma$$

By definition,

$$\begin{aligned}
&\left\{(1-\alpha) \leq P\left(\Gamma_n^* \leq c_\infty\left(1-\alpha+K_\gamma n^{-1/2}\right)|\mathcal{X}_n\right)\right\} \subseteq \\
&\subseteq \left\{\hat{c}_n^B\left(1-\alpha\right) \leq c_\infty\left(1-\alpha+K_\gamma n^{-1/2}\right)\right\} \\
&\subseteq \left\{(1-\alpha) \leq P\left(\Gamma_n^* \leq \hat{c}_n^B\left(1-\alpha\right)|\mathcal{X}_n\right) \leq P\left(\Gamma_n^* \leq c_\infty\left(1-\alpha+K_\gamma n^{-1/2}\right)|\mathcal{X}_n\right)\right\}
\end{aligned}$$

Therefore, $\exists N \in \mathbb{N}$ so that $\forall n \geq N$,

$$\begin{aligned}
&P\left(\left|P\left(\Gamma_n^* \leq \hat{c}_n^B\left(1-\alpha\right)|\mathcal{X}_n\right)-(1-\alpha)\right| \leq 2K_\gamma n^{-1/2}\right) \geq \\
&\geq P\left((1-\alpha) \leq P\left(\Gamma_n^* \leq \hat{c}_n^B\left(1-\alpha\right)|\mathcal{X}_n\right) \leq (1-\alpha)+2K_\gamma n^{-1/2}\right) \geq \\
&\geq P\left((1-\alpha) \leq P\left(\Gamma_n^* \leq c_\infty\left(1-\alpha+K_\gamma n^{-1/2}\right)|\mathcal{X}_n\right) \leq (1-\alpha)+2K_\gamma n^{-1/2}\right) \geq 1-\gamma
\end{aligned}$$

This conclusion can be extended $\forall n \in \mathbb{N}$ by appropriate choice of K_γ. $\square$

Proof of corolary 12. Part 1. Suppose that $\Theta_I \neq \varnothing$. First, notice that the event $\left\{\Theta_I \subseteq \hat{C}_n^B\left(1-\alpha\right)\right\}$ occurs if and only if the event $\left\{\Gamma_n \leq \hat{c}_n^B\left(1-\alpha\right)\right\}$ occurs. For any $K>0$, consider the following derivation,

$$\begin{aligned}
&\left\{\left|P\left(\Gamma_n \leq \hat{c}_n^B\left(1-\alpha\right)\right)-(1-\alpha)\right| > Kn^{-1/2}\right\} \\
&\subseteq \left\{\begin{array}{c} \left\{\left|P\left(\Gamma_n \leq \hat{c}_n^B\left(1-\alpha\right)\right)-P\left(\Gamma_n^* \leq \hat{c}_n^B\left(1-\alpha\right)|\mathcal{X}_n\right)\right| > (K/2)\,n^{-1/2}\right\} \cup \\ \cup \left\{\left|P\left(\Gamma_n^* \leq \hat{c}_n^B\left(1-\alpha\right)|\mathcal{X}_n\right)-(1-\alpha)\right| > (K/2)\,n^{-1/2}\right\} \end{array}\right\} \\
&\subseteq \left\{\begin{array}{c} \sup_{|h|\geq\mu}\left\{\left|P\left(\Gamma_n \leq \mu\right)-P\left(\Gamma_n^* \leq \mu|\mathcal{X}_n\right)\right| > (K/2)\,n^{-1/2}\right\} \cup \\ \cup \left\{\hat{c}_n^B\left(1-\alpha\right) < \mu\right\} \cup \left\{\left|P\left(\Gamma_n^* \leq \hat{c}_n^B\left(1-\alpha\right)|\mathcal{X}_n\right)-(1-\alpha)\right| > (K/2)\,n^{-1/2}\right\} \end{array}\right\}
\end{aligned}$$

Hence,

$$P\left(\left|P\left(\Gamma_n \leq \hat{c}_n^B\left(1-\alpha\right)\right)-\left(1-\alpha\right)\right|>Kn^{-1/2}\right)\leq$$

$$\leq\left\{\begin{array}{c} P\left(\sup_{|h|\geq\mu}\left|P\left(\Gamma_n\leq\mu\right)-P\left(\Gamma_n^*\leq\mu|\mathcal{X}_n\right)\right|>\left(K/2\right)n^{-1/2}\right)+ \\ +P\left(\hat{c}_n^B\left(1-\alpha\right)<\mu\right)+ \\ +P\left(\left|P\left(\Gamma_n^*\leq\hat{c}_n^B\left(1-\alpha\right)|\mathcal{X}_n\right)-\left(1-\alpha\right)\right|>\left(K/2\right)n^{-1/2}\right)\end{array}\right\}$$

Pick $\varepsilon>0$ arbitrarily. The right hand side is a sum of three terms. For a K large enough, the first term is smaller than $\varepsilon/3$ by theorem 11. By the arguments in corollary 10, $\hat{c}_n^B\left(1-\alpha\right)\geq\mu$, w.p.a.1 and so, $\exists N\in\mathbb{N}$ such that $\forall n\geq N$, the second term smaller than $\varepsilon/3$. The third term is smaller than $\varepsilon/3$ by corollary 42. By choosing K appropriately, we can extend $\forall n\in\mathbb{N}$.

Part 2. Now suppose that $\Theta_I=\varnothing$. By theorem 38, $\liminf\left\{\Gamma_n^*=0\right\}$, a.s., or equivalently, for any $\alpha\in\left[0,1\right]$, $\liminf\left\{\hat{c}_n^B\left(1-\alpha\right)=0\right\}$, a.s., concluding the proof. □

Lemma 43. *Let* $\tilde{H}$ *be the function in theorem 37 and assume that* $\xi\sim N\left(0,\Upsilon\right)$ *with non-singular* $\Upsilon\in\mathbb{R}^{r\times r}$*. Then,* $\forall\mu>0$*,*

$$\sup_{|h|\geq\mu}P\left(\tilde{H}\left(\xi\right)\in\left(h-n^{-1/2},h+n^{-1/2}\right]\right)\leq O\left(n^{-1/2}\ln n\right)$$

Proof. Consider the following derivation for $\varepsilon_n=n^{-1/2}$.

$$\sup_{|h|\geq\mu}P\left(\tilde{H}\left(\xi\right)\in\left(h-\varepsilon_n,h+\varepsilon_n\right]\right)=\sup_{|h|\geq\mu}P\left(\vartheta\in\Upsilon^{-1}\tilde{H}^{-1}\left(\left\{h\right\}^{\varepsilon_n}\right)\right)$$

$$\leq\left\{\sup_{|h|\geq\mu}P\left(\vartheta\in\left(\Upsilon^{-1}\tilde{H}^{-1}\left(\left\{h\right\}\right)\right)^{O\left(\varepsilon_n\sqrt{g_n}\right)}\right)+P\left(\left\|\vartheta\right\|>O\left(\sqrt{g_n}\right)\right)\right\}$$

where $\vartheta \sim N\left(0, \mathbf{I}_r\right)$.

Choose $g_n = \ln\left(n^{(1+\gamma)}\right)$ for some $\gamma > 0$. By theorem 37 and corollary 3.2 in Bhattacharya and Rao [**12**], the first term on the right side is $O\left(\varepsilon_n \sqrt{g_n}\right)$. By theorem 1 in Hüsler, Liu and Singh [**34**], $P\left(\|\vartheta\| > O\left(\sqrt{g_n}\right)\right)\left(\varepsilon_n \sqrt{g_n}\right)^{-1} \leq O\left(\exp\left(-\frac{g_n}{2}\right) g_n^{(r-3)/2} n^{1/2}\right)$ $= O(1)$. Since $\varepsilon_n \sqrt{g_n} = O\left(n^{-1/2} \ln n\right)$, the proof is completed. □

Corollary 44. *Assume (B1)-(B5) and (CF') and choose the bootstrap procedure to be the one specialized for the conditionally separable model. If $\Theta_I \neq \varnothing$ then, for any $\alpha \in [0, 0.5)$,*

$$\left|P\left(\Theta_I \subseteq \hat{C}_n^B (1-\alpha)\right) - (1-\alpha)\right| = O\left(n^{-1/2} \ln n\right)$$

and if $\Theta_I = \varnothing$ then, for any $\alpha \in [0, 1]$,

$$P\left(\liminf \left\{\hat{C}_n^B (1-\alpha) = \hat{\Theta}_I (0)\right\}\right) = 1$$

Proof. Follows from arguments used to prove theorems 11 and 12 under the result of lemma 43. □

A.2.6. Subsampling

A.2.6.1. Subsampling with recentering.

Theorem 45. *Let $\{b_n\}_{n=1}^{+\infty}$ be such that $b_n \to \infty$ and $b_n/n = o(1)$.*

If we assume (A1)-(A4), (CF') and that $\Theta_I \neq \varnothing$, then, $\Gamma_{b_n,n}^{SS} = H\left(v_{b_n,n}^{SS}(m_\theta)\right) + \delta_{b_n,n}^{SS}$, where,

(1) *for any $\varepsilon > 0$, $\lim_{n\to\infty} P^*\left(\left|\delta_{b_n,n}^{SS}\right| > \varepsilon | \mathcal{X}_n\right) = 0$, a.s..*

(2) $\{v^{SS}_{b_n,n}(m_\theta)|\mathcal{X}_n\}:\Omega_n\to l^\infty_J(\Theta)$ *is an empirical process that converges weakly to the same Gaussian process as in theorem 36, a.s..*

(3) $H:l^\infty_J(\Theta)\to\mathbb{R}$ *is the same function as in theorem 36.*

If, instead, we assume (B1)-(B4), (CF) and that $\Theta_I\neq\varnothing$*, then, for* r *as in theorem 36,* $\Gamma^{SS}_{b_n,n}=\tilde{H}\left(\sqrt{b_n}\bar{Z}^{SS}_{b_n,n}\right)+\tilde{\delta}^{SS}_{b_n,n}$*, where,*

(1) $\liminf\left\{\tilde{\delta}^{SS}_{b_n,n}=0\right\}$*, a.s..*

(2) $\{\bar{Z}^{SS}_{b_n,n}|\mathcal{X}_n\}:\Omega_n\to\mathbb{R}^r$ *is a zero mean sample average of* b_n *observations sampled without replacement from a distribution with variance covariance matrix* $\hat{V}$*. If we also assume (B5), this distribution has finite third moments, a.s., and* $\left\|\hat{V}-\mathbf{I}_r\right\|\leq O_p\left(n^{-1/2}\right)$*.*

(3) $\tilde{H}:\mathbb{R}^r\to\mathbb{R}$ *is the same function as in theorem 36.*

If, instead, we assume (A1)-(A4), (CF') and that $\Theta_I=\varnothing$ *then,* $\liminf\left\{\Gamma^{SS}_{b_n,n}=0\right\}$*, a.s..*

Proof. This proof follows the proof of theorem 36 very closely. We only focus on the differences.

<u>Part 1.</u> By the CLT for empirical processes, $v_n(m_\theta):\Omega_n\to l^\infty_J(\Theta)$ converges weakly to the tight Gaussian process we denote by ζ (see, e.g., theorem 1.5.7 in van der Vaart and Wellner [**63**]). Moreover, ζ has uniformly continuous sample paths, a.s.. Since the space of continuous functions on a compact space is separable, assumption 7.4.1 in Politis et al. [**54**] is satisfied, and so, by theorem 7.4.1 of Politis et al. [**54**], $v^{SS}_{b_n,n}$ converges weakly to ζ.

Part 2. The CLT for sample averages of bootstrapped vectors is replaced by the CLT for averages of subsampled vectors, as in theorem 2.2.1 of Politis et al. [**54**]. The conclusion that $\liminf \left\{\delta^{SS}_{b_n,n} = 0\right\}$, a.s. was already shown in theorem 36. □

Theorem 46 (Consistency of subsampling excluding zero)**.** *Assume (A1)-(A4), (CF') and let* $\{b_n\}_{n=1}^{+\infty}$ *be such that* $b_n \to \infty$ *and* $b_n/n = o(1)$*. If* $\Theta_I \neq \varnothing$ *then, for any* $\mu > 0$,

$$P\left(\lim_{n\to\infty} \sup_{|h|\geq\mu} \left| P\left(\Gamma^{SS}_{b_n,n} \leq h|\mathcal{X}_n\right) - \lim_{m\to\infty} P\left(\Gamma_m \leq h\right)\right| = 0\right) = 0$$

and if $\Theta_I = \varnothing$ *then,*

$$P\left(\liminf \left\{\sup_{h\in\mathbb{R}} \left| P\left(\Gamma^{SS}_{b_n,n} \leq h|\mathcal{X}_n\right) - \lim_{m\to\infty} P\left(\Gamma_m \leq h\right)\right| = 0\right\}\right) = 1$$

Proof. This proof follows the arguments of the proof of theorem 9. □

Proof of corollary 13. This proof follows the arguments of the proof of theorem 10. □

Lemma 47. *Assume that the distribution of* $\{Y|X = x_k\}_{k=1}^{K}$ *is strongly non-lattice and that* $b_n \to \infty$ *and* $b_n/n = o(1)$*. Then,*

$$P\left(\sqrt{b_n}\left(\mathbb{E}^{SS}_{b_n,n}(Z) - \mathbb{E}_n(Z)\right) \in S|\mathcal{X}_n\right) = \Phi(S) + K_1(S)\, b_n^{-1/2} + K_2(S)\, b_n/n + o_p\left(b_n^{-1/2} + b_n/n\right)$$

uniformly in $S \in \mathcal{C}_r$ *, where* $\sup_{S\in\mathcal{C}_r} |K_1(S)| < +\infty$ *and* $\sup_{S\in\mathcal{C}_r} |K_2(S)| < +\infty$.

For $\tilde{H}$ *as in theorem 36,* $\vartheta \sim N(0, \mathbf{I}_r)$ *and any* $\gamma > 0$*, let* h_L *and* h_H *be defined such that* $P\left(\tilde{H}(\vartheta) \leq h_L\right) = 0.72$ *and* $P\left(\tilde{H}(\vartheta) \leq h_H\right) = 1-\gamma$ *and define the function* $\Lambda(\gamma) = \left\{S \in \mathcal{C}_r : \exists h \in [h_L, h_H] : S = \left\{y \in \mathbb{R}^r : \tilde{H}(y) \leq h\right\}\right\}$*. Then,* $\inf_{S\in\Lambda(\gamma)} |K_2(S)| > 0$.

Proof. Part 1. For any $S \in \mathcal{C}_r$, Define $S_n(S) = \left\{x \in \mathbb{R}^r : (1 - b_n/n)^{-1/2} y \in S\right\}$. Notice that $S \in \mathcal{C}_r$ if and only if $S_n(S) \in \mathcal{C}_r$. By definition,

$$P\left(\sqrt{b_n}\left(\mathbb{E}^{SS}_{b_n,n}(Z) - \mathbb{E}_n(Z)\right) \in S \middle| \mathcal{X}_n\right) =$$
$$= P\left(\sqrt{b_n}\,(1 - b_n/n)^{-1/2}\left(\mathbb{E}^{SS}_{b_n,n}(Z) - \mathbb{E}_n(Z)\right) \in S_n(S) \middle| \mathcal{X}_n\right)$$

Babu and Singh [**6**] provide a Edgeworth expansion for averages of samples without replacement from a finite population, uniformly on classes of functions. For indicator functions over Borel measurable convex sets (in $\mathbb{R}^r$) combined with results in Bhattacharya and Rao [**12**], we deduce that,

$$P\left(\sqrt{b_n}\left(\mathbb{E}^{SS}_{b_n,n}(Z) - \mathbb{E}_n(Z)\right) \in S \middle| \mathcal{X}_n\right) = \Phi\left(S_n(S)\right) + b_n^{-1/2} K_1(S) + o_p\left(b_n^{-1/2}\right)$$

uniformly in $S \in \mathcal{C}_r$, $\forall S \in \mathcal{C}_r$, $K_1(S)$ is given by,

$$K_1(S) =$$
$$= \sum_{\beta \in \left\{b \in \mathbb{N}^r : \sum_{j=1}^r b_j = 3\right\}} \frac{1}{\prod_{j=1,\ldots,r} \beta_j!} \mathbb{E}\left(\prod_{j=1,\ldots,r} \left(Z_j - \mathbb{E}(Z_j)\right)^{\beta_j}\right) \int_{y \in S} \left(\prod_{j=1,\ldots,r} \frac{\partial^{\beta_j} \phi(y)}{\partial y_j}\right) dy$$

In order to deduce this result from Babu and Singh [**6**], we replace sample moments by population moments, introducing a term that is $O_p\left(n^{-1/2}\right) = o_p\left(b_n^{-1/2}\right)$, uniform in $S \in \mathcal{C}_r$, and we replace $b_n^{-1/2} K_1(S_n(S))$ by $b_n^{-1/2} K_1(S)$, which introduces the a term that is $o\left(b_n^{-1/2}\right)$, uniform in $S \in \mathcal{C}_r$. Since the normal distribution has finite absolute moments of all orders, we deduce that $\sup_{S \in \mathcal{C}_r} |K_1(S)| < +\infty$.

By change of variables and a Taylor expansion argument,

$$\Phi\left(S_n\left(S\right)\right) = \Phi\left(S\right) + K_2\left(S\right) b_n/n + o\left(b_n/n\right)$$

uniformly in $S \in \mathcal{C}_r$, where $K_2\left(S\right) = P\left(\vartheta \in S\right) \mathbb{E}\left(1 - \vartheta'\vartheta | \vartheta \in S\right)$ for $\vartheta \sim N\left(0, \mathbf{I}_r\right)$. Hence, $\sup_{S \in \mathcal{C}_r} \left|K_2\left(S\right)\right| < +\infty$. The expansion follows from combining both arguments.

<u>Part 2.</u> $\forall \gamma > 0$, we now show that $\exists C = C\left(\gamma\right) > 0$, such that $\inf_{S \in \Lambda(\gamma)} \left|K_2\left(S\right)\right| \geq C$. Consider any $S \in \Lambda\left(\gamma\right)$. Since $\tilde{H}$ is homogenous of degree β (with $\beta \geq 1$, by convexity), if $y \in S$, then $\forall \lambda \in [0,1]$, $\lambda y \in S$.

Case 1: $r = 1$. By homogeneity of degree β of $\tilde{H}$, $S = [-y_1, y_2]$ for some $y_1 \geq 0$ and $y_2 \geq 0$. By definition, $K_2\left(S\right) = K_2\left([0, y_2]\right) + K_2\left([-y_1, 0]\right)$. By inspection, $K_2\left([0, y_2]\right) \geq 0$ with strict inequality if $y_2 \in \left(0, +\infty\right)$. By symmetry, the same result applies to $K_2\left([-y_1, 0]\right)$. Since $P\left(\vartheta \in S\right) \in [0.72, 1-\gamma]$, either y_1 or y_2 is both positive and finite. Then, $\exists C_A > 0$ such that $\inf_{S \in \Lambda(\gamma)} P\left(\vartheta \in S\right) \mathbb{E}\left(1 - \vartheta'\vartheta | \vartheta \in S\right) \geq C_A$.

Case 2: $r > 1$. Since $\mathbb{E}\left(1 - \vartheta'\vartheta\right) \leq -1$ and by the homogeneity, then for a fixed probability given by $P\left(\vartheta \in S\right)$, $\mathbb{E}\left(1 - \vartheta'\vartheta | \vartheta \in S\right)$ is maximized if the probability mass is completely assigned to a circle around zero. For any S, let $c > 0$ be define by $P\left(\vartheta \in S\right) = P\left(\vartheta'\vartheta \leq c\right)$ and by construction: $\mathbb{E}\left(1 - \vartheta'\vartheta | \vartheta \in S\right) \leq \mathbb{E}\left(1 - \vartheta'\vartheta | \vartheta'\vartheta \leq c\right)$, which implies $P\left(\vartheta \in S\right) \mathbb{E}\left(1 - \vartheta'\vartheta | \vartheta \in S\right) \leq P\left(\vartheta'\vartheta \leq c\right) \mathbb{E}\left(1 - \vartheta'\vartheta | \vartheta'\vartheta \leq c\right)$. Since $P\left(\vartheta \in S\right) \geq 0.72$, then $c \geq 2.52$, which implies that $\mathbb{E}\left(1 - \vartheta'\vartheta | \vartheta'\vartheta \leq c\right) < 0$. By continuity of the function $f\left(c\right) = P\left(\vartheta'\vartheta \leq c\right) \mathbb{E}\left(1 - \vartheta'\vartheta | \vartheta'\vartheta \leq c\right)$, it follows that $\exists C_B > 0$, such that $\sup_{S \in \Lambda(\gamma)} \left\{P\left(\vartheta \in S\right) \mathbb{E}\left(1 - \vartheta'\vartheta | \vartheta \in S\right)\right\} \leq -C_B$.

The result follows by considering $C = \min\left\{C_A, C_B\right\}$. $\square$

Corollary 48. *Assume (B1)-(B5), (CF), that the distribution of* $\{Y|X=x_k\}_{k=1}^{K}$ *is strongly non-lattice and let* $b_n \to \infty$ *and* $b_n/n = o(1)$. *If* $\Theta_I \neq \varnothing$ *then, for any* $\mu > 0$,

$$\sup_{|h|\geq\mu}\left|P\left(\Gamma_{b_n,n}^{SS}\leq h|\mathcal{X}_n\right)-P\left(\Gamma_n\leq h\right)\right|\leq O_p\left(b_n^{-1/2}+b_n/n\right)$$

If $\Theta_I = \varnothing$ *then,*

$$P\left(\liminf\left\{\sup_{h\in\mathbb{R}}\left|P\left(\Gamma_{b_n,n}^{SS}\leq h|\mathcal{X}_n\right)-P\left(\Gamma_n\leq h\right)\right|=0\right\}\right)=1$$

Proof. Part 1. Consider first the case when $\Theta_I \neq \varnothing$.

As a first step, we show that: $\sup_{|h|\geq\mu}\left|P\left(\Gamma_{b_n,n}^{SS}\leq h|\mathcal{X}_n\right)-\lim_{m\to\infty}P\left(\Gamma_m\leq h\right)\right| = O_p\left(b_n^{-1/2}+b_n/n\right)$. By theorem 45, $\lim_{m\to\infty}P\left(\Gamma_m\leq h\right)=\Phi\left(\tilde{H}^{-1}\left((-\infty,h]\right)\right)$ and for any positive sequence $\varepsilon_n = O\left(n^{-1/2}\right)$,

$$\sup_{|h|\geq\mu}\left(P\left(\Gamma_{b_n,n}^{SS}\leq h|\mathcal{X}_n\right)-\Phi\left(\tilde{H}^{-1}\left((-\infty,h]\right)\right)\right)\leq$$

$$\leq\left\{\begin{array}{c}\sup_{|h|\geq\mu}\left|\begin{array}{c}P\left(\sqrt{b_n}\left(\mathbb{E}_{b_n,n}^{SS}(Z)-\mathbb{E}_n(Z)\right)\in\tilde{H}^{-1}\left((-\infty,h+\varepsilon_n]\right)|\mathcal{X}_n\right)+\\-\Phi\left(\tilde{H}^{-1}\left((-\infty,h+\varepsilon_n]\right)\right)\end{array}\right|+\\+P\left(\left|\tilde{\delta}_{b_n,n}^{SS}\right|>\varepsilon_n|\mathcal{X}_n\right)+\sup_{|h|\geq\mu}\Phi\left(\tilde{H}^{-1}\left((h-\varepsilon_n,h+\varepsilon_n]\right)\right)\end{array}\right\}$$

The upper bound is a sum of three terms. By lemma 47, the first term is $O_p\left(b_n^{-1/2}+b_n/n\right)$, by theorem 45, the second term is $o_p\left(n^{-1/2}\right)$ and by lemma 6, the third term is $O_p\left(n^{-1/2}\right)$. Thus, the whole expression is $O_p\left(b_n^{-1/2}+b_n/n\right)$. The step is completed by repeating the argument with the reverse inequality. The following step would be to show that: $\sup_{|h|\geq\mu}\left|\lim_{m\to\infty}P\left(\Gamma_m\leq h\right)-P\left(\Gamma_n\leq h\right)\right| = O_p\left(n^{-1/2}\right)$, which was shown in the proof of theorem 11. Combining both steps, we complete the result.

Part 2. The case when $\Theta_I = \varnothing$ follows from the arguments in the proof of theorem 11. $\square$

Corollary 49. *Assume (B1)-(B5), (CF), that the distribution of* $\{Y|X = x_k\}_{k=1}^{K}$ *is strongly non-lattice and let* $b_n \to \infty$ *and* $b_n/n = o(1)$*. For any* $\alpha \in [0, 0.5)$, $q^{SS}_{b_n,n}(1-\alpha) \equiv P\left(\Gamma^{SS}_{b_n,n} \leq \hat{c}^{SS}_{b_n,n}(1-\alpha)|\mathcal{X}_n\right)$. *If* $\Theta_I \neq \varnothing$, *then,*

$$\left|q^{SS}_{b_n,n}(1-\alpha) - (1-\alpha)\right| \leq O_p\left(b_n^{-1/2} + b_n/n\right)$$

Proof. This proof follows the same arguments as in corollary 42. $\square$

Proof of corolary 14. This proof follows the same arguments as in corollary 12. $\square$

Corollary 50. *Assume that the distribution of* $\{Y|X = x_k\}_{k=1}^{K}$ *is strongly non-lattice and let* $b_n \to \infty$ *and* $b_n/n = o(1)$*. Assume that* $K_1\left(\tilde{H}^{-1}\left((-\infty, c_\infty(1-\alpha)]\right)\right) > 0$*, where* $K_1 : \mathcal{C}_r \to \mathbb{R}$ *is defined as in lemma 47,* $c_\infty(1-\alpha)$ *is defined by* $P\left(\tilde{H}(\vartheta) \leq c_\infty(1-\alpha)\right) = 1-\alpha$ *and where* $\tilde{H}$ *is the function defined in theorem 36 and* $\vartheta \sim N(0, \mathbf{I}_r)$*.*

If $\Theta_I \neq \varnothing$ *then,* $\forall \varepsilon > 0$, $\exists \eta > 0$, $\exists C > 0$ *and* $\exists N \in \mathbb{N}$ *such that* $\forall n \geq N$,

$$P\left(\inf_{h \in [c_\infty(1-\alpha)-\eta, c_\infty(1-\alpha)+\eta]} \left|P\left(\Gamma^{SS}_{b_n,n} \leq h|\mathcal{X}_n\right) - \lim_{m\to\infty} P\left(\Gamma_m \leq h\right)\right| \geq C\left(b_n^{-1/2} + b_n/n\right)\right) \geq 1-\varepsilon$$

and also $\forall \varepsilon > 0$, $\exists \eta > 0$, $\exists C' > 0$ *and* $\exists N' \in \mathbb{N}$ *such that* $\forall n \geq N'$,

$$P\left(\inf_{h \in [c_\infty(1-\alpha)-\eta, c_\infty(1-\alpha)+\eta]} \left|P\left(\Gamma^{SS}_{b_n,n} \leq h|\mathcal{X}_n\right) - P\left(\Gamma_n \leq h\right)\right| \geq C'\left(b_n^{-1/2} + b_n/n\right)\right) \geq 1-\varepsilon$$

Proof. Denote γ_n as follows,

$$\gamma_n = \sup_{h \in \mathbb{R}} \left| P\left(\Gamma^{SS}_{b_n,n} \le h | \mathcal{X}_n\right) - P\left(\tilde{H}\left(\sqrt{b_n}\left(\mathbb{E}^{SS}_n(Z) - \mathbb{E}_n(Z)\right)\right) \le h | \mathcal{X}_n\right)\right|$$

By theorem 45, $\Gamma^{SS}_{b_n,n} = \tilde{H}\left(\sqrt{b_n}\left(\mathbb{E}^{SS}_n(Z) - \mathbb{E}_n(Z)\right)\right) + \tilde{\delta}^{SS}_{b_n,n}$ with $\liminf \left\{\tilde{\delta}^{SS}_{b_n,n} = 0\right\}$, a.s.. Since $\left\{\tilde{\delta}^{SS}_{b_n,n} = 0\right\}$ implies $\left\{n^{1/2}\gamma_n = 0\right\}$, it follows that $\gamma_n = o_p\left(n^{-1/2}\right)$.

Combining this result with lemma 47, and since $\forall |h| \ge \mu$, $\tilde{H}^{-1}((-\infty, h]) \in \mathcal{C}_r$, we deduce that,

$$\begin{aligned} &P\left(\Gamma^{SS}_{b_n,n} \le h | \mathcal{X}_n\right) - \lim_{m \to \infty} P\left(\Gamma_m \le h\right) = \\ &= K_1\left(\tilde{H}^{-1}((-\infty, h])\right) b_n^{-1/2} + K_2\left(\tilde{H}^{-1}((-\infty, h])\right) b_n/n + o_p\left(b_n^{-1/2} + b_n/n\right) \end{aligned}$$

uniformly in $|h| \ge \mu$.

The absolute value of the right hand side is minimized by $b_n = \Psi n^{2/3}$, where $\Psi = \Psi(h)$ is selected to minimize $\left|K_1\left(\tilde{H}^{-1}((-\infty, h])\right)\Psi(h)^{-1/2} + K_2\left(\tilde{H}^{-1}((-\infty, h])\right)\right|$, subject to $\Psi(h) > 0$. By the definition of $K_1(S)$ for $S \in \mathcal{C}_r$, by properties of the function $\tilde{H}$ and by the same arguments used in lemma 41, it is not hard to show that $K_1\left(\tilde{H}^{-1}((-\infty, h])\right)$ is continuous in h for $|h| \ge \mu$. Since $K_1\left(\tilde{H}^{-1}((-\infty, c_\infty(1-\alpha)])\right) > 0$ then, by continuity, $\exists \eta > 0$ such that $\forall h \in [c_\infty(1-\alpha) - \eta, c_\infty(1-\alpha) + \eta]$, $K_1\left(\tilde{H}^{-1}((-\infty, h])\right) > 0$ and $K_2\left(\tilde{H}^{-1}((-\infty, h])\right) > 0$, which implies that the expression to be minimized is positive.

Then, for any subsampling size, $\exists \tilde{C} > 0$ and $\exists N \in \mathbb{N} : \forall n \geq N$,

$$P\left(\inf_{h \in [c_\infty(1-\alpha)-\eta, c_\infty(1-\alpha)+\eta]} \left| \begin{array}{c} K_1\left(\tilde{H}^{-1}\left((-\infty, h]\right)\right) b_n^{-1/2} + \\ +K_2\left(\tilde{H}^{-1}\left((-\infty, h]\right)\right) b_n/n \\ +o_p\left(b_n^{-1/2} + b_n/n\right) \end{array} \right| > \tilde{C}\left(b_n^{-1/2} + b_n/n\right)\right)$$
$$> 1 - \varepsilon$$

which implies the first result, by taking $C \in \left(0, \tilde{C}\right)$. To get the second result, combine the first result with: $\sup_{|h| \geq \mu} \left|P\left(\Gamma_n \leq h\right) - \lim_{m \to \infty} P\left(\Gamma_m \leq h\right)\right| = O_p\left(n^{-1/2}\right)$ and choose $C' \in (0, C)$. □

Lemma 51. *For any μ_L, μ_H such that $(\mu_L, \mu_H) \subset (\mu, 1)$, let h_L and h_H be such that $P\left(\tilde{H}(\vartheta) \leq h_L\right) = \mu_L$ and $P\left(\tilde{H}(\vartheta) \leq h_H\right) = \mu_H$, where $\tilde{H}$ is the function defined in theorem 36 and $\vartheta \sim N(0, \mathbf{I}_r)$. If $(1-\alpha) \in (\mu_L, \mu_H)$, then,*

$$\lim_{n \to \infty} P\left(\hat{c}_{b_n,n}^{SS}(1-\alpha) \in (h_L, h_H)\right) = 1$$

Proof. This follows from corollary 48 and the arguments in corollary 42. □

Corollary 52. *Assume that the distribution of $\{Y|X = x_k\}_{k=1}^K$ is continuous and let $b_n \to \infty$ and $b_n/n = o(1)$. Let $(1-\alpha) \in [0.72, 1)$ denote the level of interest and let $c_\infty(1-\alpha)$ be defined by $P\left(\tilde{H}(\vartheta) \leq c_\infty(1-\alpha)\right) = 1-\alpha$, where $\tilde{H}$ is the function defined in theorem 36 and $\vartheta \sim N(0, \mathbf{I}_r)$. Moreover, assume that $K_1\left(\tilde{H}^{-1}\left((-\infty, c_\infty(1-\alpha)]\right)\right) > 0$ where K_1 is defined as in lemma 47. If the identified set is non-empty, then, $\exists C > 0$*

and $\exists N \in \mathbb{N}$ *such that* $\forall n \geq N$,

$$\left|P\left(\Theta_I \subset C^{SS}_{b_n,n}(1-\alpha)\right)-(1-\alpha)\right| \geq C\left(b_n^{-1/2}+b_n/n\right)$$

Proof. By the result in theorem 45 and since $(1-\alpha) \in [0, 0.5)$, the $(1-\alpha)$ quantile of $P\left(\Gamma^{SS}_{b_n,n} \leq h|\mathcal{X}_n\right)$ will be positive, w.p.a.1. By properties of the function $\tilde{H}$, w.p.a.1, the $(1-\alpha)$ quantile of $P\left(\Gamma^{SS}_{b_n,n} \leq h|\mathcal{X}_n\right)$ can represent a level higher than $(1-\alpha)$ only if $\mathbb{E}^{SS}_{b_n,n}(Z)$ coincides for at least two subsamples. If the original sample is continuously distributed, then no two subsamples will coincide, a.s.. Therefore, for any $\varepsilon > 0$, $\exists N \in \mathbb{N}$ such that $\forall n \geq N_0$, $P\left(q^{SS}_{b_n,n}(1-\alpha)=(1-\alpha)\right) > 1-\varepsilon/2$.

By corollary 50 and lemma 51, $\exists \eta > 0$ such that $\forall \varepsilon > 0$, $\exists K > 0$ and $N_1 \in \mathbb{N}$ such that $\forall n \geq N_1$,

$$\begin{aligned} P\left(\inf_{h \in [c_\infty(1-\alpha)-\eta, c_\infty(1-\alpha)+\eta]}\left|\begin{array}{c} P\left(\Gamma^{SS}_{b_n,n} \leq h|\mathcal{X}_n\right)+ \\ -P\left(\Gamma_n \leq h\right) \end{array}\right| \geq C\left(b_n^{-1/2}+b_n/n\right)\right) &\geq 1-\varepsilon/4 \\ P\left(\hat{c}^{SS}_{b_n,n}(1-\alpha) \notin [c_\infty(1-\alpha)-\eta, c_\infty(1-\alpha)+\eta]\right) &\leq \varepsilon/4 \end{aligned}$$

Hence $\forall n \geq N_1$,

$$\begin{aligned}
&1-\varepsilon/4 \leq \\
&\leq P\left(\inf_{h\in[c_\infty(1-\alpha)-\eta,c_\infty(1-\alpha)+\eta]}\left|P\left(\Gamma^{SS}_{b_n,n}\leq h|\mathcal{X}_n\right)-P\left(\Gamma_n\leq h\right)\right|\geq C\left(b_n^{-1/2}+b_n/n\right)\right) \\
&\leq \left\{\begin{array}{c} P\left(\left|\begin{array}{c} P\left(\Gamma^{SS}_{b_n,n}\leq \hat{c}^{SS}_{b_n,n}\left(1-\alpha\right)|\mathcal{X}_n\right)+ \\ -P\left(\Gamma_n\leq \hat{c}^{SS}_{b_n,n}\left(1-\alpha\right)\right)\end{array}\right|\geq C\left(b_n^{-1/2}+b_n/n\right)\right)+ \\ +P\left(\hat{c}^{SS}_{b_n,n}\left(1-\alpha\right)\notin\left[c_\infty\left(1-\alpha\right)-\eta,c_\infty\left(1-\alpha\right)+\eta\right]\right)\end{array}\right\} \\
&\leq \left\{P\left(\begin{array}{c}\left|P\left(\Gamma^{SS}_{b_n,n}\leq \hat{c}^{SS}_{b_n,n}\left(1-\alpha\right)|\mathcal{X}_n\right)-P\left(\Gamma_n\leq \hat{c}^{SS}_{b_n,n}\left(1-\alpha\right)\right)\right| \\ \geq C\left(b_n^{-1/2}+b_n/n\right)\end{array}\right)+\varepsilon/4\right\}
\end{aligned}$$

which implies that $\forall n \geq N_1$,

$$P\left(\left|P\left(\Theta_I\subset\hat{C}^{SS}_{b_n,n}\left(1-\alpha\right)\right)-q^{SS}_{b_n,n}\left(1-\alpha\right)\right|\geq C\left(b_n^{-1/2}+b_n/n\right)\right)\geq 1-\varepsilon/2$$

Therefore: $\forall\varepsilon>0$, $\exists N=\max\{N_0,N_1\}\in\mathbb{N}:\forall n\geq N$,

$$\begin{aligned}
\varepsilon &\geq \left\{\begin{array}{c} P\left(\left|P\left(\Theta_I\subset\hat{C}^{SS}_{b_n,n}\left(1-\alpha\right)\right)-q^{SS}_{b_n,n}\left(1-\alpha\right)\right|<C\left(b_n^{-1/2}+b_n/n\right)\right)+ \\ +P\left(q^{SS}_{b_n,n}\left(1-\alpha\right)\neq\left(1-\alpha\right)\right)\end{array}\right\} \\
&\geq P\left(\left|P\left(\Theta_I\subset\hat{C}^{SS}_{b_n,n}\left(1-\alpha\right)\right)-\left(1-\alpha\right)\right|<C\left(b_n^{-1/2}+b_n/n\right)\right)
\end{aligned}$$

Since the event inside the probability is non-random, then $\exists N\in\mathbb{N}$ such that $\forall n\geq N$ it does not occur. □

A.2.6.2. Subsampling with no recentering (a la CHT).

Theorem 53. *Let* $\{b_n\}_{n=1}^{+\infty}$ *be such that* $b_n\to\infty$ *and* $b_n/n=o(1)$.

If we assume (A1)-(A4), (CF') and that $\Theta_I \neq \varnothing$*, then,* $\Gamma^{SS,CHT}_{b_n,n} = H\left(v^{SS}_{b_n,n}(m_\theta)\right) + \delta^{SS,CHT}_{b_n,n}$*, where,*

(1) *for any* $\varepsilon > 0$, $\lim_{n\to\infty} P^*\left(\left|\delta^{SS,CHT}_{b_n,n}\right| > \varepsilon|\mathcal{X}_n\right) = 0$, *a.s..*

(2) $\left\{v^{SS}_{b_n,n}(m_\theta)|\mathcal{X}_n\right\} : \Omega_n \to l^\infty_J(\Theta)$ *is an empirical process that converges weakly to the same Gaussian process as in theorem 36, a.s..*

(3) $H : l^\infty_J(\Theta) \to \mathbb{R}$ *is the same function as in theorem 36.*

Proof. This proof is similar to the one in theorem 45. We only point out how the proof changes given that there is no recentering.

By theorem 45 and the properties of the function H, it follows that,

$$\Gamma^{SS,CHT}_{b_n,n} = H\left(\left\{v^{SS}_{b_n,n}(m_{j,\theta}) + \sqrt{b_n}\left(\bar{m}_j(\theta) - \mathbb{E}(m_j(\theta))\right)\right\}_{j=1}^{J}\right) + \delta^{SS}_{b_n,n}$$

where $\forall\varepsilon > 0$, $\lim_{n\to\infty} P^*\left(\left|\delta^{SS,CHT}_{b_n,n}\right| > \varepsilon/2|\mathcal{X}_n\right) = 0$, a.s.. Therefore, in order to complete this step, it suffices to show that $\forall\varepsilon > 0$,

$$\lim_{n\to\infty} P^*\left(\left|\begin{array}{c} H\left(\left\{v^{SS}_{b_n,n}(m_{j,\theta}) + \sqrt{b_n}\left(\bar{m}_j(\theta) - \mathbb{E}(m_j(\theta))\right)\right\}_{j=1}^{J}\right) + \\ -H\left(\left\{v^{SS}_{b_n,n}(m_{j,\theta})\right\}_{j=1}^{J}\right) \end{array}\right| > \varepsilon/2\right) = 0, \text{ a.s.}$$

Since the function H is assumed to be continuous, it is sufficient to show that $\forall\eta > 0$, $\liminf\left\{\sup_{\theta\in\Theta}\left\|\sqrt{b_n}\left(\bar{m}(\theta) - \mathbb{E}(m(\theta))\right)\right\| < \eta\right\}$, a.s.. Since $(\ln\ln n)\, b_n/n = o(1)$, this follows from the LIL for empirical processes (see, e.g., Kuelbs [**39**]). □

Lemma 54 (Comparison with CHT's subsampling)**.** *Assume (A1)-(A4), (CF'). Let* $\{b_n\}_{n=1}^{+\infty}$ *be such that* $b_n \to \infty$ *and* $b_n/n = o(1)$ *and let* $\Gamma^{SS,CHT}_{b_n,n}$ *denote the subsampling* *statistic proposed by CHT* [**23**]. *If* $\Theta_I \neq \varnothing$, *then* $\forall \varepsilon > 0$,

$$P\left(\lim_{n\to\infty} P^*\left(\left|\Gamma^{SS,CHT}_{b_n,n} - \Gamma^{SS}_{b_n,n}\right| > \varepsilon|\mathcal{X}_n\right) = 0\right) = 1$$

If $\Theta_I = \varnothing$, *then,*

$$P\left(\liminf\left\{\Gamma^{SS}_{b_n,n} \leq \Gamma^{SS,CHT}_{b_n,n}\right\}\right) = 1$$

Proof. Part 1. Follows directly from theorem 53.

Part 2. Since $\Gamma^{SS,CHT}_{n,b_n} \geq 0$ and since $\liminf\left\{\Gamma^{SS}_{n,b_n} = 0\right\}$, a.s. then it follows that $\liminf\left\{\Gamma^{SS}_{n,b_n} \leq \Gamma^{SS,CHT}_{n,b_n}\right\}$, a.s.. □

A.2.7. Asymptotic approximation

Theorem 55. *Assume (A1)-(A4), (CF') and that* $\Theta_I \neq \varnothing$. *Then,* $\Gamma^{AA}_n = H\left(\hat{Z}(\theta)\right) + \delta^{AA}_n$, *where,*

(1) *for any* $\varepsilon > 0$, $\lim_{n\to\infty} P^*\left(\left|\delta^{AA}_n\right| > \varepsilon|\mathcal{X}_n\right) = 0$, *a.s.,*

(2) $\left\{\hat{Z}(\theta)|\mathcal{X}_n\right\} : \Omega_n \to l^\infty_J(\Theta)$ *is an empirical process that converges weakly to the same Gaussian process as in theorem 36, a.s.,*

(3) $H : l^\infty_J(\Theta) \to \mathbb{R}$ *is the same function as in theorem 36.*

If, instead, we assume (B1)-(B4), (CF), $\Theta_I \neq \varnothing$ *and we construct* Γ^{AA}_n *by simulating* $\hat{Z}$ *independently from a zero mean normally distributed vector with variance covariance matrix* $\hat{\Sigma}$ *then, for r as in theorem 36,* $\Gamma^{AA}_n = \tilde{H}\left(\hat{\zeta}\right) + \tilde{\delta}^{AA}_n$, *where,*

(1) *for any* $\varepsilon_n = O\left(n^{-1/2}\right)$, $P\left(\left|\tilde{\delta}^{AA}_n\right| > \varepsilon_n|\mathcal{X}_n\right) = o\left(n^{-1/2}\right)$, *a.s.,*

(2) $\left\{\hat{\zeta}|\mathcal{X}_n\right\} : \Omega_n \to \mathbb{R}^r$ *is a zero mean normally distributed vector with variance covariance matrix* $\hat{V}$. *If we also assume (B5),* $\left\|\hat{V} - \mathbf{I}_r\right\| \leq O_p\left(n^{-1/2}\right)$,

(3) $\tilde{H} : \mathbb{R}^r \to \mathbb{R}$ is the same function as in theorem 36.

If, instead, we assume (A1)-(A4), (CF') and that $\Theta_I = \varnothing$ then, $\liminf \left\{\Gamma_n^{AA} = 0\right\}$, *a.s..*

Proof. This proof follows the proof of theorem 36 very closely. $\square$

Theorem 56 (Consistency of asymptotic approximation excluding zero)**.** *Assume (A1)-(A4) and (CF'). If $\Theta_I \neq \varnothing$ then, for any $\mu > 0$,*

$$P\left(\lim_{n\to\infty} \sup_{|h|\geq\mu} \left|P\left(\Gamma_n^{AA} \leq h|\mathcal{X}_n\right) - \lim_{m\to\infty} P\left(\Gamma_m \leq h\right)\right| = 0\right) = 0$$

and if $\Theta_I = \varnothing$ then,

$$P\left(\liminf\left\{\sup_{h\in\mathbb{R}} \left|P\left(\Gamma_n^{AA} \leq h|\mathcal{X}_n\right) - \lim_{m\to\infty} P\left(\Gamma_m \leq h\right)\right| = 0\right\}\right) = 1$$

Proof. This proof follows the arguments of the proof of theorem 9. $\square$

Proof of corollary 15. This proof follows the arguments of the proof of theorem 10. $\square$

Theorem 57 (Rate of convergence - asymptotic approximation)**.** *Assume (B1)-(B4) and (CF). If the identified set is non-empty then,*

$$\sup_{|h|\geq\mu} \left|P\left(\Gamma_n^{AA} \leq h|\mathcal{X}_n\right) - P\left(\Gamma_n \leq h\right)\right| \leq O_p\left(n^{-1/2}\right)$$

If the identified set is empty then,

$$P\left(\liminf\left\{\sup_{h\in\mathbb{R}} \left|P\left(\Gamma_n^{AA} \leq h|\mathcal{X}_n\right) - P\left(\Gamma_n \leq h\right)\right| = 0\right\}\right) = 1$$

Proof. This proof follows the arguments of the proof of corollary 11. □

Proof of corollary 16. This proof follows the arguments of the proof of corollary 12. □

A.3. Confidence sets for each element of the identified set

A.3.1. Differences with the naive bootstrap

The bootstrap procedure to cover each element of the identified set differs qualitatively from replacing the subsampling scheme in CHT [**23**] with the traditional bootstrap. Exactly as in section A.2.1, we refer to the procedure of replacing the subsampling scheme in CHT [**23**] with the traditional bootstrap as *naive bootstrap*.

In general the naive bootstrap will produce inconsistent inference. Consider the following example. Let $\Theta_I = \{\theta \in \Theta : \mathbb{E}(Y) \leq \theta\}$ and assume that $Y \sim N(0,1)$ and $0 \in \Theta$. The criterion function is given by $Q(\theta) = G\left([\mathbb{E}(Y) - \theta]_+\right)$, and its sample analogue is $Q_n(\theta) = G\left([\mathbb{E}_n(Y) - \theta]_+\right)$. Hence, the distribution of interest is given by: $\Gamma_n(\theta) = \sqrt{n} Q_n(\theta) \xrightarrow{d} G\left([\vartheta]_+\right) 1[\theta = 0]$, where $\vartheta \sim N(0,1)$.

Consider performing inference for $\theta = 0$ combining the bootstrap and the analogy principle criterion function, that is, the naive bootstrap. In this setting, we will obtain the following statistic,

$$\Gamma_n^{*,AP}(0) = G\left(\left[\sqrt{n}\left(\mathbb{E}_n^*(Y) - \mathbb{E}_n(Y)\right) + \sqrt{n}\left(\mathbb{E}_n(Y) - \mathbb{E}(Y)\right)\right]_+\right)$$

For any $c > 0$, consider the following sets,

$$\begin{aligned} A &= \left\{ \left\{ \sqrt{n} \left(\mathbb{E}_n^* \left(Y \right) - \mathbb{E}_n \left(Y \right) \right) | \mathcal{X}_n \right\} \xrightarrow{d} N \left(0, 1 \right) \right\} \\ B &= \liminf \left\{ \sqrt{n} \left(\mathbb{E}_n \left(Y \right) - \mathbb{E} \left(Y \right) \right) < -c \right\} \end{aligned}$$

Let $\omega \in \{A \cap B\}$. Since $\omega \in B$, $\exists N \in \mathbb{N}$ such that $\forall n \geq N$, the conditional distribution of $\Gamma_n^{*,AP}\left(0\right)$ satisfies,

$$\Gamma_n^{*,AP}\left(0\right) \leq G\left(\left[\sqrt{n}\left(\mathbb{E}_n^*\left(Y\right) - \mathbb{E}_n\left(Y\right)\right) - c\right]_+\right)$$

Since $\omega \in A$, the conditional distribution of the right hand side converges weakly to $G\left(\left[\vartheta - c\right]_+\right)$, where $\vartheta \sim N\left(0,1\right)$. By the LIL we can deduce that $P\left(B\right) = 1$ and by theorem 2.1 in Bickel and Freedman [**13**], $P\left(A\right) = 1$. Hence, this procedure leads to inconsistent inference.

Instead, suppose that we use our proposed bootstrap procedure to perform inference for $\theta = 0$. In this case,

$$\Gamma_n^*\left(0\right) = 1\left[0 \in \hat{\Theta}_I\left(\tau_n\right)\right] G\left(\left[\sqrt{n}\left(\mathbb{E}_n^*\left(Y\right) - \mathbb{E}_n\left(Y_1\right)\right)\right]_+ 1\left[\sqrt{n}\mathbb{E}_n\left(Y\right) \geq -\tau_n\right]\right)$$

Let $\omega \in \{A \cap B'\}$ where B'' is given by,

$$B' = \liminf\left\{\left|\sqrt{n}\mathbb{E}_n\left(Y\right)\right| \leq \tau_n\right\}$$

Since $\omega \in B'$, $\exists N \in \mathbb{N}$ such that $\forall n \geq N$, the conditional distribution of $\Gamma_n^{*,AP}\left(0\right)$ satisfies,

$$\Gamma_n^*\left(0\right) = G\left(\left[\sqrt{n}\left(\mathbb{E}_n^*\left(Y\right) - \mathbb{E}_n\left(Y_1\right)\right)\right]_+\right)$$

Since $\omega \in A$, the conditional distribution of the right hand side converges weakly to $G\left([\vartheta]_+\right)$, where $\vartheta \sim N(0,1)$. By the LIL we can deduce that $P(B') = 1$ and by theorem 2.1 in Bickel and Freedman [**13**], $P(A) = 1$. Hence, the proposed bootstrap inference leads to consistent inference.

Finally, we note that the inconsistency of the naive bootstrap is a consequence of using bootstrap instead of subsampling. When we use subsampling (with sample size b_n) with the analogue principle criterion function, we obtain the following statistic,

$$\Gamma_{b_n,n}^{SS,AP}(0) = G\left(\left[\sqrt{b_n}\left(\mathbb{E}_{b_n,n}^{SS}(Y) - \mathbb{E}_n(Y)\right) + \sqrt{b_n}\left(\mathbb{E}_n(Y) - \mathbb{E}(Y)\right)\right]_+\right)$$

For any $\varepsilon > 0$, let B'' be given by,

$$B'' = \liminf\left\{\left|\sqrt{b_n}\left(\mathbb{E}_n(Y) - \mathbb{E}(Y)\right)\right| \leq (1+\varepsilon)\sqrt{2\left(b_n \ln\ln n\right)/n}\right\}$$

Let $\omega \in \{A \cap B''\}$. If $\left(b_n \ln\ln n\right)/n = o(1)$ and using previous arguments, $\Gamma_{b_n,n}^{SS,AP}(0)$ converges weakly to $G\left([\zeta]_+\right)$, where $\zeta \sim N(0,1)$. Since $P(A \cap B'') = 1$, we deduce subsampling generates consistent inference in level.

A.3.2. Representation results

Theorem 58. *Assume (C1)-(C4), (CF').*

(1) *If* $\theta \in \partial\Theta_I$ *then, for some* $r \in \mathbb{N} \cap [1, J]$, $\Gamma_n(\theta) = H_\theta\left(\sqrt{n}\bar{Z}(\theta)\right) + \delta_n(\theta)$, *where:*
   (a) *for any* $\varepsilon_n = O\left(n^{-1/2}\right)$, $P\left(\left|\delta_n(\theta)\right| > \varepsilon_n\right) = o\left(n^{-1/2}\right)$,

   (b) $\bar{Z}(\theta) : \Omega_n \to \mathbb{R}^r$ *is a zero mean sample average of* n *i.i.d. observations from a distribution with non-singular variance-covariance matrix* $V = \mathbf{I}_r$ *. If assumption (C5) is added, then this distribution has finite fourth moments,*

   (c) $H_\theta : \mathbb{R}^r \to \mathbb{R}$ *is continuous, non-negative, weakly quasi-convex and homogeneous of degree* β*.* $H_\theta(y) = 0$ *implies for some non-zero vector* $b \in \mathbb{R}^r$, $b'y \leq 0$*. For any* $\mu > 0$*, any* $|h| \geq \mu > 0$ *and any sequence* $\varepsilon_n = o(1)$, $\left\{H_\theta^{-1}(\{h\}^{\varepsilon_n}) \cap \|y\| \leq O\left(\sqrt{g_n}\right)\right\} \subseteq \left\{H_\theta^{-1}(\{h\})\right\}^{\eta_n}$ *where* $\eta_n = O\left(\sqrt{g_n}\varepsilon_n\right)$*. If we add (CF), then any* $\mu > 0$*, any* $|h| \geq \mu > 0$ *and any sequence* $\varepsilon_n = o(1)$, $\left\{H_\theta^{-1}(\{h\}^{\varepsilon_n})\right\} \subseteq \left\{H_\theta^{-1}(\{h\})\right\}^{\gamma_n}$ *where* $\gamma_n = O(\varepsilon_n)$.

(2) *If* $\theta \in Int(\Theta_I)$, *then* $\liminf \{\Gamma_n(\theta) = 0\}$, *a.s..*

(3) *If* $\theta \notin \Theta_I$*, then* $\lim \Gamma_n(\theta) = +\infty$*, a.s..*

Proof. This proof follows the proof of theorems 36 and 37 very closely. □

Theorem 59. *Assume (C1)-(C4), (CF').*

(1) *If* $\theta \in \partial\Theta_I$ *then, for* r *as in theorem 58,* $\Gamma_n^*(\theta) = H_\theta\left(\sqrt{n}\bar{Z}^*(\theta)\right) + \delta_n^*(\theta)$*, where:*

   (a) *for any* $\varepsilon_n = O\left(n^{-1/2}\right)$, $P\left(|\delta_n^*(\theta)| > \varepsilon_n | \mathcal{X}_n\right) = o\left(n^{-1/2}\right)$, *a.s.,*

   (b) $\left\{\bar{Z}^*(\theta) | \mathcal{X}_n\right\} : \Omega_n \to \mathbb{R}^r$ *is a zero mean sample average of* n *i.i.d. observations from a distribution with a variance-covariance matrix* $\hat{V}$ *which is a.s. non-singular. If assumption (C5) is added, then* $\left\|\hat{V} - \mathbf{I}_r\right\| \leq O_p\left(n^{-1/2}\right)$ *and the distribution has finite fourth moments, a.s.,*

   (c) $H_\theta : \mathbb{R}^r \to \mathbb{R}$ *is the same function as in theorem 58.*

(2) *If* $\theta \in Int(\Theta_I)$, *then* $\liminf \{\Gamma_n^*(\theta) = 0\}$, *a.s..*

(3) *If* $\theta \notin \Theta_I$*, then* $\liminf \{\Gamma_n^*(\theta) = 0\}$, *a.s..*

Proof. This proof follows the proof of theorems 38 and 39 very closely. □

A.3.3. Consistency results

Theorem 60 (Bootstrap consistency on any set excluding zero)**.** *Assume (C1)-(C4), (CF').*

(1) *If* $\theta \in \partial\Theta_I$ *then,* $\forall \mu > 0$,

$$P\left(\lim_{n\to\infty}\sup_{|h|\geq\mu}\left|P\left(\Gamma_n^*\left(\theta\right)\leq h|\mathcal{X}_n\right)-\lim_{m\to\infty}P\left(\Gamma_m\left(\theta\right)\leq h\right)\right|=0\right)=1$$

(2) *If* $\theta \in Int\left(\Theta_I\right)$ *then,*

$$P\left(\lim_{n\to\infty}\sup_{h\in\mathbb{R}}\left|P\left(\Gamma_n^*\left(\theta\right)\leq h|\mathcal{X}_n\right)-\lim_{m\to\infty}P\left(\Gamma_m\left(\theta\right)\leq h\right)\right|=0\right)=1$$

Proof. This proof follows the proof of theorem 9. □

Proof of corollary 17. Part 1. Follows the proof of theorem 10.

Part 2. By the case under consideration, $\exists\varpi > 0$ such that $\max_{j\leq J}\mathbb{E}\left(m_j\left(Z,\theta\right)\right)\leq-\varpi$. Hence, by the LIL, $\liminf\left\{\max_{j\leq J}\sqrt{n}\mathbb{E}_n\left(m_j\left(Z,\theta\right)\right)<-\tau_n\right\}$, a.s.. Since the event $\left\{\max_{j\leq J}\sqrt{n}\mathbb{E}_n\left(m_j\left(Z,\theta\right)\right)<-\tau_n\right\}$ implies the events $\left\{\sqrt{n}Q_n\left(\theta\right)=0\right\}$ and $\left\{\Gamma_n^*\left(\theta\right)=0\right\}$, $P\left(\liminf\left\{\theta\in\hat{C}_n^B\left(1-\alpha\right)\right\}\right)=1$.

Part 3. $\exists j=1,2,...,J$, $\mathbb{E}\left(m_j\left(Z,\theta\right)\right)=\varpi>0$. By LIL, $\liminf\left\{\sqrt{n}\mathbb{E}_n\left(m_j\left(Z,\theta\right)\right)>\tau_n\right\}$, a.s.. The event $\left\{\sqrt{n}\mathbb{E}_n\left(m_j\left(Z,\theta\right)\right)>\tau_n\right\}$ implies $\theta\notin\hat{\Theta}_I\left(\tau_n\right)$ which, in turn, implies that $\Gamma_n^*\left(\theta\right)=0$. Also, $\left\{\sqrt{n}\mathbb{E}_n\left(m_j\left(Z,\theta\right)\right)>\tau_n\right\}$ implies that $\sqrt{n}Q_n\left(\theta\right)>0$. Therefore, $P\left(\liminf\left\{\theta\notin\hat{C}_n^B\left(1-\alpha\right)\right\}\right)=1$. □

A.3.4. Rates of convergence results

Theorem 61 (Rates of convergence). *Assume (C1)-(C5), (CF'). Then,*

(1) *If* $\theta \in \partial \Theta_I$ *then* $\forall \mu > 0$,

$$\sup_{|h| \geq \mu} \left| P\left(\Gamma_n^*\left(\theta\right) \leq h | \mathcal{X}_n\right) - P\left(\Gamma_n\left(\theta\right) \leq h\right)\right| \leq O_p\left(n^{-1/2} \ln n\right)$$

and if we further assume (CF), then $\forall \mu > 0$,

$$\sup_{|h| \geq \mu} \left| P\left(\Gamma_n^*\left(\theta\right) \leq h | \mathcal{X}_n\right) - P\left(\Gamma_n\left(\theta\right) \leq h\right)\right| \leq O_p\left(n^{-1/2}\right)$$

(2) *If* $\theta \in Int\left(\Theta_I\right)$, *then,*

$$P\left(\liminf \left\{\sup_{h \in \mathbb{R}} \left| P\left(\Gamma_n^*\left(\theta\right) \leq h | \mathcal{X}_n\right) - P\left(\Gamma_n\left(\theta\right) \leq h\right)\right| = 0\right\}\right) = 1$$

Proof of theorem 61. This proof follows the proof of theorem 11. □

Proof of corollary 18. Part 1. Follows the proof for corollary 12.

Part 2. $\exists \varpi > 0$ such that $\max_{j=1,\ldots,J} \mathbb{E}\left(m_j\left(Z,\theta\right)\right) \leq -\varpi$. By definition, the event $\left\{\theta \notin \hat{C}_n^B\left(1-\alpha\right)\right\}$ equals the event $\left\{\sqrt{n}Q_n\left(\theta\right) > \hat{c}_n^B\left(\theta, 1-\alpha\right)\right\}$. The latter event implies that $\exists j = 1, \ldots, J$ such that $\left\{\sqrt{n}\mathbb{E}_n\left(m_j\left(Z,\theta\right)\right) > 0\right\}$, which, in turn, implies that $\exists j = 1, \ldots, J$ such that the event $\left\{\sqrt{n}\left(\mathbb{E}_n\left(m_j\left(Z,\theta\right)\right) - \mathbb{E}\left(m_j\left(Z,\theta\right)\right)\right) > \sqrt{n}\varpi\right\}$ occurs. Thus, it suffices to show that $\forall j = 1, \ldots, J$, $P\left(\sqrt{n}\left(\mathbb{E}_n\left(m_j\left(Z,\theta\right)\right) - \mathbb{E}\left(m_j\left(Z,\theta\right)\right)\right) > \sqrt{n}\varpi\right) = O\left(n^{-1}\right)$, which follows from Chebyshev's inequality.

Part 3. $\exists j = 1, 2, \ldots, J$, $\mathbb{E}\left(m_j\left(Z,\theta\right)\right) = \varpi > 0$. By definition, $\left\{\theta \in \hat{C}_n^B\left(1-\alpha\right)\right\}$ is equivalent to $\left\{\sqrt{n}Q_n\left(\theta\right) \leq \hat{c}_n^B\left(\theta, 1-\alpha\right)\right\}$, which implies that either $\left\{\sqrt{n}Q_n\left(\theta\right) = 0\right\}$ or $\left\{\hat{c}_n^B\left(\theta, 1-\alpha\right) > 0\right\}$. Notice that $\left\{\sqrt{n}Q_n\left(\theta\right) = 0\right\}$ implies $\left\{\sqrt{n}\mathbb{E}_n\left(m_j\left(Z,\theta\right)\right) \leq 0\right\}$ and

$\left\{\hat{c}_n^B(\theta, 1-\alpha) > 0\right\}$ implies $\left\{\sqrt{n}\mathbb{E}_n(m_j(Z,\theta)) \leq \tau_n\right\}$. Since $\forall n \in \mathbb{N}$, $\tau_n > 0$, their union implies $\left\{\sqrt{n}\mathbb{E}_n(m_j(Z,\theta)) \leq \tau_n\right\}$. Thus, $P\left(\sqrt{n}\left(\mathbb{E}_n(m_j(Z,\theta)) - \mathbb{E}(m_j(Z,\theta))\right) > \sqrt{n}\frac{\varpi}{2}\right)$ $= O\left(n^{-1}\right)$, which follows from Chebyshev's inequality. □

A.3.5. Subsampling procedure

(1) Choose a positive sequence $\{\tau_n\}_{n=1}^{+\infty}$ such that $\sqrt{\ln\ln n}/\tau_n = o(1)$ and $\tau_n/\sqrt{n} = o(1)$ a.s.,

(2) Estimate of the identified set with,

$$\hat{\Theta}_I(\tau_n) = \left\{\theta \in \Theta : \left\{\mathbb{E}_n(m_j(Z,\theta)) \leq \tau_n/\sqrt{n}\right\}_{j=1}^J\right\}$$

(3) For every $\theta \in \Theta$ and for $s = 1, 2, ..., S$, repeat the following steps,

   (a) Construct subsamples of size b_n (with $b_n \to \infty$ and $b_n/n = o(1)$) by sampling randomly without replacement from the data. For each subsample, denote the observations by $\left\{Z_i^{SS}\right\}_{i=1}^{b_n}$, and compute $\mathbb{E}_{b_n,n}^{SS}(m_j(Z,\theta)) = b_n^{-1}\sum_{i=1}^{b_n} m\left(Z_i^{SS},\theta\right)$.

   (b) For each subsample, compute,

$$\Gamma_{b_n,n}^{SS}(\theta) = 1\left[\theta \in \hat{\Theta}_I(\tau_n)\right] * G\left(\left\{\begin{array}{c} \left[\sqrt{b_n}\left(\mathbb{E}_{b_n,n}^{SS}(m_j(Z,\theta)) - \mathbb{E}_n(m_j(Z,\theta))\right)\right]_+ * \\ *1\left[\mathbb{E}_n(m_j(Z,\theta)) \geq -\tau_n/\sqrt{n}\right] \end{array}\right\}_{j=1}^J\right)$$

(4) Let $\hat{c}_{b_n,n}^{SS}(\theta, 1-\alpha)$ be the $(1-\alpha)$ quantile of the simulated distribution of $\Gamma_{b_n,n}^{SS}(\theta)$, simulated with arbitrary accuracy from the step 3,

(5) The asymptotic approximation estimate of the $(1-\alpha)$ coverage region for the each parameter in the identified set is given by,

$$\hat{C}^{SS}_{b_n,n}\left(1-\alpha\right)=\left\{\theta\in\Theta:\sqrt{n}Q_n\left(\theta\right)\leq\hat{c}^{SS}_{b_n,n}\left(\theta,1-\alpha\right)\right\}$$

A.3.6. Asymptotic approximation

(1) Choose a positive sequence $\{\tau_n\}_{n=1}^{+\infty}$ such that $\sqrt{\ln\ln n}/\tau_n=o\left(1\right)$ and $\tau_n/\sqrt{n}=o\left(1\right)$ a.s.,

(2) Estimate of the identified set with,

$$\hat{\Theta}_I\left(\tau_n\right)=\left\{\theta\in\Theta:\left\{\mathbb{E}_n\left(m_j\left(Z,\theta\right)\right)\leq\tau_n/\sqrt{n}\right\}_{j=1}^{J}\right\}$$

(3) For every $\theta\in\Theta$ and for $s=1,2,...,S$, repeat the following steps,

  (a) Obtain random observation from the $N\left(\vec{0},\hat{\Sigma}\left(\theta\right)\right)$, denoted by $\hat{Z}\left(\theta\right)$.

  (b) Compute,

$$\Gamma^{AA}_n\left(\theta\right)=1\left[\theta\in\hat{\Theta}_I\left(\tau_n\right)\right]*G\left(\left\{\left[\hat{Z}_j\left(\theta\right)\right]_+*1\left[\mathbb{E}_n\left(m_j\left(Z,\theta\right)\right)\geq-\tfrac{\tau_n}{\sqrt{n}}\right]\right\}_{j=1}^{J}\right)$$

(4) Let $\hat{c}^{AA}_n\left(\theta,1-\alpha\right)$ be the $(1-\alpha)$ quantile of the simulated distribution of $\Gamma^{AA}_n\left(\theta\right)$, simulated with arbitrary accuracy from the step 3,

(5) The asymptotic approximation estimate of the $(1-\alpha)$ coverage region for each parameter in the identified set is given by,

$$\hat{C}^{AA}_n\left(1-\alpha\right)=\left\{\theta\in\Theta:\sqrt{n}Q_n\left(\theta\right)\leq\hat{c}^{AA}_n\left(\theta,1-\alpha\right)\right\}$$

A.3.7. Monte Carlo simulations

In order to evaluate the finite sample behavior of the inferential methods, we construct confidence sets that cover each element of the identified set for each of the designs of section 1.4.4. For a detailed description of the four Monte Carlo designs, as well as comments on the choice of $\{\tau_n\}_{n=1}^{+\infty}$ and the subsampling size, the reader is referred to that section.

Our objective is to cover each element of the identified set. In order to evaluate the performance of each of the inferential methods, we compare the coverage of one point in the identified set. If this point is chosen in the interior of the identified set, then the coverage will converge to one at a relatively fast rate, regardless of the method used. In order to make the comparison interesting, we choose a point in the boundary of the identified set, where the inferential methods differ in their convergence rates.

A.3.7.1. Design 1. Table A.1 compares the coverage probability for $\beta = (-0.5, 0)$. The table shows that the subsampling procedures lack a satisfactory finite sample performance. The subsampling procedure proposed by CHT [**23**] suffers from overcoverage whereas the subsampling procedure of section 1.4.3.1 suffers from undercoverage. In small samples, CHT's [**23**] subsampling might be affected by the same problems as the boundary problems affecting the naive bootstrap. The naive bootstrap procedure suffers from significant overcoverage. This is to be expected, given that it is inconsistent.

Our bootstrap and our asymptotic approximation procedures present a satisfactory finite sample performance.

	Empirical coverage			
Procedure	75%	90%	95%	99%
CHT's subsampling ($b_n = n/2$)	93.5%***	99.5%***	99.9%***	100%***
CHT's subsampling ($b_n = n/3$)	89.2%***	98.7%***	99.5%***	100%***
Our subsampling ($b_n = n/2, \tau_n = \ln \ln n$)	64.1%***	79.0%***	87.4%***	94.2%***
Our subsampling ($b_n = n/3, \tau_n = \ln \ln n$)	67.6%***	84.5%***	90.8%***	95.9%***
Naive bootstrap	100%***	100%***	100%***	100%***
Our bootstrap ($\tau_n = \ln \ln n$)	75.0%	91.2%	94.5%	98.5%
Our asymptotic approximation ($\tau_n = \ln \ln n$)	78.5%***	92.9%**	96.6%***	99.3%

Table A.1. Results of first Monte Carlo design

	Empirical coverage			
Procedure	75%	90%	95%	99%
CHT's subsampling ($b_n = n/2$)	93.1%***	99.6%***	100%***	100%***
CHT's subsampling ($b_n = n/3$)	87.2%***	98.8%***	99.9%***	99.9%***
Our subsampling ($b_n = n/2, \tau_n = \ln \ln n$)	57.2%***	74.5%***	82.3%***	91.9%***
Our subsampling ($b_n = n/3, \tau_n = \ln \ln n$)	63.5%***	80.8%***	89.3%***	96.1%***
Naive bootstrap	100%***	100%***	100%***	100%***
Our bootstrap ($\tau_n = \ln \ln n$)	73.6%	89.6%	94.8%	98.7%
Our asymptotic approximation ($\tau_n = \ln \ln n$)	78.4%**	91.9%**	96.1%	98.6%

Table A.2. Results of second Monte Carlo design

A.3.7.2. Design 2. Table A.2 compares the coverage probability for $\beta = (-0.5, -0.5)$. The performance of the different inferential methods is similar to the one in the first design.

A.3.7.3. Design 3. Table A.3 compares the coverage probability for $\beta = (0, 0)$. The performance of the different inferential methods is similar to the one in the first design, except that the asymptotic approximation seems to be affected by overcoverage.

A.3.7.4. Design 4. Table A.4 compares the coverage probability for $\beta = (0, 0)$. In this design, the identified set is empty or equivalently, the model is misspecified. This, in this case, covering the point $(0, 0)$ implies making a type II error. For this reason, we do not

	Empirical coverage			
Procedure	75%	90%	95%	99%
CHT's subsampling ($b_n = n/2$)	93.5%***	99.9%***	100%***	100%***
CHT's subsampling ($b_n = n/3$)	87.4%***	99.2%***	100%***	100%***
Our subsampling ($b_n = n/2, \tau_n = \ln \ln n$)	69.2%***	81.3%***	86.7%***	94.4%***
Our subsampling ($b_n = n/3, \tau_n = \ln \ln n$)	73.2%	85.4%***	91.8%***	97.1%***
Naive bootstrap	100%***	100%***	100%***	100%***
Our bootstrap ($\tau_n = \ln \ln n$)	79.3%***	91.5%	96.1%	99.2%
Our asymptotic approximation ($\tau_n = \ln \ln n$)	86.1%***	95.7%***	98.3%***	99.7%**

Table A.3. Results of third Monte Carlo design

	Empirical coverage			
Procedure	75%	90%	95%	99%
CHT's subsampling ($b_n = n/2$)	49.8%	91.5%	99.3%	100%
CHT's subsampling ($b_n = n/3$)	36.4%	79.0%	92.9%	99.8%
Our subsampling ($b_n = n/2, \tau_n = \ln \ln n$)	15.2%	26.0%	34.2%	51.4%
Our subsampling ($b_n = n/3, \tau_n = \ln \ln n$)	18.5%	32.1%	42.6%	61.0%
Naive bootstrap	100%	100%	100%	100%
Our bootstrap ($\tau_n = \ln \ln n$)	24.0%	43.4%	56.4%	75.3%
Our asymptotic approximation ($\tau_n = \ln \ln n$)	32.3%	55.0%	68.4%	81.7%

Table A.4. Results of fourth Monte Carlo design

test whether the empirical coverage coincides with the desired coverage level but, instead compare the methods in terms of which is providing lowest coverage.

In this case, our bootstrap and our asymptotic approximation lead to coverage sets that are smaller than the one generated by CHT [**23**]'s subsampling but larger than the one generated by the subsampling procedure of section 1.4.3.1. As expected, the naive bootstrap leads to coverage sets that are relatively too large.

APPENDIX B

Specification Test for Missing Functional Data

B.1. Notation

- Throughout this appendix, we abbreviate "cumulative distribution function" by CDF,
- For $x_1, x_2 \in \mathbb{R}^K$, we define $\|x_1 - x_2\| \equiv \sqrt{\sum_{k=1}^{K} (x_{1,k} - x_{2,k})^2}$ and for $G_1, G_2 \in L_2(\mathcal{D})$, we define $\|G_1 - G_2\|_\mu \equiv \int (G_1(x) - G_2(x))^2 \, d\mu(x)$.

B.2. Identification analysis for missing functional data

Proof of lemma 19. The construction of the bounds follows directly from derivations provided in the main text, so it only remains to be show that they are sharp. For any $x \in L_2(\mathcal{D})$ and a value $F(x)$ that satisfies the worst case scenario bounds, define the vector,

$$\{P(\{X(t) \leq x(t), \forall t \in U(j)\} | \pi = j, \{X(t) \leq x(t), \forall t \in O(j)\})\}_{j=2}^{2^K}$$

in any way such that the following equation is satisfied,

(B.1)

$$\begin{aligned} & F(x) - F_X(x|\pi = 1) P(\pi = 1) = \\ & = \sum_{j=2}^{2^K} \left\{ \begin{array}{c} P(\{X(t) \leq x(t), \forall t \in U(j)\} | \pi = j, \{X(t) \leq x(t), \forall t \in O(j)\}) * \\ * F_X(x, O(j) | \pi = j) P(\pi = j) \end{array} \right\} \end{aligned}$$

To provide a concrete example, consider the case when the vector is a vector of constants, that is, $\forall j = 2, 3, ..., 2^K$,

$$\begin{aligned} &P\left(\left\{X(t) \leq x(t), \forall t \in U(j)\right\} | \pi = j, \left\{X(t) \leq x(t), \forall t \in O(j)\right\}\right) = \\ &= \frac{F(x) - F_X(x|\pi = 1)\, P(\pi = 1)}{\sum_{s=2}^{2^K} F_X(x, O(s) | \pi = s)\, P(\pi = s)} \end{aligned}$$

Since $F(x)$ satisfies the worst case scenario bounds then, as long as equation (B.1) is satisfied, then $\forall j = 2, 3, ..., 2^K$,

$$P\left(\left\{X(t) \leq x(t), \forall t \in U(j)\right\} | \pi = j, \left\{X(t) \leq x(t), \forall t \in O(j)\right\}\right) \in [0, 1]$$

and hence it is a valid number for a probability. Hence, every value inside the worst case scenario bounds is feasible, completing the proof. $\square$

Proof of lemma 20. In order to prove that $\mathcal{H}(F_X) = \left\{\Gamma \cap \left\{G : F_X^L \leq G \leq F_X^H\right\}\right\}$, we need to show two statements: (1) $F_X \in \left\{\Gamma \cap \left\{G : F_X^L \leq G \leq F_X^H\right\}\right\}$ and (2) if $F \in \left\{\Gamma \cap \left\{G : F_X^L \leq G \leq F_X^H\right\}\right\}$, then F is a valid CDF for X.

Part 1. Since F_X is a CDF, $F_X \in \Gamma$ and by lemma 19, $F_X \in \left\{G : F_X^L \leq G \leq F_X^H\right\}$.

Part 2. Suppose that $F \in \left\{\Gamma \cap \left\{G : F_X^L \leq G \leq F_X^H\right\}\right\}$. Then, $\forall x \in L_2(\mathcal{D})$, we can define,

$$\begin{aligned} &P\left(\left\{X(t) \leq x(t), \forall t \in U(j)\right\} | \pi = j, \left\{X(t) \leq x(t), \forall t \in O(j)\right\}\right) = \\ &= \frac{F(x) - F_X(x|\pi = 1)\, P(\pi = 1)}{\sum_{s=2}^{2^K} F_X(x, O(s) | \pi = s)\, P(\pi = s)} \end{aligned}$$

for every $j = 2, 3, ..., 2^K$. Given that $F \in \left\{\Gamma \cap \left\{G : F_X^L \leq G \leq F_X^H\right\}\right\}$, it is not hard to verify that $\forall j = 2, 3, ..., 2^K$ $P\left(\left\{X(t) \leq x(t), \forall t \in U(j)\right\} | \pi = j, \left\{X(t) \leq x(t), \forall t \in O(j)\right\}\right)$

$\in \Gamma$. Hence, we have constructed the CDF for the missing data such that F is the CDF of X. $\square$

Proof of theorem 21. This proof is trivial. To show that the bounds are sharp, we need to clarify the definition of sharpness. Sharpness of the upper (lower) bound means that $\exists G \in \mathcal{H}(F_X)$ such that $\int (G(x) - F_Y(x|\theta_0))^2 \mu(dx)$ is arbitrary close to the upper (lower) bound (the bounds need not be exactly achieved). $\square$

Proof of theorem 22. Part 1. Lower bound. For every $x \in L_2(\mathcal{D})$, consider the problem,

$$\inf_{F_X^L(x) \leq f \leq F_X^H(x)} (f - F_Y(x|\theta_0))^2 \tag{B.2}$$

The unique solution to problem (B.2) is,

$$\begin{aligned} G_L^{**}(x) &= F_Y(x|\theta_0) 1\left[F_X^L(x) \leq F_Y(x|\theta_0) \leq F_X^H(x)\right] + \\ &\quad + F_X^L(x) 1\left[F_X^L(x) > F_Y(x|\theta_0)\right] + F_X^H(x) 1\left[F_Y(x|\theta_0) > F_X^H(x)\right] \end{aligned}$$

We can use this function to construct a function $G_L^* : L_2(\mathcal{D}) \to \mathbb{R}$ such that $\forall x \in L_2(\mathcal{D})$, $G_L^*(x) = G_L^{**}(x)$. By definition, $\forall x \in L_2(\mathcal{D})$, $F_X^L(x) \leq G_L^*(x) \leq F_X^H(x)$ or, equivalently, $G_L^* \in \mathcal{H}'(F_X)$ and therefore,

$$\begin{aligned} T_L'(X, Y_{\theta_0}) &= \int \inf_{F_X^L(x) \leq f \leq F_X^H(x)} (f - F_Y(x|\theta_0))^2 \mu(dx) \\ &= \int \left\{ \begin{array}{c} 1\left[F_Y(x|\theta_0) < F_X^L(x)\right] \left(F_X^L(x) - F_Y(x|\theta_0)\right)^2 + \\ +1\left[F_Y(x|\theta_0) > F_X^H(x)\right] \left(F_X^H(x) - F_Y(x|\theta_0)\right)^2 \end{array} \right\} \mu(dx) \end{aligned}$$

Part 2. Upper bound. For every $x \in L_2(\mathcal{D})$, consider the problem,

$$\text{(B.3)} \qquad \sup_{F_X^L(x) \leq f \leq F_X^H(x)} \left(f - F_Y(x|\theta_0)\right)^2$$

There are eight possible solutions for the problem (B.3), which we denote by $\left\{G_{H,j}^{**}(x)\right\}_{j=1}^{8}$. These solutions can be characterized by,

$$G_{H,j}^{**}(x) = \left\{ \begin{array}{c} F_X^L(x) \left\{ \begin{array}{c} 1 \left[\begin{array}{c} F_X^L(x) \lessdot_{(2.a)} F_Y(x|\theta_0) \lessdot_{(1.a)} F_X^H(x) \\ F_Y(x|\theta_0) - F_X^L(x) \gtrdot_{(3.a)} F_X^H(x) - F_Y(x|\theta_0) \end{array} \right] + \\ +1\left[F_Y(x|\theta_0) \gtrdot_{(1.b)} F_X^H(x)\right] \end{array} \right\} + \\ +F_X^H(x) \left\{ \begin{array}{c} 1 \left[\begin{array}{c} F_X^L(x) \lessdot_{(2.a)} F_Y(x|\theta_0) \lessdot_{(1.a)} F_X^H(x) \\ F_Y(x|\theta_0) - F_X^L(x) \lessdot_{(3.b)} F_X^H(x) - F_Y(x|\theta_0) \end{array} \right] + \\ +1\left[F_X^L(x) \gtrdot_{(2.b)} F_Y(x|\theta_0)\right] \end{array} \right\} \end{array} \right\}$$

where $\lessdot_{(1.a)}$, $\gtrdot_{(1.b)}$, $\lessdot_{(2.a)}$, $\gtrdot_{(2.b)}$, $\gtrdot_{(3.a)}$ and $\lessdot_{(3.b)}$ denote either strict or weak inequalities, giving rise to the eight solutions. These eight solutions are shown in the following table,

Solution	$\lessdot_{(1.a)}$	$\gtrdot_{(1.b)}$	$\lessdot_{(2.a)}$	$\gtrdot_{(2.b)}$	$\gtrdot_{(3.a)}$	$\lessdot_{(3.b)}$
$j = 1$	$\leq$	$>$	$\leq$	$>$	$\geq$	$<$
$j = 2$	$\leq$	$>$	$\leq$	$>$	$>$	$\leq$
$j = 3$	$\leq$	$>$	$<$	$\geq$	$\geq$	$<$
$j = 4$	$\leq$	$>$	$<$	$\geq$	$>$	$\leq$
$j = 5$	$<$	$\geq$	$\leq$	$>$	$\geq$	$<$
$j = 6$	$<$	$\geq$	$\leq$	$>$	$>$	$\leq$
$j = 7$	$<$	$\geq$	$<$	$\geq$	$\geq$	$<$
$j = 8$	$<$	$\geq$	$<$	$\geq$	$>$	$\leq$

We can use any of these solutions to construct a function $G_H^* : L_2(\mathcal{D}) \to \mathbb{R}$ such that $\forall x \in L_2(\mathcal{D})$, $\exists j \in \{1, 2, ..., 8\}$ such that $G_H^*(x) = G_{H,j}^{**}(x)$. By definition, $\forall x \in L_2(\mathcal{D})$, $\forall j \in \{1, 2, ..., 8\}$, $F_X^L(x) \leq G_{H,j}^{**}(x) \leq F_X^H(x)$ so, $G_H^* \in \mathcal{H}'(F_X)$ and therefore,

$$\begin{aligned} T_H'(X, Y_{\theta_0}) &= \int \sup_{F_X^L(x) \leq f \leq F_X^H(x)} (f - F_Y(x|\theta_0))^2 \mu(dx) \\ &= \int \max\left\{\left(F_X^L(x) - F_Y(x|\theta_0)\right)^2, \left(F_X^H(x) - F_Y(x|\theta_0)\right)^2\right\} \mu(dx) \end{aligned}$$

completing the proof. $\square$

Proof of theorem 23. Part 1. Lower bound.

Step 1. For every $x \in L_2(\mathcal{D})$, consider $G_L^*(x)$ defined according to the proof of theorem 22. We now show that $G_L^* \in \left\{\Gamma \cap \left\{G : F_X^L \leq G \leq F_X^H\right\}\right\}$. By construction,

$G_L^* \in \left\{G : F_X^L \leq G \leq F_X^H\right\}$, so we only need to verify that $G_L^* \in \Gamma$. We verify the properties one by one:

(i) *Monotonicity.* Let $x_1, x_2 \in L_2(\mathcal{D})$ and $x_1(t) \leq x_2(t)$ $\forall t \in \mathcal{D}$. Since $F_Y(x|\theta_0)$, $F_X^L(x)$ and $F_X^H(x)$ are weakly increasing in x, $F_Y(x_1|\theta_0) \leq F_Y(x_2|\theta_0)$, $F_X^L(x_1) \leq F_X^L(x_2)$ and $F_X^H(x_1) \leq F_X^H(x_2)$.

Case 1: Suppose that $F_X^L(x_1) > F_Y(x_1|\theta_0)$, so $G^*(x_1) = F_X^L(x_1)$. Case 1.1: If $F_X^L(x_2) > F_Y(x_2|\theta_0)$, then $G_L^*(x_2) = F_X^L(x_2) \geq F_X^L(x_1) = G_L^*(x_1)$. Case 1.2: If $F_X^L(x_2) \leq F_Y(x_2|\theta_0) \leq F_X^H(x_2)$, then $G_L^*(x_2) = F_Y(x_2|\theta_0) \geq F_X^L(x_2) \geq F_X^L(x_1) = G_L^*(x_1)$. Case 1.3: If $F_X^H(x_2) < F_Y(x_2|\theta_0)$, then $G_L^*(x_2) = F_X^H(x_2) \geq F_X^L(x_2) \geq F_X^L(x_1) = G_L^*(x_1)$.

Case 2: Suppose that $F_X^L(x_1) \leq F_Y(x_1|\theta_0) \leq F_X^H(x_1)$, so $G_L^*(x_1) = F_Y(x_1|\theta_0)$. Case 2.1: If $F_X^L(x_2) > F_Y(x_2|\theta_0)$, then $G_L^* = F_X^L(x_2) > F_Y(x_2|\theta_0) \geq F_Y(x_1|\theta_0) = G_L^*(x_1)$. Case 2.2: If $F_X^L(x_2) \leq F_Y(x_2|\theta_0) \leq F_X^H(x_2)$, then $G_L^*(x_2) = F_Y(x_2|\theta_0) \geq F_Y(x_1|\theta_0) = G_L^*(x_1)$. Case 2.3: If $F_X^H(x_2) < F_Y(x_2|\theta_0)$, then $G_L^*(x_2) = F_X^H(x_2) \geq F_X^H(x_1) \geq F_Y(x_1|\theta_0) = G_L^*(x_1)$.

Case 3: Suppose that $F_Y(x_1|\theta_0) > F_X^H(x_1)$, so $G_L^*(x_1) = F_X^H(x_1)$. Case 3.1: If $F_X^L(x_2) > F_Y(x_2|\theta_0)$, then $G_L^*(x_2) = F_X^L(x_2) > F_Y(x_2|\theta_0) \geq F_X^H(x_1) = G_L^*(x_1)$. Case 3.2: If $F_X^L(x_2) \leq F_Y(x_2|\theta_0) \leq F_X^H(x_2)$, then $G_L^*(x_2) = F_Y(x_2|\theta_0) \geq F_Y(x_1|\theta_0) > F_X^H(x_1) = G_L^*(x_1)$. Case 3.3: If $F_X^H(x_2) < F_Y(x_2|\theta_0)$, then $G_L^*(x_2) = F_X^H(x_2) \geq F_X^H(x_1) = G_L^*(x_1)$.

(ii) Right Continuity. Consider an arbitrary convergent sequence $\{x_n : n \in \mathbb{N}\}$ such that $\forall n \in \mathbb{N}$, $x_n \in L_2(\mathcal{D})$ and $x_n \geq x_{n+1}$. Denote $\lim_{n\to\infty} x_n = \bar{x}$. The objective is to show that $\lim_{n\to\infty} G_L^*(x_n) = G_L^*(\bar{x})$. Since $G_L^*(x)$ is a combination of $F_Y(x|\theta_0)$, $F_X^L(x)$

and $F_X^H(x)$, which are right continuous, the limit can only be discontinuous if, at the limit, the definition of the function G_L^* switches from one function to another one and the value of these two functions differ. There are three cases to consider.

Case 1: $\forall n \in \mathbb{N}$, $F_Y(x_n|\theta_0) < F_X^L(x_n)$ and so $G_L^*(x_n) = F_X^L(x_n)$. Taking limits and using right continuity, $F_Y(\bar{x}|\theta_0) \leq F_X^L(\bar{x})$. In this case, in the limit, the function G_L^* either does not switch its definition from function or, if it does, then the value of these two functions coincide. As a consequence, $\lim_{n\to\infty} G_L^*(x_n) = G_L^*(\bar{x})$.

Case 2: $\forall n \in \mathbb{N}$, $F_X^L(x_n) \leq F_Y(x_n|\theta_0) \leq F_X^H(x_n)$ and so $G_L^*(x_n) = F_Y(x_n|\theta_0)$. Taking limits and using right continuity, $F_X^L(\bar{x}) \leq F_Y(\bar{x}|\theta_0) \leq F_X^H(\bar{x})$. In this case, in the limit, the function G_L^* does not switch its definition and so $\lim_{n\to\infty} G_L^*(x_n) = F_Y(\bar{x}|\theta_0) = G_L^*(\bar{x})$.

Case 3: $\forall n \in \mathbb{N}$, $F_Y(x_n|\theta_0) > F_X^H(x_n)$ and so $G_L^*(x_n) = F_X^H(x)$. The rest of the argument is analogous to case 1.

(iii) $\lim_{x\to x_{-\infty}} G_L^*(x) = 0$. This follows from $\lim_{x\to x_{-\infty}} F_X^H(x) = 0$, $\lim_{x\to x_{-\infty}} F_X^L(x) = 0$, and $\lim_{x\to x_{-\infty}} F_Y(x|\theta_0) = 0$.

(iv) $\lim_{x\to x_{+\infty}} G_L^*(x) = 1$. Note that $\lim_{x\to x_{+\infty}} F_X^H(x) = 1$, $\lim_{x\to x_{+\infty}} F_Y(x|\theta_0) = 1$, but, in general, $\lim_{x\to x_{-\infty}} F_X^L(x) < 1$. Thus, only verify $\lim_{x\to x_{+\infty}} 1\left[F_X^L(x) > F_Y(x|\theta_0)\right] = 0$, which follows from $\lim_{x\to x_{-\infty}} F_X^L(x) < 1 = \lim_{x\to x_{+\infty}} F_Y(x|\theta_0)$.

Step 2. From the previous two steps,

$$\begin{aligned}
\left\|G_L^* - F_Y(\cdot|\theta_0)\right\|_\mu &= \int \min_{F_X^L(x)\leq f\leq F_X^H(x)} \left(f - F_Y(x|\theta_0)\right)^2 d\mu(x) \\
&\leq \inf_{F\in\left\{\Gamma\cap\left\{G:F_X^L\leq G\leq F_X^H\right\}\right\}} \int \left(F(x) - F_Y(x|\theta_0)\right)^2 d\mu(x) \\
&\leq \left\|G_L^* - F_Y(\cdot|\theta_0)\right\|_\mu
\end{aligned}$$

where the first equality holds by definition of G_L^* and last inequality follows from $G_L^* \in \left\{\Gamma \cap \left\{G : F_X^L \leq G \leq F_X^H\right\}\right\}$. As a consequence,

$$T_L\left(X, Y_{\theta_0}\right) = \left\|G_L^* - F_Y\left(\cdot|\theta_0\right)\right\|_\mu = \int \left(G_L^*\left(x\right) - F_Y\left(x|\theta_0\right)\right)^2 d\mu\left(x\right)$$

completing the part.

<u>Step 3.</u> For every $x \in L_2\left(\mathcal{D}\right)$, recall the functions $\left\{G_{H,j}^{**}\left(x\right)\right\}_{j=1}^{8}$ defined according to the proof of theorem 22. We now show that $\nexists G_H^* \in \Gamma$ such that $\forall x \in L_2\left(\mathcal{D}\right)$, $\exists j \in \left\{1, 2, ..., 8\right\}$ such that $G_H^*\left(x\right) = G_{H,j}^{**}\left(x\right)$. If this were the case, then $\lim_{x \rightarrow x_{+\infty}} G_H^*\left(x\right) = 1$. Since $\lim_{x \rightarrow x_{+\infty}} F_X^L\left(x\right) < 1$ and $\lim_{x \rightarrow x_{+\infty}} F_X^H\left(x\right) = 1$, $\lim_{x \rightarrow x_{+\infty}} G_H^*\left(x\right) = 1$ requires that,

$$\text{(B.4)} \qquad \lim_{x \rightarrow x_{+\infty}} \left\{ \begin{array}{c} 1 \left[\begin{array}{c} F_X^L\left(x\right) \lessdot_{(2.a)} F_Y\left(x|\theta_0\right) \lessdot_{(1.a)} F_X^H\left(x\right) \\ F_Y\left(x|\theta_0\right) - F_X^L\left(x\right) \lessdot_{(3.b)} F_X^H\left(x\right) - F_Y\left(x|\theta_0\right) \end{array} \right] + \\ +1\left[F_X^L\left(x\right) \gtrdot_{(2.b)} F_Y\left(x|\theta_0\right)\right] \end{array} \right\} = 1$$

Since $\lim_{x \rightarrow x_{+\infty}} \left\{F_Y\left(x|\theta_0\right) - F_X^L\left(x\right)\right\} > 0$ and $\lim_{x \rightarrow x_{+\infty}} \left\{F_X^H\left(x\right) - F_Y\left(x|\theta_0\right)\right\} = 0$, regardless of how $\lessdot_{(3.b)}$ is defined along the sequence $x \rightarrow x_{+\infty}$, $\lim_{x \rightarrow x_{+\infty}} G_H^*\left(x\right) = 1$ requires that,

$$\text{(B.5)} \qquad \lim_{x \rightarrow x_{+\infty}} 1 \left[\begin{array}{c} F_X^L\left(x\right) \lessdot_{(2.a)} F_Y\left(x|\theta_0\right) \lessdot_{(1.a)} F_X^H\left(x\right) \\ F_Y\left(x|\theta_0\right) - F_X^L\left(x\right) \lessdot_{(3.b)} F_X^H\left(x\right) - F_Y\left(x|\theta_0\right) \end{array} \right] = 0$$

Also, since $\lim_{x \rightarrow x_{+\infty}} F_Y\left(x|\theta_0\right) = 1 < \lim_{x \rightarrow x_{+\infty}} F_X^L\left(x\right) = P\left(\pi = 1\right)$, regardless of how $\gtrdot_{(2.b)}$ is defined along the sequence $x \rightarrow x_{+\infty}$,

$$\text{(B.6)} \qquad \lim_{x \rightarrow x_{+\infty}} 1\left[F_X^L\left(x\right) \gtrdot_{(2.b)} F_Y\left(x|\theta_0\right)\right] = 0$$

Hence, we obtain a contradiction between the requirements of equations (B.4), (B.5) and (B.6).

Part 2. Upper bound. We now show that the inequality in the following expression,

$$T_H\left(X, Y_{\theta_0}\right) \leq \int \max_{F_X^L(x) \leq f \leq F_X^H(x)} \left(f - F_Y\left(x|\theta_0\right)\right)^2 d\mu\left(x\right)$$

can be both an equality or a strict inequality, depending on $F_Y\left(\cdot|\theta_0\right), F_X^H$ and F_X^H.

Example 1. Let $F_Y\left(\cdot|\theta_0\right), F_X^H$ and F_X^H be such that $\forall x \in L_2\left(\mathcal{D}\right), F_Y\left(x|\theta_0\right) = F_X^H\left(x\right) > F_X^L\left(x\right)$ and $\int \left(F_X^L\left(x\right) - F_Y\left(x|\theta\right)\right)^2 d\mu\left(x\right) = M < +\infty$. In this case, no matter how $\gtrdot_{(2.b)}$ and $\lessdot_{(3.b)}$ are defined in the solution of problem (B.3), $1\left[F_X^L\left(x\right) \gtrdot_{(2.b)} F_Y\left(x|\theta_0\right)\right] = 0$ and $1\left[F_Y\left(x|\theta_0\right) - F_X^L\left(x\right) \lessdot_{(3.b)} F_X^H\left(x\right) - F_Y\left(x|\theta_0\right)\right] = 0$, which implies that: $\forall x \in L_2\left(\mathcal{D}\right)$ and $\forall j \in \{1, 2, ..., 8\}, G_{H,j}^{**}\left(x\right) = F_X^L\left(x\right)$, and so, $G_H^* = F_X^L$.

By the derivation in step 2, G_H^* does not satisfy with the properties of a CDF, so we cannot use the argument of the lower bound to claim our result. Instead, we will construct a sequence of functions that will be CDFs and will be arbitrarily close to G_H^*. Now consider the following alternative sequence of functions: $\{G_m^* : m \geq 1\}$ such that $\forall x \in L_2\left(\mathcal{D}\right)$,

$$G_m^*\left(x\right) = F_X^L\left(x\right) 1\left[x\left(t\right) \leq m : \forall t \in \mathcal{D}\right] + \left(1 - 1\left[x\left(t\right) \leq m : \forall t \in \mathcal{D}\right]\right)$$

It is easy to verify that $\forall m \in \mathbb{N}$, G_m^* satisfies the property of a CDF and moreover, $m \to +\infty$, $\left\|G_m^* - G_H^*\right\|_\mu = \int \left(1 - 1\left[x\left(t\right) \leq m : \forall t \in \mathcal{D}\right]\right) d\mu\left(x\right) = o\left(1\right)$.

Fix $\varepsilon > 0$ and set η so that $\max\left\{\eta + 2\eta^{1/2} M, \eta + 2\eta^{1/2}\left(\eta + 2\eta^{1/2} M + M\right)^{1/2}\right\} \leq \varepsilon$. Since $\left\|G_m^* - G_H^*\right\|_\mu = o\left(1\right), \exists N\left(\varepsilon\right) \in \mathbb{N}$ such that $\forall m \geq N\left(\varepsilon\right), \left\|G_m^* - G_H^*\right\|_\mu \leq \eta$. As a consequence, $\forall m \geq N(\varepsilon)$,

$$\begin{aligned}
\|G_H^* - F_Y(\cdot|\theta_0)\|_\mu &= \int \left(G_H^*(x) - G_m^*(x) + G_m^*(x) - F_Y(x|\theta_0)\right)^2 d\mu(x) \\
&\leq \left\{ \begin{array}{c} \|G_m^* - G_H^*\|_\mu + \|G_m^* - F_Y(\cdot|\theta_0)\|_\mu + \\ +2\int \left(G_m^*(x) - G_H^*(x)\right)\left(G_m^*(x) - F_Y(x|\theta_0)\right) d\mu(x) \end{array} \right\} \\
&\leq \|G_m^* - F_Y(\cdot|\theta_0)\|_\mu + \eta + 2\eta^{1/2}\left(\eta + 2\eta^{1/2}M + M\right)^{1/2} \\
&\leq \|G_m^* - F_Y(\cdot|\theta_0)\|_\mu + \varepsilon
\end{aligned}$$

As a consequence, $\forall m \geq N(\varepsilon)$,

$$\begin{aligned}
\|G_H^* - F_Y(\cdot|\theta_0)\|_\mu &= \int \max_{F_X^L(x) \leq f \leq F_X^H(x)} \left(f - F_Y(x|\theta_0)\right)^2 d\mu(x) \\
&\geq \sup_{F \in \left\{\Gamma \cap \left\{G : F_X^L \leq G \leq F_X^H\right\}\right\}} \int \left(F(x) - F_Y(x|\theta_0)\right)^2 d\mu(x) \\
&\geq \|G_m^* - F_Y(\cdot|\theta_0)\|_\mu \\
&\geq \|G_H^* - F_Y(\cdot|\theta_0)\|_\mu - \varepsilon
\end{aligned}$$

where the third inequality follows from the fact that $G_m^* \in \left\{\Gamma \cap \left\{G : F_X^L \leq G \leq F_X^H\right\}\right\}$. As a consequence of the previous chain of inequalities, we deduce that,

$$T_H(X, Y_{\theta_0}) = \int \max_{F_X^L \leq f \leq F_X^H} \left(f(x) - F_Y(x|\theta_0)\right)^2 d\mu(x)$$

completing the example.

Example 2. Let $F_X^H > F_X^L$ and let $F_Y(\cdot|\theta_0)$ be defined as follows: $\forall x \in L_2(\mathcal{D})$,

$$F_Y(x|\theta_0) = F_X^L(x) 1\left[F_X^H(x) < 0.5\right] + F_X^H(x) 1\left[F_X^H(x) \geq 0.5\right]$$

Let $\bar{x}_1 \in L_2(\mathcal{D})$ such that $F_X^H(\bar{x}_1) = 0.5 > F_X^L(\bar{x}_1)$ and let $\bar{x}_2 \in L_2(\mathcal{D})$ such that $\forall t \in \mathcal{D}$, $\bar{x}_2(t) < \bar{x}_1(t)$ and $F_X^L(\bar{x}_2) < F_X^L(\bar{x}_1) < F_X^H(\bar{x}_2) < 0.5$. Notice that $F_X^H(x) - F_X^L(x)$ is defined as follows,

$$F_X^H(x) - F_X^L(x) = \sum_{j=2}^{2^K} P\left(X(t) \le x(t), \forall t \in O(j) \,|\pi = j\right) P(\pi = j)$$

By this definition, $F_X^H(\bar{x}_1) - F_X^L(\bar{x}_1) \ge F_X^H(\bar{x}_2) - F_X^L(\bar{x}_2)$ and we will further assume that $F_X^H(\bar{x}_2) - F_X^L(\bar{x}_2) \ge \eta > 0$ for some $\eta > 0$.

If $x \in L_2(\mathcal{D})$ is such that $F_X^H(x) < 0.5$, then $F_Y(x|\theta_0) = F_X^L(x) < F_X^H(x)$ and so, $\forall j \in \{1, 2, ..., 8\}$ $G_{H,j}^{**}(x) = F_X^H(x)$. Similarly, if $x \in L_2(\mathcal{D})$ is such that $F_X^H(x) \ge 0.5$, then $F_X^L(x) < F_Y(x|\theta_0) = F_X^H(x)$ and so, $\forall j \in \{1, 2, ..., 8\}$, $G_{H,j}^{**}(x) = F_X^L(x)$. In other words, the solution to problem (B.3) is unique and, therefore, G_H^* is given by the following expression: $\forall x \in L_2(\mathcal{D})$

$$G_H^*(x) = F_X^H(x)\, 1\left[F_X^H(x) < 0.5\right] + F_X^L(x)\, 1\left[F_X^H(x) \ge 0.5\right]$$

and as a result,

$$\int \max_{F_X^L \le f \le F_X^H} \left(f(x) - F_Y(x|\theta_0)\right)^2 d\mu(x) = \int \left(F_X^H(x) - F_X^L(x)\right)^2 d\mu(x)$$

Notice that G_H^* is not weakly increasing and so $G_H^* \notin \Gamma$.

The next step is to show that: $T_H(X, Y_{\theta_0}) < \int \max_{F_X^L \le f \le F_X^H} \left(f - F_Y(x|\theta_0)\right)^2 d\mu(x)$. Suppose not. Therefore $\forall \varepsilon > 0$, $\exists F \in \left\{\Gamma \cap \left\{G : F_X^L \le G \le F_X^H\right\}\right\}$ such that,

$$\int \left(F_X^H(x) - F_X^L(x)\right)^2 d\mu(x) - \int \left(F(x) - F_Y(x|\theta_0)\right)^2 d\mu(x) < \varepsilon$$

Hence, it suffices to show that $\forall F \in \left\{\Gamma \cap \left\{G : F_X^L \leq G \leq F_X^H\right\}\right\}$ there is a positive lower bound for $\int \left(F_X^H(x) - F_X^L(x)\right)^2 d\mu(x) - \int \left(F(x) - F_Y(x|\theta_0)\right)^2 d\mu(x)$. Consider the following derivation,

$$\int \left(F_X^H(x) - F_X^L(x)\right)^2 d\mu(x) - \int \left(F(x) - F_Y(x|\theta_0)\right)^2 d\mu(x) =$$

$$= \left\{ \begin{array}{c} \int_{\left\{F_X^H < 0.5\right\}} \left(F_X^H(x) - F_X^L(x)\right)^2 - \left(F(x) - F_X^L(x)\right)^2 d\mu(x) + \\ + \int_{\left\{F_X^H \geq 0.5\right\}} \left(F_X^H(x) - F_X^L(x)\right)^2 - \left(F(x) - F_X^H(x)\right)^2 d\mu(x) \end{array} \right\} =$$

$$= \left\{ \begin{array}{c} \int \left(F_X^H(x) - F(x)\right)\left(F(x) - F_X^L(x)\right) d\mu(x) + \\ + \int_{\left\{F_X^H < 0.5\right\}} \left(F_X^H(x) - F(x)\right)\left(F_X^H(x) - F_X^L(x)\right) d\mu(x) + \\ + \int_{\left\{F_X^H \geq 0.5\right\}} \left(F_X^H(x) - F_X^L(x)\right)\left(F(x) - F_X^L(x)\right) d\mu(x) \end{array} \right\}$$

Since $F \in \left\{G : F_X^L \leq G \leq F_X^H\right\}$, the right hand side is a sum of three non-negative terms. Therefore,

$$\int \left(F_X^H(x) - F_X^L(x)\right)^2 d\mu(x) - \int \left(F(x) - F_Y(x|\theta_0)\right)^2 d\mu(x) \geq$$

$$\geq \eta \max \left\{ \begin{array}{c} \int_{\left\{F_X^H < 0.5 \cap \left\{F_X^H - F_X^L > \eta\right\}\right\}} \left(F(x) - F_X^L(x)\right) d\mu(x), \\ \int_{\left\{F_X^H < 0.5 \cap \left\{F_X^H - F_X^L > \eta\right\}\right\}} \left(F_X^H(x) - F(x)\right) d\mu(x) \end{array} \right\}$$

We divide the rest of the analysis into two cases. Case 1: $F(\bar{x}_1) > F_X^L(\bar{x}_1)$. Let $F(\bar{x}_1) - F_X^L(\bar{x}_1) = \delta > 0$. By right continuity of F and F_X^L, there is a set of functions sufficiently close to the function $\bar{x}_1$ (in the $L_2(\mathcal{D})$ metric) such that for every function of this set $F - F_X^L \geq \delta/2 > 0$. This set of functions could be constructed from functions such that $\forall t \in D$, $y(t) \geq \bar{x}_1(t)$ that are sufficiently close to $\bar{x}_1$. By our definition of the measure μ, this set can be constructed so that it has positive μ measure. This construction provides a positive lower bound for the expression for the first term of the maximum on the right hand side, which would provide a positive lower bound for the left hand side. Case 2: $F(\bar{x}_1) = F_X^L(\bar{x}_1)$. Since $F(\bar{x}_2) \leq F(\bar{x}_1)$ and we assumed that $F_X^L(\bar{x}_1) < F_X^H(\bar{x}_2)$, then $F(\bar{x}_2) < F_X^H(\bar{x}_2)$. Let $F_X^H(\bar{x}_2) - F(\bar{x}_2) = \delta > 0$. In the same way as in case 1, by right continuity of F and F_X^H, there is a set of functions with positive μ measure that are sufficiently close to the function $\bar{x}_2$ (in the $L_2(\mathcal{D})$ metric) such that for every function in this set $F_X^H - F \geq \delta/2 > 0$. This provides a positive lower bound for the expression for the second term of the maximum on the right hand side, which would provide a positive lower bound for the left hand side.

This completes the analysis of the example and the proof. □

Proof of lemma 24. This proof follows same arguments as the proof of lemma 20. □

Proof of theorem 25. This proof follows same arguments as the proof of theorem 21. □

B.3. Specification test for missing functional data

Proof of lemma 28. Consider the optimization problem to compute the estimate of the sharp worst case scenario lower bound. The value of the objective function depends on the function G only thought V values: $\{G(Z_j)\}_{j=1}^V$. Hence, we need to check that (1) if $G \in \hat{\mathcal{H}}(F_X)$, then $\{G(Z_j)\}_{j=1}^V \in \hat{\mathcal{S}}$ and (2) If $g \in \hat{\mathcal{S}}$, then $\exists G \in \hat{\mathcal{H}}(F_X)$ such that $\forall j \in \{1, 2, ...V\}, G(Z_j) = g_j$.

Part 1. Since $G \in \left\{G : \hat{F}_X^L \leq G \leq \hat{F}_X^H\right\}$, then $\forall j \in \{1, 2, ...V\}$, $\hat{F}_X^L(Z_j) \leq G(Z_j) \leq \hat{F}_X^H(Z_j)$. Moreover, since $G \in \Gamma$, G is monotonic and hence, $\forall j, k \in \{1, 2, ...V\}$, $Z_j \leq Z_k$ implies $g_j \leq g_k$.

Part 2. If $g \in \hat{\mathcal{S}}$, then $\exists G \in \hat{\mathcal{H}}(F_X)$ such that $\forall j \in \{1, 2, ...V\}$, $G(Z_j) = g_j$. For any $\varepsilon > 0$, define $Z_{V+1} = (1 + \varepsilon)\max_{j \in \{1,2,...V\}}\{Z_j\}$ and define $g_{V+1} = 1$. It is evident that $Z_{V+1} \in L_2(\mathcal{D})$. Then, consider the following function $G : L_2(\mathcal{D}) \to \mathbb{R}$, $G(Z) = \max_{j \in \{1,2,...V+1\}}\{g_j 1[Z_j \leq Z]\}$. We claim that this is a CDF that satisfies with our requirements. We check these requirements one by one.

(a) $\forall j \in \{1, 2, ...V\}$, $G(Z_j) = g_j$. Suppose this is not true for some $j \in \{1, 2, ...V\}$. Then, $\exists h \in \{1, 2, ...V\}$ such that $Z_j \geq Z_h$ and $g_j < g_h$, violating monotonicity.

(b) Monotonicity. For $Z_1, Z_2 \in L_2(\mathcal{D})$, if $Z_1 \leq Z_2$, then $G(Z_1) \leq G(Z_2)$. If $Z_1 \leq Z_2$, then $\forall j \in \{1, 2, ...V\}$, $g_j 1[Z_j \leq Z_1] \leq g_j 1[Z_j \leq Z_2]$ and therefore, $G(Z_1) \leq G(Z_2)$.

(c) Right continuity. Consider an arbitrary convergent sequence $\{x_n : n \in \mathbb{N}\}$ such that $\forall n \in \mathbb{N}$, $x_n \in L_2(\mathcal{D})$ and $x_n \geq x_{n+1}$. Denote $\lim_{n\to\infty} x_n = \bar{x}$. This result follows from the fact that $1[Z_j \leq x_n] \to 1[Z_j \leq x]$.

(d) $\lim_{x \to x_{-\infty}} G(x) = 0$. If $x \to x_{-\infty}$, then, eventually, $\forall j \in \{1, 2, ...V\}$, $x < Z_j$ and so, $G(x) = 0$.

(e) $\lim_{x \to x_{+\infty}} G(x) = 1$. If $x \to x_{+\infty}$, then, eventually, $\forall j \in \{1, 2, ...V\}$, $x > Z_j$. By definition $g_{V+1} = \max_{j \in \{1,2,...V+1\}}\{g_j\} = 1$. □

Proof of lemma 29. This proof follows same arguments as the proof of theorem 22. □

Proof of theorem 30. The proof is exactly the same as the proof of theorem 23. □

Proof of corollary 31. This proof follows from the result in theorem 30**.** Since $\hat{T}_L'\left(X,Y_{\hat{\theta}_0}\right) = \hat{T}_L\left(X,Y_{\hat{\theta}_0}\right),$ then when $t_{\hat{\theta}_0}(1-\alpha) < n\hat{T}_L\left(X,Y_{\hat{\theta}_0}\right) \leq n\hat{T}_H\left(X,Y_{\hat{\theta}_0}\right)$ it is also the case that $t_{\hat{\theta}_0}(1-\alpha) < n\hat{T}_L'\left(X,Y_{\hat{\theta}_0}\right) \leq n\hat{T}_H'\left(X,Y_{\hat{\theta}_0}\right).$ On the other hand, if $n\hat{T}_L\left(X,Y_{\hat{\theta}_0}\right) \leq n\hat{T}_H\left(X,Y_{\hat{\theta}_0}\right) \leq t_{\hat{\theta}_0}(1-\alpha)$ then it is still possible that $n\hat{T}_L'\left(X,Y_{\hat{\theta}_0}\right) \leq t_{\hat{\theta}_0}(1-\alpha) < n\hat{T}_H'\left(X,Y_{\hat{\theta}_0}\right)$ although it is impossible that $t_{\hat{\theta}_0}'(1-\alpha) < n\hat{T}_L'\left(X,Y_{\hat{\theta}_0}\right) \leq n\hat{T}_H'\left(X,Y_{\hat{\theta}_0}\right).$ □

Proof of theorem 32. From theorem 2 in BHHN [**19**] we know that,

$$\lim_{n\to\infty} P\left(n\hat{T}\left(X,Y_{\hat{\theta}_0}\right) > t_{\hat{\theta}_0}(1-\alpha)\right) = 1-\alpha$$

where $\hat{T}\left(X,Y_{\hat{\theta}_0}\right)$ represents the (unknown) test statistic that we would compute if we were to observe the complete dataset. By definition, $\hat{T}_L\left(X,Y_{\hat{\theta}_0}\right) \leq \hat{T}\left(X,Y_{\hat{\theta}_0}\right)$ and hence, for every n,

$$P\left(n\hat{T}_L\left(X,Y_{\hat{\theta}_0}\right) > t_{\hat{\theta}_0}(1-\alpha)\right) \leq P\left(n\hat{T}\left(X,Y_{\hat{\theta}_0}\right) > t_{\hat{\theta}_0}(1-\alpha)\right)$$

computing the limit infimum on both sides, the result follows. □

Proof of lemma 33. In this proof we will focus of the lower bounds, that is, we will show that: $\left|\hat{T}_L\left(X,Y_{\hat{\theta}_0}\right) - T_L\left(X,Y_{\theta_0}\right)\right| = o_p(1)$, but analogous arguments can be used to show that the analogous result for the upper bounds, that is, $\left|\hat{T}_H\left(X,Y_{\hat{\theta}_0}\right) - T_H\left(X,Y_{\theta_0}\right)\right| = o_p(1)$.

Step 1. Let Φ be the space of pairs of functions (F_1,F_2) defined by the following properties: (i) $F_1 : L_2(\mathcal{D}) \to \mathbb{R}$, $F_2 : L_2(\mathcal{D}) \to \mathbb{R}$ and (ii) $\forall x \in L_2(\mathcal{D}),\ 0 \leq F_1(x) \leq F_2(x) \leq 1$, (iii) F_1 and F_2 are weakly increasing, (iv) $\lim_{x\to\to-\infty} F_1(x) = 0$ and $\lim_{x\to+\infty} F_2(x) = 1$.

Define the functions $\Psi_H : \Phi \to \mathbb{R}_+$, $\check{\Psi}_H : \Phi \to \mathbb{R}_+$, $\tilde{\Psi}_H : \Phi \to \mathbb{R}_+$ and $\hat{\Psi}_H : \Phi \to \mathbb{R}_+$ as follows,

$$\begin{aligned}
\Psi_H(F_1, F_2) &= \sup_{F \in \{\Gamma \cap \{G : F_1 \leq G \leq F_2\}\}} \int (F(x) - F_Y(x|\theta_0))^2 \, d\mu(x) \\
\check{\Psi}_H(F_1, F_2) &= \sup_{F \in \{\Gamma \cap \{G : F_1 \leq G \leq F_2\}\}} \int \left(F(x) - F_Y\left(x|\hat{\theta}_0\right)\right)^2 d\mu(x) \\
\tilde{\Psi}_H(F_1, F_2) &= \sup_{F \in \{\Gamma \cap \{G : F_1 \leq G \leq F_2\}\}} \int \left(F(x) - \hat{F}_Y\left(x|\hat{\theta}_0\right)\right)^2 d\mu(x) \\
\hat{\Psi}_H(F_1, F_2) &= \sup_{F \in \{\Gamma \cap \{G : F_1 \leq G \leq F_2\}\}} \tfrac{1}{V} \sum_{j=1}^{V} \left(F(Z_j) - \hat{F}_Y\left(Z_j|\hat{\theta}_0\right)\right)^2
\end{aligned}$$

where $\{Z_j : j = 1, 2, ..., V\}$ is a random sample distributed according to μ. By definition: $T_H(X, Y_{\theta_0}) = \Psi_H\left(F_X^L, F_X^H\right)$ and $\hat{T}_H\left(X, Y_{\hat{\theta}_0}\right) = \hat{\Psi}_H\left(\hat{F}_X^L, \hat{F}_X^H\right)$. Therefore,

$$\left|\hat{T}_H\left(X, Y_{\hat{\theta}_0}\right) - T_H(X, Y_{\theta_0})\right| \leq \left\{ \begin{array}{c} \left|\hat{\Psi}_H\left(\hat{F}_X^L, \hat{F}_X^H\right) - \tilde{\Psi}_H\left(\hat{F}_X^L, \hat{F}_X^H\right)\right| + \\ + \left|\tilde{\Psi}_H\left(\hat{F}_X^L, \hat{F}_X^H\right) - \check{\Psi}_H\left(\hat{F}_X^L, \hat{F}_X^H\right)\right| + \\ + \left|\check{\Psi}_H\left(\hat{F}_X^L, \hat{F}_X^H\right) - \Psi_H\left(\hat{F}_X^L, \hat{F}_X^H\right)\right| + \\ + \left|\Psi_H\left(\hat{F}_X^L, \hat{F}_X^H\right) - \Psi_H\left(F_X^L, F_X^H\right)\right| \end{array} \right\}$$

In the remaining steps, we show that the right hand side of both expressions is $o_p(1)$.

Step 2. In this step, we show that $\left|\Psi_H\left(\hat{F}_X^L, \hat{F}_X^H\right) - \Psi_H\left(F_X^L, F_X^H\right)\right| = o_p(1)$.

Step 2.1. Consider any $(F_1, F_2) \in \Phi$ such that $\{\Gamma \cap \{G : F_1 \leq G \leq F_2\}\}$ is non-empty and $\Psi_H(F_1, F_2) < +\infty$. We want to show that for any sequence of $\{(F_{1,n}, F_{2,n}) : n \in \mathbb{N}\}$ such that,

$$\|F_{1,n} - F_1\|_\mu + \|F_{2,n} - F_2\|_\mu = o(1)$$

then: $\Psi_L(F_{1,n}, F_{2,n}) - \Psi_L(F_1, F_2) = o(1)$ and $\Psi_H(F_{1,n}, F_{2,n}) - \Psi_H(F_1, F_2) = o(1)$.

Fix $\varepsilon > 0$ and set $\varepsilon_0 \in (0, \varepsilon)$. Since $\Psi_H(F_1, F_2) < +\infty$, pick $\varepsilon_1 > 0$ such that,

$$\varepsilon_1 + 2(\varepsilon_1)^{1/2}\left(\varepsilon_0 + \varepsilon_1 + 2(\varepsilon_1)^{1/2}(\Psi_H(F_1, F_2))^{1/2} + \Psi_H(F_1, F_2)\right)^{1/2} = \varepsilon - \varepsilon_0$$

By definition of supremum, $\exists G_n^* \in \{\Gamma \cap \{F_{1,n} \leq G \leq F_{2,n}\}\}$ such that,

$$\Psi_H(F_{1,n}, F_{2,n}) - \varepsilon_0 \leq \|G_n^* - F_Y(\cdot|\theta_0)\|_\mu$$

Moreover, we can find $G_n' \in \{\Gamma \cap \{G : F_1 \leq G \leq F_2\}\}$ such that, $\exists N(\varepsilon_1) \in \mathbb{N}$, such that $\forall n \geq N$,

$$\|G_n' - G_n^*\|_\mu = \int (G_n'(x) - G_n^*(x))^2 \, d\mu(x) \leq \varepsilon_1$$

To show this, define, $\forall n \in \mathbb{N}$,

$$G_n'(x) = \left\{ \begin{array}{c} G_n^*(x) 1[F_1(x) \leq G_n^*(x) \leq F_2(x)] + \\ +F_1(x) 1[G^*(x) > F_1(x)] + F_2(x) 1[F_2(x) > G^*(x)] \end{array} \right\}$$

and so, $\forall n \geq N(\varepsilon_1)$,

$$\begin{aligned} \|G_n' - G_n^*\|_\mu &= \int \left(\begin{array}{c} (F_2(x) - G_n^*(x)) 1[G_n^*(x) > F_2(x)] + \\ +(F_1(x) - G_n^*(x)) 1[F_1(x) > G_n^*(x)] \end{array} \right)^2 d\mu(x) \\ &\leq \|F_{n,2} - F_2\|_\mu + \|F_{n,1} - F_1\|_\mu \leq \varepsilon_1 \end{aligned}$$

The part is completed by showing that $\forall n \in \mathbb{N}$, $G_n' \in \{\Gamma \cap \{G : F_1 \leq G \leq F_2\}\}$. It is clear that $G_n' \in \{G : F_1 \leq G \leq F_2\}$ by construction. Since $G_n^* \in \Gamma$, $F_{1,n} \leq F_{2,n}$, $\lim_{x \to -\infty} F_{1,n} = 0$, $\lim_{x \to +\infty} F_{2,n} = 1$ and $F_{1,n}$ and $F_{2,n}$ are weakly increasing and right continuous, it follows that $G_n' \in \Gamma$.

As a consequence,

$$\begin{aligned}\Psi_H\left(F_{1,n},F_{2,n}\right)-\varepsilon_0 &\leq \int\left(G_n^*\left(x\right)-F_Y\left(x|\theta_0\right)\right)^2 d\mu\left(x\right)\\ &= \int\left(G_n^*\left(x\right)-G_n'\left(x\right)+G_n'\left(x\right)-F_Y\left(x|\theta_0\right)\right)^2 d\mu\left(x\right)\\ &\leq \varepsilon_1+\Psi_H\left(F_1,F_2\right)+2\left(\varepsilon_1\right)^{1/2}\left(\Psi_H\left(F_1,F_2\right)\right)^{1/2}\end{aligned}$$

And so, by definition of ε_1, $\forall n\geq N\left(\varepsilon_1\right),\ \Psi_H\left(F_{1,n},F_{2,n}\right)\leq\varepsilon+\Psi_H\left(F_1,F_2\right).$

By reversing the roles, and repeating the argument, we deduce that,

$$\begin{aligned}&\Psi_H\left(F_1,F_2\right)-\varepsilon_0\leq\\ &\leq\left\{\begin{array}{c}\varepsilon_1+\Psi_H\left(F_{1,n},F_{2,n}\right)+\\ +2\left(\varepsilon_1\right)^{1/2}\left(\varepsilon_0+\varepsilon_1+2\left(\varepsilon_1\right)^{1/2}\left(\Psi_H\left(F_1,F_2\right)\right)^{1/2}+\Psi_H\left(F_1,F_2\right)\right)^{1/2}\end{array}\right\}\end{aligned}$$

And so, by our definition of $\varepsilon_1, \forall n\geq N\left(\varepsilon_1\right)$, then $\Psi_H\left(F_1,F_2\right)\leq\varepsilon+\Psi_H\left(F_{1,n},F_{2,n}\right)$. Finally, define $N\left(\varepsilon\right)=N\left(\varepsilon_1\right)$. For arbitrary $\varepsilon>0$, $\exists N\left(\varepsilon\right)$ such that $\forall n\geq N\left(\varepsilon\right)$, $\left|\Psi_H\left(F_{1,n},F_{2,n}\right)-\Psi_H\left(F_1,F_2\right)\right|\leq\varepsilon$.

Step 2.2. Let $\Phi_0\subset\Phi$, be such that $\forall\left(F_1,F_2\right)\in\Phi_0$, $\Psi\left(F_1,F_2\right)<+\infty$. By the law of large numbers, $\left(\hat{F}_X^L,\hat{F}_X^H\right)\xrightarrow{p}\left(F_X^L,F_X^H\right)$ and so, by Slutzky's Lemma, $\left\|\hat{F}_X^L-F_X^L\right\|_\mu+\left\|\hat{F}_X^H-F_X^H\right\|_\mu=o_p\left(1\right)$. As a consequence of step 1.1. and arguments similar to those used in the proof of Slutzky's Lemma: $\Psi_H\left(\hat{F}_X^L,\hat{F}_X^H\right)=\Psi_H\left(F_X^L,F_X^H\right)+o_p\left(1\right)$, completing this step.

Step 3. In this step, we show that $\left|\check{\Psi}_H\left(\hat{F}_X^L,\hat{F}_X^H\right)-\Psi_H\left(\hat{F}_X^L,\hat{F}_X^H\right)\right|=o_p\left(1\right)$.

Step 3.1. For any $(\theta, x) \in \{\Theta, L_2(\mathcal{D})\}$ denote $\dot{F}_Y(x|\theta) = \partial F_Y(x|\theta)/\partial\theta$. A Taylor series expansion gives,

$$\hat{F}_Y\left(x|\hat{\theta}_0\right) - F_Y(x|\theta_0) = \dot{F}_Y(x|\theta_0)\left(\hat{\theta}_0 - \theta_0\right) + \left(\dot{F}_Y\left(x|\tilde{\theta}\right) - \dot{F}_Y(x|\theta_0)\right)\left(\hat{\theta}_0 - \theta_0\right)$$

where $\tilde{\theta}$ is between θ_0 and $\hat{\theta}_0$. Consider the following derivation,

$$\int\left(F_Y\left(x|\hat{\theta}_0\right) - F_Y(x|\theta_0)\right)^2 d\mu(x)$$

$$\leq \left\{\begin{array}{c} \left\|\left(\hat{\theta}_0 - \theta_0\right)\right\|^2 \int \dot{F}_Y(x|\theta_0)' \dot{F}_Y(x|\theta_0)\, d\mu(x) + \\ + \left\|\left(\hat{\theta}_0 - \theta_0\right)\right\|^2 \int \left\{\begin{array}{c} \sum_{i,j=1}^{p} \left|\dot{F}_{Y,i}\left(x|\tilde{\theta}\right) - \dot{F}_{Y,i}(x|\theta_0)\right| * \\ * \left|\dot{F}_{Y,j}\left(x|\tilde{\theta}\right) - \dot{F}_{Y,j}(x|\theta_0)\right| \end{array}\right\} d\mu(x) + \\ + \left\{\begin{array}{c} \left\|\left(\hat{\theta}_0 - \theta_0\right)\right\|^2 * \sqrt{\int \dot{F}_Y(x|\theta_0)' \dot{F}_Y(x|\theta_0)\, d\mu(x)} * \\ 2\sqrt{\int \sum_{i,j=1}^{p} \sup\limits_{\|\theta - \theta_0\| < \varepsilon} \left\{\begin{array}{c} \left|\dot{F}_{Y,i}\left(x|\tilde{\theta}\right) - \dot{F}_{Y,i}(x|\theta_0)\right| * \\ * \left|\dot{F}_{Y,j}\left(x|\tilde{\theta}\right) - \dot{F}_{Y,j}(x|\theta_0)\right| \end{array}\right\} d\mu(x)} \end{array}\right\} + o_p(1) \end{array}\right\}$$

and by assumptions 2 and 3, the right hand side is $O_p(n^{-1})$.

Step 3.2. Consider the following derivation,

$$\int\left(F(x) - F_Y\left(x|\hat{\theta}_0\right)\right)^2 d\mu(x) =$$

$$= \int\left(F(x) - F_Y(x|\theta_0) + F_Y(x|\theta_0) - F_Y\left(x|\hat{\theta}_0\right)\right)^2 d\mu(x)$$

$$\leq \left\{\begin{array}{c} \int (F(x) - F_Y(x|\theta_0))^2 d\mu(x) + \int\left(F_Y\left(x|\hat{\theta}_0\right) - F_Y(x|\theta_0)\right)^2 d\mu(x) + \\ + \sqrt{\int (F(x) - F_Y(x|\theta_0))^2 d\mu(x)} \sqrt{\int\left(F_Y\left(x|\hat{\theta}_0\right) - F_Y(x|\theta_0)\right)^2 d\mu(x)} \end{array}\right\}$$

Taking supremum with respect to $F \in \left\{ \Gamma \cap \left\{ G : F_X^L \leq G \leq F_X^H \right\} \right\}$ on both sides, using steps 2 and the fact that $T_H\left(X, Y_{\theta_0}\right) < \infty$, it follows that,

$$\begin{aligned} &\check{\Psi}_H\left(\hat{F}_X^L, \hat{F}_X^H\right) - \Psi_H\left(\hat{F}_X^L, \hat{F}_X^H\right) \leq \\ &\leq \left\{ \begin{array}{c} \int \left(F_Y\left(x|\hat{\theta}_0\right) - F_Y\left(x|\theta_0\right)\right)^2 d\mu\left(x\right) + \\ + \sqrt{T_H\left(X, Y_{\theta_0}\right) + o_p\left(1\right)} \sqrt{\int \left(F_Y\left(x|\hat{\theta}_0\right) - F_Y\left(x|\theta_0\right)\right)^2 d\mu\left(x\right)} \end{array} \right\} \end{aligned}$$

and the right hand side is $o_p\left(1\right)$ by step 2.1.

Reversing the roles and repeating the argument it follows that,

$$\begin{aligned} &\Psi_H\left(\hat{F}_X^L, \hat{F}_X^H\right) \leq \\ &\leq \left\{ \begin{array}{c} \check{\Psi}_H\left(\hat{F}_X^L, \hat{F}_X^H\right) + \int \left(F_Y\left(x|\hat{\theta}_0\right) - F_Y\left(x|\theta_0\right)\right)^2 d\mu\left(x\right) + \\ + \sqrt{\Psi_H\left(F_X^L, F_X^H\right) + o_p\left(1\right)} \sqrt{\int \left(F_Y\left(x|\hat{\theta}_0\right) - F_Y\left(x|\theta_0\right)\right)^2 d\mu\left(x\right)} \end{array} \right\} \end{aligned}$$

from where we can deduce that: $\check{\Psi}_H\left(\hat{F}_X^L, \hat{F}_X^H\right) - \Psi_H\left(\hat{F}_X^L, \hat{F}_X^H\right) \geq o_p\left(1\right)$. Combining this with the previous result, we complete the step.

Step 4. In this step, we show that $\left| \tilde{\Psi}_H\left(\hat{F}_X^L, \hat{F}_X^H\right) - \check{\Psi}_H\left(\hat{F}_X^L, \hat{F}_X^H\right) \right| = o_p\left(1\right)$.

By the law of large numbers, $\forall \left(x, \theta\right) \in \left\{ L_2\left(\mathcal{D}\right), \Theta \right\}$, $\hat{F}_Y\left(x|\theta\right) \xrightarrow{p} F_Y\left(x|\theta\right)$. This asymptotic result holds for any fixed sample by increasing the number of simulations used to approximate $F_Y\left(x|\theta\right)$. Following the arguments in BHHN [**19**], it follows that,

$$\int \left(\hat{F}_Y\left(x|\hat{\theta}_0\right) - F_Y\left(x|\hat{\theta}_0\right)\right)^2 d\mu\left(x\right) = o_p\left(1\right)$$

The rest of the arguments follows from arguments that are very similar to those used in step 2.

Step 5. By theorem 3 in BHHN [**19**], $\left|\hat{\Psi}_H\left(\hat{F}_X^L,\hat{F}_X^H\right)-\tilde{\Psi}_H\left(\hat{F}_X^L,\hat{F}_X^H\right)\right|=o_p(1)$. This completes the proof. □

Proof of theorem 34. By theorem 2 in BHHN [**19**], $t^*_{\hat{\theta}_0}(1-\alpha)=t_{\theta_0}(1-\alpha)+\gamma_n$, where $\gamma_n=o_p(1)$ and $t_{\theta_0}(1-\alpha)$ is given by,

$$t_{\theta_0}(1-\alpha)=\inf_{h\in\mathbb{R}}\left\{P\left(V_0>h\right)\geq(1-\alpha)\right\}$$

where $V_0=\int\left(\zeta(x)+\left(\partial F_Y(x|\theta_0)/\partial\theta'\right)\xi\right)^2 d\mu(x)$, ζ is a Gaussian process on $[0,1]$ having the same covariance structure as the indicator process $1\left[X(t)\leq x(t):t\in\mathcal{D}\right]$ and ξ is a p-variate random variable whose mean is 0, covariance matrix is $cov\left(\Omega(X)\right)$ and satisfies $E\left(\xi\zeta(x)\right)=E\left(\Omega(X)\left(1\left[X(t)\leq x(t):t\in\mathcal{D}\right]-F_X(x)\right)\right)$. From these conditions, we deduce that $t_{\theta_0}(1-\alpha)$ is a positive and finite number.

Fix $\varepsilon=T_H\left(X,Y_{\theta_0}\right)/2$ and consider the following derivation.

$$\begin{aligned}
&P\left(n\hat{T}_H\left(X,Y_{\hat{\theta}_0}\right)\leq t^*_{\hat{\theta}_0}(1-\alpha)\right)=\\
&=\left\{\begin{array}{c}P\left(n\hat{T}_H\left(X,Y_{\hat{\theta}_0}\right)\leq t_{\theta_0}(1-\alpha)+\gamma_n\cap|\gamma_n|>\varepsilon\right)+\\ P\left(n\hat{T}_H\left(X,Y_{\hat{\theta}_0}\right)\leq t_{\theta_0}(1-\alpha)+\gamma_n\cap|\gamma_n|\leq\varepsilon\right)\end{array}\right\}\\
&\leq P\left(|\gamma_n|>\varepsilon\right)+P\left(\hat{T}_H\left(X,Y_{\hat{\theta}_0}\right)\leq\left(t_{\theta_0}(1-\alpha)+\varepsilon\right)/n\right)
\end{aligned}$$

Since $t_{\theta_0}(1-\alpha)<\infty$, $\exists N\in\mathbb{N}$ such that $\forall n\geq N$,

$$\left(t_{\theta_0}(1-\alpha)+\varepsilon\right)/n-T_H\left(X,Y_{\theta_0}\right)<-T_H\left(X,Y_{\theta_0}\right)/2=-\varepsilon$$

Therefore, $\forall n \geq N$,

$$P\left(n\hat{T}_H\left(X, Y_{\hat{\theta}_0}\right) \leq t^*_{\hat{\theta}_0}\left(1-\alpha\right)\right) \leq P\left(\left|\gamma_n\right| > \varepsilon\right) + P\left(\left|\hat{T}_H\left(X, Y_{\hat{\theta}_0}\right) - T_H\left(X, Y_{\theta_0}\right)\right| > \varepsilon\right)$$

taking limits on both sides, we deduce that $P\left(n\hat{T}_H\left(X, Y_{\hat{\theta}_0}\right) \leq t^*_{\hat{\theta}_0}\left(1-\alpha\right)\right) = o\left(1\right)$, which completes the proof. □

Proof of theorem 35. This proof follows same arguments as the proof of theorem 34. □