MONASH University

MONASH BUSINESS SCHOOL

DEPARTMENT OF ECONOMETRCS AND BUSINESS STATISTICS

ISSN 1440-771X

WORKING PAPER SERIES

Bootstrap Hausdorff Confidence Regions for Average Treatment Effect Identified Sets

D.S. Poskitt and Xueyan Zhao

Working Paper No. 09/23
May 2023

Bootstrap Hausdorff Confidence Regions for Average Treatment Effect Identified Sets

D. S. Poskitt and Xueyan Zhao*

Version of May 26, 2023

Abstract

This paper introduces a new bootstrap approach to the construction of confidence regions for Average Treatment Effect (ATE) identified sets. Minimum Hausdorff distance bootstrap confidence regions are developed and shown to be valid under suitable regularity. A novel measure of the discrepancy between a confidence region and the target identified set is advanced that contains two components analogous to conventional hypothesis test Type I and Type II errors. Monte Carlo experimentation is employed to compare the behaviour of the new confidence regions with an existing state of the art approach and the impact of different features on the properties of the alternative techniques are investigated. Properties arising from the application of quasi-maximum likelihood estimation as a tool for conducting inference on ATEs are also examined.

Keywords: binary models, bounds, coverage, partial identification.

1 Introduction

In many policy analysis and program evaluation studies the focus of interest centres on conducting inference on the average treatment effect (ATE) using models that allow for meaningful restrictions to be placed on the possible values of the ATE in the form of an identified set (or bounds). And various treatment effect bounds have been derived in the literature under different assumptions concerning the underlying structure that gives rise

*Department of Econometrics and Business Statistics, Monash University, Victoria 3800, Australia. Email: Donald.Poskitt@monash.edu and Xueyan.Zhao@monash.edu

to the observed data. For a comparison of the properties of different ATE bounds and reviews of the extant literature, see *inter alios* Kitagawa (2009), Tamer (2010), Ho and Rosen (2013), Flores and Chen (2018) and Swanson et al. (2018). Our aim in this paper is to contribute to the literature on conducting inference on identified sets in such partially identified models.

It is now reasonably well known that asymptotically valid confidence regions for identified sets can be derived, for such methods have received considerable attention in the recent literature (See Imbens and Manski, 2004; Chernozhukov et al., 2007; Rosen, 2008; Romano and Shaikh, 2010; Andrews and Soares, 2010; Canay, 2010; Andrews and Barwick, 2012, and the references contained therein, for example.). As pointed out in Imbens and Manski (2004), there are two conceptually different definitions of a confidence region for a partially identified set. One relates to a point-based definition of coverage. If we denote ATE, the parameter of interest, by η, and the ATE identified set by $[\eta_l, \eta_u]$, then the point based definition results in a random set $CR_{(1-\alpha)}^{[\eta_l,\eta_u]}$ that depends on the data and satisfies

$$\lim_{n\to\infty} \inf_{\eta\in[\eta_l,\eta_u]} \Pr\left(\eta \in CR_{(1-\alpha)}^{[\eta_l,\eta_u]}\right) \geq 1-\alpha\,. \tag{1}$$

That is, a confidence region $CR_{(1-\alpha)}^{[\eta_l,\eta_u]}$ that covers each point in the identified set with a prescribed probability of at least $1-\alpha$. The other is a set-based concept, defined as a confidence region that covers the entire identified set with a given probability. Such a confidence region for $[\eta_l, \eta_u]$ satisfies

$$\lim_{n\to\infty} \Pr\left([\eta_l, \eta_u] \subseteq CR_{(1-\alpha)}^{[\eta_l,\eta_u]}\right) \geq 1-\alpha\,. \tag{2}$$

which is a stricter condition than that prescribed in Equation (1).

In this paper we are interested in using the bootstrap to construct confidence regions for identified sets that satisfy the set-based definition. Horowitz and Manski (2000) proposed employing confidence regions of the form $CR_{(1-\alpha)}^{[\eta_l,\eta_u]} = [\eta_{ln} - c_{\alpha n}, \eta_{un} + c_{\alpha n}]$ where; (i) η_{ln} and η_{un} are point estimates of the lower and upper bounds obtained by replacing the population probabilities by their empirical counterparts, and (ii) $c_{\alpha n}$ is derived by re-sampling the observations on the treatment and response of interest, together with their associated regressor and instrument values, and using the bootstrap distribution of the lower and upper bound estimates to determine the appropriate adjustment, a procedure akin to Efron's percentile method. Bugni (2010) also uses the bootstrap to construct confidence regions for identified sets. Bugni builds upon the criterion function approach and a resampling technique by Chernozhukov et al. (2007) to construct confidence regions for identified sets defined by finitely many moment inequalities. He shows that if one is working within this

framework then his procedure will yield a confidence region that satisfies Equation (2). Chernozhukov et al. (2013) point out, however, that analog estimates of $[\eta_l, \eta_u]$ based on intersection bounds will be biased due to the convexity and concavity of the supremum and infimum operations involved in their construction. To address this issue they advocate using "precision corrected" estimates to take account of sampling variation in the bounding functions and construct confidence regions for ATE identified sets based on these.

This paper presents an adaptation of the previous approaches and use the bootstrap to construct what we have christened a minimum Hausdorff distance (MHD) bootstrap confidence region for the ATE identified set $[\eta_l, \eta_u]$. We are able thereby to allow for sampling variation in the bounding functions, and to also minimise the length of the confidence region whilst maintaining a specified coverage probability. We also propose a measure of the performance of a confidence region for an identified set that is analogous to a volumetric analysis. The measure separately evaluates two types of errors that we designate as a Type I discrepancy and a Type II discrepancy. Using our proposed performance measures, we evaluate and compare the performance of our MHD bootstrap confidence region with that of an alternative confidence region in the literature, via simulastion.

The paper proceeds by outlining in Section 2 the Chesher model (Chesher, 2005, 2010) which we use as the framework for the development and demonstration of our general methodology. Section 3 then presents our first contribution to the literature on inference for identified sets. In Section 3 we present the MHD bootstrap algorithm and establish that the confidence region will satisfy Equation (2) for any given confidence level $1-\alpha$ when implemented using the empirical distribution function. We also show that, under suitable regularity, MHD bootstrap confidence regions maintain their validity when constructed using alternative non-parametric estimators and M-estimators.

The second contribution of this paper is to provide in Section 4 a novel characterisation of the performance features of a confidence region $CR_{(1-\alpha)}^{[\eta_l,\eta_u]}$ for an identified set $[\eta_l, \eta_u]$ that is analogous to a volumetric analysis. We present measures of the discrepancy between $CR_{(1-\alpha)}^{[\eta_l,\eta_u]}$ and the target $[\eta_l, \eta_u]$ that recognise that although conventional notions of 'accuracy' and 'precision' are sufficient to characterise the performance of a confidence interval for a point, they do not provide a complete picture of the properties of a confidence region for a set. We outline six possible outcomes that delineate the relative locations of $CR_{(1-\alpha)}^{[\eta_l,\eta_u]}$ and $[\eta_l, \eta_u]$, and separately measure two types of errors that we designate: *omitted coverage*, those parts of $[\eta_l, \eta_u]$ that are not included in $CR_{(1-\alpha)}^{[\eta_l,\eta_u]}$, called a Type I discrepancy; and *false coverage*, those parts of the parameter space that are included in $CR_{(1-\alpha)}^{[\eta_l,\eta_u]}$ but are not part of $[\eta_l, \eta_u]$, called a Type II discrepancy.

In Section 5 the proposed discrepancy measures are employed in numerical simulations designed to illustrate the finite sample properties of MHD bootstrap confidence regions when using correctly specified bounding functions, and thus to demonstrate confirmation of concept. As a third contribution to the literature, we also present a comparison of the finite sample performance of the confidence region introduced in Chernozhukov et al. (2013) (labeled CLR) with that of our proposed MHD bootstrap confidence region. Using coverage probability and the Type I and Type II discrepancy measures as a basis for comparison, the performance characteristics of MHD confidence regions are shown to compare favourably with those of CLR.

The fourth contribution of this paper is to investigate in Section 6 the consequences of using a bivariate probit (BVP) quasi maximum likelihood estimator (QMLE) as a basis for conducting inference on ATEs. The BVP model is commonly employed by empirical researchers as a tool for conducting causal inference with binary treatment and binary outcome data, and Manski (1988) and Vytlacil (2006) demonstrate that under relatively mild conditions a wide class of latent index processes can be represented using a BVP specification. Therefore, when data scarcity renders the use of nonparametric techniques impracticable, the use of the BVP model to estimate ATE bounds is an obvious choice that practitioners might make. Experimental results are presented that investigate the consequences of making such a choice in scenarios where the true structure is unknown and the misspecification in the BVP model is severe. We present hitherto undocumented mathematical properties of the BVP model via an examination of a Gram-Charlier expansion. This yields properties of the BVP model that form a technical explanation of the somewhat incongruous finite sample performance characteristics of BVP QMLE based CLR and MHD confidence regions observed in the experimental outcomes.

Finally, Section 7 presents a summary of the paper and is followed by a supplementary appendix that presents the proofs.

2 The Chesher Model

In this paper we will use the ATE bounds developed in Chesher (2005, 2010) as the vehicle to develop and demonstrate our general methodology. Bearing in mind the adage in Manski (2003) that the credibility of inference decreases with the strength of the assumptions maintained, we are motivated in this choice by the broad applicability of the Chesher model. As observed above, several different treatment effect bounds have been discussed in the extant literature. Practitioners can avail themselves of any one of these, and subject to suitable adaptation the MHD bootstrap methodology presented here could

then be applied to the treatment effect bounds employed in order to conduct appropriate inference.

Consider, then, a model for a scalar binary outcome variable Y and a binary endogenous treatment D with a structural equation $Y = h(D, \mathbf{X}, U)$, where $\mathbf{X}$ denotes a vector of exogenous numerical features or explanatory variables with support $\Omega_{\mathbf{X}}$, and U is an unobserved scalar random variable that is continuously distributed and normalised to be uniformly distributed on $(0, 1)$. Suppose that the following assumptions hold:
Assumption C1: the structure function $h(d, \mathbf{x}, u)$ is weakly monotonic in $u \in (0, 1)$ for every $d \in \{0, 1\}$ and $\mathbf{x} \in \Omega_{\mathbf{X}}$;
Assumption C2: there exists a vector of instruments $\mathbf{Z}$ that is independently distributed from U such that $P[U \leq \tau | \mathbf{Z} = \mathbf{z}] = \tau$ for all $\tau \in (0, 1)$ and all $\mathbf{z} \in \Omega_{\mathbf{Z}}$.

Given Assumption C1, there exists a threshold function $p(d, \mathbf{x})$ such that

$$Y = h(D, \mathbf{X}, U) = \begin{cases} 0, \ 0 < U \leq p(D, \mathbf{X}) \\ 1, \ p(D, \mathbf{X}) < U \leq 1 \end{cases} \tag{3}$$

where the support of Y and the dummy treatment variable D is taken as $\{0, 1\}$. Following the potential outcome framework of Neyman (1923) and Rubin (1974), from Equation (3) we have, trivially, that the distributions of the binary potential outcomes Y_0 and Y_1 can be expressed as $Y_d = h(d, \mathbf{X}, U) = \mathbf{1}\{ \ p(d, \mathbf{X}) < U \leq 1\}$, $d \in \{0, 1\}$, where $\mathbf{1}\{A\}$ denotes the usual indicator function for the event A. It then follows that the ATE for an individual with characteristics $\mathbf{X} = \mathbf{x} \in \Omega_{\mathbf{X}}$ is $ATE(\mathbf{x}) = \mathbb{E}[Y_1 - Y_0 | \, \mathbf{x}] = p(0, \mathbf{x}) - p(1, \mathbf{x})$.

For a given conditional distribution function of (Y, D) given $\mathbf{Z}$ and $\mathbf{X}$, denoted by $F_{YD|\mathbf{ZX}}$, there can be more than one admissible structure $\mathfrak{S} \equiv \left\{h, F_{UD|\mathbf{ZX}}\right\}$ with distinct structure function h that is consistent with the model and that can deliver an identical $F_{YD|\mathbf{ZX}}$. Such structures are observationally equivalent and the model is set identified. Crucially, the existence of instruments as in Assumption C2 places limitations on the variation of $p(D, \mathbf{X})$. In particular, it can be shown (Chesher, 2005, 2010) that for any given distribution $F_{YD|\mathbf{ZX}}$, members of the family of observationally equivalent structures admitted by the model must satisfy probability inequalities that provide bounds for the threshold functions $p(0, \mathbf{x})$ and $p(1, \mathbf{x})$. Intersecting these bounds across all values of $\mathbf{z} \in \Omega_{\mathbf{Z}}$ yields the identified set of $ATE(\mathbf{x})$ as an interval $[\eta_l(\mathbf{x}), \eta_u(\mathbf{x})]$, where the lower bound $\eta_l(\mathbf{x})$

and upper bound $\eta_u(\mathbf{x})$ are as given in the following expressions:

$$\left[\sup_{\mathbf{z}\in\Omega_\mathbf{Z}} \Pr(\{Y=0, D=0\}|\mathbf{x},\mathbf{z}) - \inf_{\mathbf{z}\in\Omega_\mathbf{Z}} \{\Pr(\{D=0\}|\mathbf{x},\mathbf{z}) + \Pr(\{Y=0, D=1\}|\mathbf{x},\mathbf{z})\}, \right.$$
$$\left. \inf_{\mathbf{z}\in\Omega_\mathbf{Z}} \{\Pr(\{Y=0\}|\mathbf{x},\mathbf{z})\} - \sup_{\mathbf{z}\in\Omega_\mathbf{Z}} \{\Pr(\{Y=0\}|\mathbf{x},\mathbf{z})\}\right], \tag{4}$$

in the case where $p(0,\mathbf{x}) \leq p(1,\mathbf{x})$ and $ATE(\mathbf{x}) \leq 0$, and

$$\left[\sup_{\mathbf{z}\in\Omega_\mathbf{Z}} \{\Pr(\{Y=0\}|\mathbf{x},\mathbf{z})\} - \inf_{\mathbf{z}\in\Omega_\mathbf{Z}} \{\Pr(\{Y=0\}|\mathbf{x},\mathbf{z})\}, \right.$$
$$\left. \inf_{\mathbf{z}\in\Omega_\mathbf{Z}} \{\Pr(\{D=1\}|\mathbf{x},\mathbf{z}) + \Pr(\{Y=0, D=0\}|\mathbf{x},\mathbf{z})\} - \sup_{\mathbf{z}\in\Omega_\mathbf{Z}} \Pr(\{Y=0, D=1\}|\mathbf{x},\mathbf{z})\right], \tag{5}$$

in the case where $p(0,\mathbf{x}) > p(1,\mathbf{x})$ and $ATE(\mathbf{x})$ is positive. For any $\mathbf{x} \in \Omega_\mathbf{X}$, the intersection bounds are obtained by evaluating (4) and (5) using the probabilities given by the distribution $F_{YD|\mathbf{ZX}}$ of the DGP observational equivalence class $\mathfrak{S}$. Given data, estimation of the ATE bounds and subsequent inference obviously turn on the evaluation of Equations (4) and (5).[1]

3 Minimum Hausdorff Distance ATE Confidence Regions

3.1 The Methodological Framework

Suppose that an admissible structure $\mathfrak{S}$ admitted by the Chesher (2010) model delivers the probability distribution $F_{YD|\mathbf{XZ}}$ for the random variable $(Y, D|\mathbf{X}, \mathbf{Z})$. Then the ATE bounds are functionals of the probabilities in Equations (4) and (5), which are in turn derived from

$$F_{YD|\mathbf{XZ}}(y(\mathbf{x},\mathbf{z}), d(\mathbf{x},\mathbf{z})) = \Pr(Y \leq y, D \leq d|\mathbf{X}=\mathbf{x}, \mathbf{Z}=\mathbf{z}),$$

where $(y,d) \in \Omega_{(Y,D)} = \{0,1\} \times \{0,1\}$ and $(\mathbf{x},\mathbf{z}) \in \Omega_\mathbf{X} \times \Omega_\mathbf{Z}$. Let $[\eta_l, \eta_u] = \psi(F)$ denote the mapping from F to the ATE identified set $[\eta_l, \eta_u]$, using an obvious notational

[1] As we have previously observed, Chernozhukov et al. (2013) advocate adjusting analog estimates of $[\eta_l, \eta_u]$ based on intersection bounds using "half-median unbiased" estimates to correct for bias. In the illustrations presented below we have therefore constructed precision corrected estimates of the bounding functions using Algorithm 1 of Chernozhukov et al. (2013, p.708).

simplification in which the subscript and arguments of the distribution are suppressed. Let $[\eta_{ln}^{(a)}, \eta_{un}^{(a)}] = \psi(F_n^{(a)})$, where $F_n^{(a)}$ denotes alternative estimators of F that can be constructed from the data $\mathbf{W}_i = (Y_i, D_i, \mathbf{X}_i, \mathbf{Z}_i)$, $i = 1, \ldots, n$, and that can be used to estimate the probabilities in Equations (4) and (5) from which the intersection bounds are calculated.

In order to take into account sampling variation in $F_n^{(a)}$ and construct Hausdorff confidence regions for the ATE intersection bounds, we employ a bootstrap re-sampling procedure in which observations on the treatment and response of interest, together with their associated regressor and instrument values, are sampled randomly with replacement from $\mathbf{W}_1 \ldots, , \mathbf{W}_n$. From each bootstrap sample $\mathbf{W}_1^* \ldots, , \mathbf{W}_n^*$ an estimate of the ATE identified set $[\eta_{ln}^{*(a)}, \eta_{un}^{*(a)}]$ is evaluated, and by repeated bootstrap sampling a bootstrap empirical distribution $[\eta_{ln,b}^{*(a)}, \eta_{un,b}^{*(a)}]$, $b = 1, \ldots, B$, is generated. The MHD bootstrap confidence region is then determined by removing up to $\alpha 100\%$ of the bootstrap intervals so as to yield an interval $cr_{(1-\alpha)}^* = [c_{ln(1-\alpha)}^*, c_{un(1-\alpha)}^*]$, comprised of the union of the remaining intervals, such that $c_{ln(1-\alpha)}^* \leq \eta_{ln}^{(a)}$ and $c_{un(1-\alpha)}^* \geq \eta_{un}^{(a)}$, the cardinality of the set $\{[\eta_{ln,b}^{*(a)}, \eta_{un,b}^{*(a)}] : [\eta_{ln,b}^{*(a)}, \eta_{un,b}^{*(a)}] \subseteq cr_{(1-\alpha)}^*\}$ is at least $(1-\alpha)B$, and the Hausdorff (Lebesgue) measure of the interval, $\nu(cr_{(1-\alpha)}^*)$, is the smallest. This effectively entails removing $\alpha 100\%$ of the bootstrap intervals that have the largest Hausdorff distance from the union of those remaining. The steps required to construct a $(1-\alpha)100\%$ MHD bootstrap confidence region can be described as follows:

(1) Draw an independent random sample of bootstrap values $\mathbf{W}_1^* \ldots, , \mathbf{W}_n^*$ from the empirical distribution function F_n of the data $\mathbf{W}_1 \ldots, , \mathbf{W}_n$, and construct $[\eta_{ln}^{*(a)}, \eta_{un}^{*(a)}] = \psi(F_n^{*(a)})$ where $F_n^{*(a)}$ denotes the probability distribution estimate obtained from the bootstrap sample $\mathbf{W}_1^* \ldots, , \mathbf{W}_n^*$ when employing the estimator $F_n^{(a)}$;

(2) Repeat Step (1) B times, and for each bootstrap sample evaluate $[\eta_{ln}^{*(a)}, \eta_{un}^{*(a)}]$ and index each of these bootstrap identified set estimates as $[\eta_{ln,b}^{*(a)}, \eta_{un,b}^{*(a)}]$, $b = 1, \ldots, B$.

(3) Initiate recursions at $r = 0$ with $cr_{(1-\alpha)}^* = [c_{ln(1-\alpha)}^*, c_{un(1-\alpha)}^*]$ where

$$[c_{ln(1-\alpha)}^*, c_{un(1-\alpha)}^*] = \left\{ \bigcup_{b=1}^{B} [\eta_{ln,b}^{*(a)}, \eta_{un,b}^{*(a)}] \right\} \bigcup [\eta_{ln}^{(a)}, \eta_{un}^{(a)}] \,.$$

(4) (a) From the $B-r$ bootstrap intervals currently contained in the interval $cr_{(1-\alpha)}^*$, determine the bootstrap interval $[\eta_{ln,b}^{*(a)}, \eta_{un,b}^{*(a)}]$, with index $b = b_{r+1}$, whose removal reduces the magnitude of $\nu(cr_{(1-\alpha)}^*)$ the most.

(b) Set B_r equal to the set $\{b_1, \ldots, b_{r+1}\}$ and update the interval $cr^*_{(1-\alpha)}$ to

$$cr^*_{(1-\alpha)} = [c^*_{ln(1-\alpha)}, c^*_{un(1-\alpha)}] = \left\{ \bigcup_{\substack{b=1 \\ b \notin B_r}}^{B} [\eta^{*(a)}_{ln,b}, \eta^{*(a)}_{un,b}] \right\} \bigcup [\eta^{(a)}_{ln}, \eta^{(a)}_{un}] .$$

(5) If $(B - r - 1)/B > (1 - \alpha)$ set $r = r + 1$ and repeat Step (4). Otherwise, set $MHD^{(a)}_{(1-\alpha)} = cr^*_{(1-\alpha)}$ and stop.

This yields, by construction, a confidence region $[c^*_{ln(1-\alpha)}, c^*_{un(1-\alpha)}]$ spanning $[\eta^{(a)}_{ln}, \eta^{(a)}_{un}]$ that; (i) contains at least $(1 - \alpha)100\%$ of the B bootstrap estimates $[\eta^{*(a)}_{ln,b}, \eta^{*(a)}_{un,b}]$, $b = 1, \ldots, B$, (ii) has the shortest Hausdorff distance to the intervals it contains, and (iii) has the smallest Hausdorff measure.

In studies of partial identification it is not uncommon to examine the performance of estimators by evaluating the Hausdorff distance between the true identified set and the estimated identified set (see, *inter alios*, Beresteanu and Molinari, 2008; Menzel, 2014) and the use of MHD confidence sets as here seems natural.

3.2 Asymptotic Validity

Thus far, $F^{(a)}_n$ has been used to denote that alternative estimates of F can be constructed from data $\mathbf{w}_i = (y_i, d_i, \mathbf{x}_i, \mathbf{z}_i)$, $i = 1, \ldots, n$, and any one of these might be used to estimate the probabilities from which the ATE and ATE intersection bounds are calculated. For a given $F^{(a)}_n$ the ATE intersection bounds are obtained by inserting the required probability estimates into Equations (4) and (5), respectively.

Assume that $\mathbf{w}_1 \ldots, , \mathbf{w}_n$ is a simple random sample of values of $\mathbf{W}$, a random variable whose stochastic properties can be characterised by a structure $\mathfrak{S}$ with probability distribution F that belongs to a set of probability measures dominated by a Borel σ-finite measure. Let

$$\mathbb{F}_n(y, d) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{y_i \leq y\} \mathbf{1}\{d_i \leq d\} \mathbf{1}\{\mathbf{X} = \mathbf{x}_i\} \mathbf{1}\{\mathbf{Z} = \mathbf{z}_i\} \tag{6}$$

denote the empirical distribution for the observed data $\mathbf{w}_i = (y_i, d_i, \mathbf{x}_i, \mathbf{z}_i)$, $i = 1, \ldots, n$. By the Glivenko-Cantelli theorem the bootstrap empirical distribution $[\eta^{*(a)}_{ln,b}, \eta^{*(a)}_{un,b}]$, $b = 1, \ldots, B$, obtained using the raw non-parametric estimates in (6) will, as $B \to \infty$, provide a consistent estimate of the bootstrap distribution of $[\eta^{*(a)}_{ln}, \eta^{*(a)}_{un}]$ with respect to the multinomial distribution of $\mathbf{W}^*_1 \ldots, , \mathbf{W}^*_n$ obtained by independently sampling with re-

placement from $\mathbf{W}_1 \ldots, , \mathbf{W}_n$. The following result confirms that the coverage probability of MHD bootstrap intersection bounds based upon using $F_n^{(a)} = \mathbb{F}_n(y, d)$ will achieve the required significance level as $n \to \infty$.

Theorem 1 *If the estimator $F_n^{(a)} = \mathbb{F}_n$, in other words if $[\eta_{ln}^{(a)}, \eta_{un}^{(a)}] = \psi(F_n^{(a)}) = \psi(\mathbb{F}_n)$, and the MHD bootstrap confidence intervals are constructed using the empirical distribution function, then*

$$\lim_{n\to\infty} Pr\left([\eta_l, \eta_u] \subseteq \textit{MHD}_{(1-\alpha)}^{(a)}\right) \geq 1 - \alpha\,.$$

That evaluation of the MHD using the empirical distribution yields the correct asymptotic coverage probability has theoretical import in that, by the Dynkin-Lehmann-Scheffé theorem, the relative frequencies that make up the empirical distribution constitute minimal sufficient statistics for the corresponding probabilities; but practical implementation of MHD using the empirical distribution function can be rendered unfeasible due to data paucity. In situations where $\mathbf{X}$ and $\mathbf{Z}$ contain several discrete valued variables, such as categorical and qualitative indicators, or essentially continuous variables that are measured on a discrete scale (age broken down into decades, or income into deciles, for example), there can be so many possible combinations that cross tabulation cells required in the estimation of various conditional probabilities via observed relative frequencies can have too few observations or be empty.

Remark 3.1. Data paucity of this kind is not uncommon with survey data. For example, in an analysis of dental care utilisation using Australian National Health Survey (ANHS) data, Li et al. (2019b) found that there were insufficient observations to evaluate non-parametric estimates for several different types of individual even though the ANHS data contained a representative sample of over 19,500 adult records. ∎

Data paucity issues have undoubtedly contributed to practitioners commonly employing other non-parametric, and parametric, techniques to calculate ATEs and ATE bounds. A consideration of the properties of more general estimators $F_n^{(a)}$ other than the raw non-parametric estimator $\mathbb{F}_n$ therefore seems pertinent. An estimator $F_n^{(a)}$ is obviously by definition a function of the data and can be expressed as $F_n^{(a)} = \lambda^{(a)}(\mathbb{F}_n)$, and therefore $\psi(F_n^{(a)}) = \psi \circ \lambda^{(a)}(\mathbb{F}_n) = \psi(\lambda^{(a)}(\mathbb{F}_n))$.

Corollary 1 *Suppose that the estimator $F_n^{(a)}$ used to calculate the ATE bounds and implement the MHD is such that $F_n^{(a)} = \lambda^{(a)}(\mathbb{F}_n)$ is a Fréchet-differentiable function of*

$\mathbb{F}_n$ *and* $\sqrt{n}\|F_n^{(a)} - \mathbb{F}\|_\infty = o(\sqrt{\log\log n})$. *Then*

$$\lim_{n\to\infty} Pr\left([\eta_l, \eta_u] \subseteq \textit{MHD}^{(a)}_{(1-\alpha)}\right) \geq 1 - \alpha\,.$$

The requirement in Corollary 1 that the order of magnitude of $\|F_n^{(a)} - \mathbb{F}\|_\infty$ be smaller than $(\log\log n/n)^{\frac{1}{2}}$ is stringent, it implies that fluctuations in the uniform norm of the empirical process $\sqrt{n}(\mathbb{F} - F)$ in its passage to a Gaussian limit exceed those of $\sqrt{n}(F_n^{(a)} - \mathbb{F})$. Regularity conditions under which it will hold will obviously have to be established on a case-by-case basis according to the specification of $F_n^{(a)}$ and the postulated structure underlying the data.

3.2.1 Implementation Using M-Estimators

In light of the issues raised by data paucity, consider a scenario where an applied worker uses a parametric model $F_n^{(\boldsymbol{\theta})}$ to characterise $F_{YD|\mathbf{XZ}}$ where the parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)' \in \Theta \subset \mathbb{R}^p$ is given by $\widehat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\Theta}\{M_n(\boldsymbol{\theta})\}$ where $M_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n m(\boldsymbol{\theta}, \mathbf{w}_i)$. Thus, in terms of our current notation we have $[\eta_{ln}^{(\widehat{\boldsymbol{\theta}}_n)}, \eta_{un}^{(\widehat{\boldsymbol{\theta}}_n)}] = \psi(F_n^{(\widehat{\boldsymbol{\theta}}_n)})$, where the lower bound $\eta_{ln}^{(\boldsymbol{\theta})}(\mathbf{x})$ and upper bound $\eta_{un}^{(\boldsymbol{\theta})}(\mathbf{x})$ for any $\mathbf{x} \in \Omega_{\mathbf{X}}$ are known functions parameterized by the p-dimensional vector $\boldsymbol{\theta}$, which is set equal to $\widehat{\boldsymbol{\theta}}_n$, the value given by the M-estimator.

A natural choice for the M-function is $n^{-1}\sum_{i=1}^n \log p_n^{(\boldsymbol{\theta})}(\mathbf{w}_i)$ where $p_n^{(\boldsymbol{\theta})}(\mathbf{w}_i)$ is the probability function of the applied worker's postulated model, in which case of course $\widehat{\boldsymbol{\theta}}_n$ corresponds to the MLE of the assumed parametric model. Researchers will have access to a range of different choices for $M_n(\boldsymbol{\theta})$ however. Those concerned with robustness, or those interested in quantile treatment effects, may select different influence functions for example. Suppose that the functions that might be used are such that the following assumption holds.

Assumption M1: $M_n(\boldsymbol{\theta})$ is a measurable function of the data $\mathbf{w}_1 \ldots, , \mathbf{w}_n$ for all $\boldsymbol{\theta} \in \Theta$, a convex subset of $\mathbb{R}^p$. There exists a non-stochastic quasi-concave function $\bar{M}_n(\boldsymbol{\theta})$ such that $|M_n(\boldsymbol{\theta}) - \bar{M}_n(\boldsymbol{\theta})| \to 0$ holds on sets $\mathfrak{B}_n$ with $\Pr(\mathfrak{B}_n) \to 1$ as $n \to \infty$ for all $\boldsymbol{\theta} \in \Theta$.

Then for any sequence $\widehat{\boldsymbol{\theta}}_n$ such that $M_n(\widehat{\boldsymbol{\theta}}_n) \geq \sup_{\boldsymbol{\theta}\in\Theta} M_n(\boldsymbol{\theta}) - \delta_n$ where $\delta_n > 0$ and $\delta_n \to 0$ as $n \to \infty$, we have $\|\widehat{\boldsymbol{\theta}}_n - \bar{\Theta}_n\| = \inf_{\boldsymbol{\theta}\in\bar{\Theta}_n}\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\| \to 0$ in probability where $\bar{\Theta}_n = \{\boldsymbol{\theta} \in \Theta : \bar{M}_n(\boldsymbol{\theta}) \geq \bar{c}_n\}$, $\bar{c}_n = \sup_{\boldsymbol{\theta}\in\Theta}\bar{M}_n(\boldsymbol{\theta})$. Assume that $\bar{M}_n(\boldsymbol{\theta})$ is twice continuously differentiable with respect to $\boldsymbol{\theta}$ with a negative-definite second derivative for all $\boldsymbol{\theta} \in \bar{\Theta}_n$ where $\bar{\Theta}_n \subset \Theta\circ$. Furthermore, suppose that the centered process $M_n(\boldsymbol{\theta}) - \bar{M}_n(\boldsymbol{\theta})$ satisfies $\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_{0n}\|<\epsilon}\sqrt{n}|M_n(\boldsymbol{\theta}) - \bar{M}_n(\boldsymbol{\theta}) - M_n(\boldsymbol{\theta}_{0n}) - \bar{M}_n(\boldsymbol{\theta}_{0n})| = O_p(\epsilon)$, $\epsilon > 0$, and

that $M_n(\widehat{\boldsymbol{\theta}}_n) \geq M_n(\boldsymbol{\theta}_{0n}) - \delta_n$ where $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{0n}\| = \|\widehat{\boldsymbol{\theta}}_n - \bar{\Theta}_n\|$ and $\delta_n = O_p(n^{-1})$. Set the centered and re-scaled gradient $\mathbb{G}_n(\boldsymbol{\theta}) = n^{\frac{1}{2}} \left\{\partial(M_n(\boldsymbol{\theta}) - \bar{M}_n(\boldsymbol{\theta}))/\partial\boldsymbol{\theta}\right\}$ and assume that $\mathbb{G}_n(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta$ yields an empirical process with covariance kernel $\bar{\boldsymbol{\Sigma}}_n(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = nE[\mathbb{G}_n(\boldsymbol{\theta}_1)\mathbb{G}_n(\boldsymbol{\theta}_2)']$, $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$, that satisfies the following functional central limit property;

Assumption M2: The empirical process $\mathbb{G}_n(\boldsymbol{\theta}) = \bar{\boldsymbol{\Sigma}}_n^{\frac{1}{2}}(\boldsymbol{\theta})\mathbb{B}_n(\boldsymbol{\theta})$ where $\mathbb{B}_n(\boldsymbol{\theta}) \Rightarrow \mathbb{B}(\boldsymbol{\theta})$, a zero mean Gaussian stochastic process indexed by $\boldsymbol{\theta} \in \Theta$ with bounded continuous sample paths and an identity covariance kernel, the symbol $\Rightarrow$ denotes weak convergence of random functions on Θ with respect to the supremum norm, and $\bar{\boldsymbol{\Sigma}}_n^{\frac{1}{2}}(\boldsymbol{\theta})$ denotes the symmetric square root of $\bar{\boldsymbol{\Sigma}}_n(\boldsymbol{\theta})$ where $\bar{\boldsymbol{\Sigma}}_n(\boldsymbol{\theta}) = \bar{\boldsymbol{\Sigma}}_n(\boldsymbol{\theta}, \boldsymbol{\theta})$.

From a mean value expansion of the gradient it follows that under Assumptions M1 and M2 we have $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{0n}) = -\bar{\boldsymbol{\Delta}}_{0n}^{-1}\bar{\boldsymbol{\Sigma}}_{0n}^{\frac{1}{2}}\mathbb{B}_n(\boldsymbol{\theta}_{0n}) + o_p(1)$ where $\bar{\boldsymbol{\Delta}}_{0n} = \partial^2\bar{M}_n(\boldsymbol{\theta}_{0n})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$, $\bar{\boldsymbol{\Sigma}}_{0n} = \bar{\boldsymbol{\Sigma}}_n(\boldsymbol{\theta}_{0n})$ and $\mathbb{B}_n(\boldsymbol{\theta}_{0n}) \Rightarrow \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. Employing the Cramér-Wold device and Slutsky's theorem we can therefore conclude that the distribution of $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{0n})$ will be approximately normal with zero mean and variance-covariance $\boldsymbol{V}(\boldsymbol{\theta}_{0n}) = \bar{\boldsymbol{\Delta}}_{0n}^{-1}\bar{\boldsymbol{\Sigma}}_{0n}\bar{\boldsymbol{\Delta}}_{0n}^{-1}$.

For detailed particulars of the properties of M-estimators and the associated regularity conditions see Van der Vaart and Wellner (1996, Section 3.2) and Van der Vaart (1998, Chapters 5 and 19.4). The properties outlined above lead to the following result.

Theorem 2 *Assume that* $\mathbf{w}_1 \ldots, , \mathbf{w}_n$ *is a simple random sample of values of a process characterised by a structure* $\mathfrak{S}$ *with probability distribution* F*, and that* F *is modelled using a parametric model* $F_n^{(\boldsymbol{\theta})}$*. Suppose that* $F_n^{(\boldsymbol{\theta})}$ *is differentiable in* $\boldsymbol{\theta}$ *and that an M-estimator that satisfies Assumption M1 and Assumption M2 is used to estimate* $\boldsymbol{\theta}$*. Then the conditional distribution of* $\sqrt{n}(\psi(F_n^{(\widehat{\boldsymbol{\theta}}_n^*)}) - \psi(F_n^{(\widehat{\boldsymbol{\theta}}_n)}))$ *given* $\mathbf{w}_1 \ldots, , \mathbf{w}_n$ *converges to the same limit as* $\sqrt{n}(\psi(F_n^{(\widehat{\boldsymbol{\theta}}_n)}) - \psi(F_n^{(\boldsymbol{\theta}_{0n})}))$ *and yields an asymptotically consistent estimate of the latter's limiting law. Moreover, if* $F_n^{(\boldsymbol{\theta}_{0n})}$ *approximates the observationl equivalence class of* $\mathfrak{S}$ *and* $\sqrt{n}(\psi(F_n^{(\boldsymbol{\theta}_{0n})}) - \psi(F)) = o(1)$ *then*

$$\lim_{n\to\infty} Pr\left([\eta_l, \eta_u] \subseteq \boldsymbol{MHD}_{(1-\alpha)}^{(\widehat{\boldsymbol{\theta}}_n)}\right) \geq 1 - \alpha\,.$$

Standard sufficient conditions comprising a law of large numbers for convergence of $M_n(\boldsymbol{\theta})$ to $\bar{M}_n(\boldsymbol{\theta}) = E[M_n(\boldsymbol{\theta})]$ and a central limit theorem via a variant of Lyapounov's condition on $\mathbb{G}_n(\boldsymbol{\theta})$, plus equicontinuity and stochastic dominance constraints on the components of $\partial M_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ and $\partial^2 M_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$ over a compact subset of Θ containing $\bar{\Theta}_n$, will ensure that Assumption M1 and Assumption M2 hold.

4 Attributes of ATE Bounds Confidence Regions

In elementary discussions of confidence intervals two attributes are considered to be of intuitive importance: (i) the probability that the interval contains the true value should be large, commonly referred to as the accuracy; and (ii) the length of the interval should be small, commonly referred to as the precision. Although accuracy and precision are sufficient to characterise the performance of a confidence interval for a point, they do not provide a complete picture of the properties of a confidence region for a set. Clearly, the length of an ATE bounds confidence region must depend on the length of the true ATE identified set if the former is to cover the latter, and coverage probability alone does not account for the mutual proximity of the two intervals, features that merit further investigation.

If we continue to let $[\eta_l, \eta_u]$ denote the true ATE bounds but adopt the shorthand $[\widehat{\eta}_{ln}, \widehat{\eta}_{un}]$ for an $(1-\alpha)100\%$ confidence region for the ATE bounds, then there are six possible outcomes that characterise the relative locations of $[\widehat{\eta}_{ln}, \widehat{\eta}_{un}]$ and $[\eta_l, \eta_u]$. These are illustrated in Figure 1. We will denote these six different event categories by $\mathcal{E}_a$, $a = I, II, III, IV, V, VI$, and refer to them as the covering, encompassed, upper intersecting, lower intersecting, upper disjoint, and lower disjoint events, respectively.

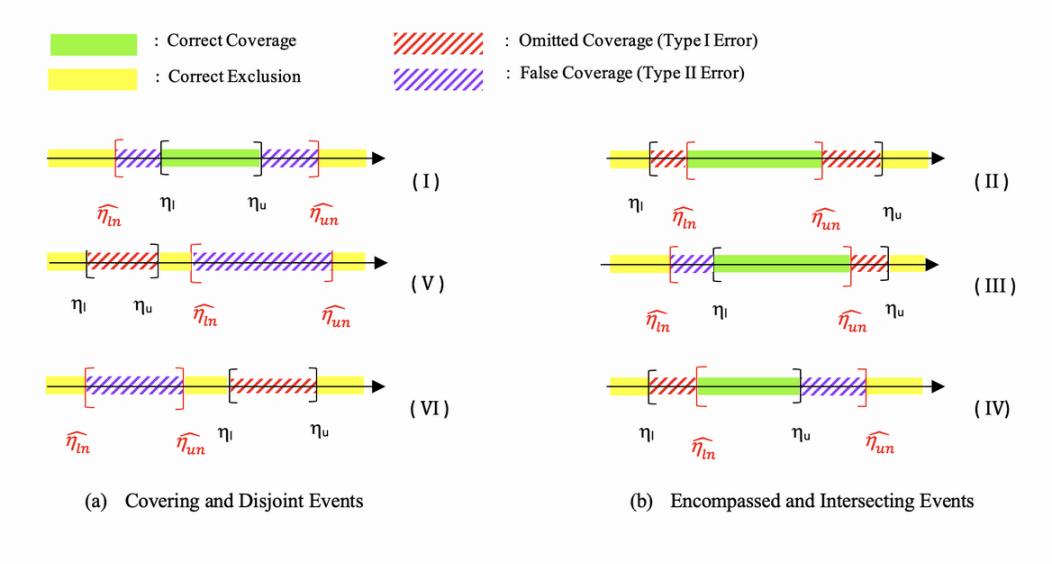


Figure 1. Confidence region event categories $[\eta_l, \eta_u] \subseteq [\widehat{\eta}_{ln}, \widehat{\eta}_{un}]$ and $[\eta_l, \eta_u] \cap [\widehat{\eta}_{ln}, \widehat{\eta}_{un}] = \emptyset$, left column, and $[\widehat{\eta}_{ln}, \widehat{\eta}_{un}] \subseteq [\eta_l, \eta_u]$ and $[\eta_l, \eta_u] \cap [\widehat{\eta}_{ln}, \widehat{\eta}_{un}] \neq \emptyset$ right column. Interval [ ] denotes the true set and [ ] the confidence region.

Turning to the performance of $[\widehat{\eta}_{ln}, \widehat{\eta}_{un}]$ as a confidence region for $[\eta_l, \eta_u]$, each event category exhibits a combination of at most four different types of region: *correct cov-*

erage $\{[\eta_l, \eta_u] \cap [\widehat{\eta}_{ln}, \widehat{\eta}_{un}]\}$ (marked in green); *correct exclusion* $\left\{\overline{[\eta_l, \eta_u]} \cap \overline{[\widehat{\eta}_{ln}, \widehat{\eta}_{un}]}\right\}$ (in yellow); *omitted coverage* $\left\{[\eta_l, \eta_u] \cap \overline{[\widehat{\eta}_{ln}, \widehat{\eta}_{un}]}\right\}$ (in red stripes); and *false coverage* $\left\{\overline{[\eta_l, \eta_u]} \cap [\widehat{\eta}_{ln}, \widehat{\eta}_{un}]\right\}$ (in purple stripes), where $\overline{[a, b]}$ denotes the complement of the interval $[a, b]$.

We can see that when $\eta_l = \eta_u = \eta$ and the ATE is point identified, the six cases in Figure 1 reduce to the three events $\mathcal{E}_I$, $\mathcal{E}_V$ or $\mathcal{E}_{VI}$. The three events $\mathcal{E}_{II}$, $\mathcal{E}_{III}$ and $\mathcal{E}_{IV}$ have measure zero, $\mathcal{E}_{II}$ corresponds to the event $\widehat{\eta}_{ln} = \widehat{\eta}_{un} = \eta$, and $\mathcal{E}_{III}$ and $\mathcal{E}_{IV}$ correspond to the cases where η falls on the right hand and left hand boundary of $[\widehat{\eta}_{ln}, \widehat{\eta}_u]$, respectively. The coverage probability $\Pr(\mathcal{E}_I) = \Pr(\eta \in [\widehat{\eta}_{ln}, \widehat{\eta}_{un}])$ equates to conventional accuracy, with η a correctly covered point. Precision in this case is obviously equivalent to the length of the interval, and $\Pr(\mathcal{E}_V) + \Pr(\mathcal{E}_{VI}) = \Pr(\eta \in \overline{(\widehat{\eta}_{ln}, \widehat{\eta}_u)} = 1 - \Pr(\mathcal{E}_I)$ with η now an omitted singleton.

When the ATE is only partially identified, however, of the six cases depicted in Figure 1 only the event $\mathcal{E}_I$ is considered when evaluating coverage probability. For event $\mathcal{E}_I$ the outcome $[\eta_l, \eta_u] \subseteq [\widehat{\eta}_{ln}, \widehat{\eta}_{un}]$ occurs and the covering event can only give rise to correct and false coverage. Any features that might arise from a consideration of omitted coverage, or a combination of both false and omitted coverage, are neglected, and neither accuracy nor precision - in the every-day sense - are accounted for if $\mathcal{E}_I$ is considered in isolation. The complement of $\mathcal{E}_I$ is comprised of five alternative mutually exclusive events. As depicted in Figure 1, of these complementary categories, the encompassed event $\mathcal{E}_{II}$ gives rise to omitted coverage, the upper and lower intersecting events $\mathcal{E}_a$, $a = III, IV$, give rise to correct, false and omitted coverage, and the upper and lower disjoint events $\mathcal{E}_a$, $a = V, VI$, give rise to false and omitted coverage. Each event obviously captures a different aspect of, and makes a different contribution to, the performance characteristics of $[\widehat{\eta}_{ln}, \widehat{\eta}_{un}]$ as a confidence region for $[\eta_l, \eta_u]$.

To accommodate all six possibilities, let us consider constructing two error measures, $\mu\left\{[\eta_l, \eta_u] \cap \overline{[\widehat{\eta}_{ln}, \widehat{\eta}_{un}]}\right\}$ for omitted coverage, and $\mu\left\{\overline{[\eta_l, \eta_u]} \cap [\widehat{\eta}_{ln}, \widehat{\eta}_{un}]\right\}$ for false coverage, where $\mu\{\cdot\}$ denotes Lebesgue measure on the real line. Broadly speaking, the omitted coverage and false coverage regions relate to Type I and Type II errors respectively, and we designate the measures as Type I discrepancy and Type II discrepancy respectively. So our discrepancy measures become

$$\Delta_I = \mu\left\{[\eta_l, \eta_u] \cap \overline{[\widehat{\eta}_{ln}, \widehat{\eta}_{un}]}\right\} \quad \text{and} \quad \Delta_{II} = \mu\left\{\overline{[\eta_l, \eta_u]} \cap (\widehat{\eta}_{ln}, \widehat{\eta}_{un})\right\}.$$

Since the omitted coverage and false coverage regions are disjoint, $\Delta = \Delta_I + \Delta_{II}$ yields a measure of the extent of the discrepancy between $[\widehat{\eta}_{ln}, \widehat{\eta}_{un}]$ and $[\eta_l, \eta_u]$. The

overall measure Δ can be viewed as an interval counterpart to conventional point wise precision. We summarise overall performance via the expected discrepancy $E[\Delta] = \sum_{a=I}^{VI} \{\Pr((\widehat{\eta}_{ln}, \widehat{\eta}_{un}) \in \mathcal{E}_a) \cdot E[\Delta|\mathcal{E}_a]\}$, the weighted sum of the partial average of the discrepancy for each of the event categories, where the conditional expectations $E[\Delta|\mathcal{E}_a]$, $a = I, II, II, IV, V, VI$, are taken with respect to the sampling distribution of $[\widehat{\eta}_{ln}, \widehat{\eta}_{un}]$. We also decompose the overall expected discrepancy into Type I and Type II expected discrepancies as $E[\Delta] = E[\Delta_I] + E[\Delta_{II}]$. Each of these provides a natural measure of the contribution of the different event categories and error types to performance.[2]

5 Experimental Illustrations

In this section we provide simulation results that illustrate the finite sample properties of the proposed MHD bootstrap confidence region in scenarios that provide confirmation of concept. We also compare the performance of the MHD confidence region with that of CLR, using the performance measures outlined in Section 4. To ensure that regularity conditions required for the confidence regions to be valid are satisfied, we have generated data from bivariate probit (BVP) models and constructed the CLR and MHD confidence regions using the correctly specified BVP MLE estimates in the bounding functions in Equations (4) and (5).

The BVP model can be expressed in the form of a linear additive latent structure where

$$Y = \mathbf{1}\{\mathbf{X}\boldsymbol{\beta} + \gamma D + \varepsilon_Y > 0\}, \quad D = \mathbf{1}\{\mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\boldsymbol{\delta} + \varepsilon_D > 0\}, \tag{7}$$

and the joint distribution function of ε_D and ε_Y is $\Phi_2(\varepsilon_D, \varepsilon_Y; \rho)$, where $\Phi_2(\cdot, \cdot; \rho)$ denotes the cumulative distribution function of a bivariate standard normal distribution with correlation ρ. That this specification falls within the ambit of the Chesher structural equation model in Section 2 follows by setting the random variable $U = \Phi(\varepsilon_Y)$, where $\Phi(\cdot)$ is the scalar standard normal distribution function, and $h(D, \mathbf{X}, U) = \mathbf{1}\{U > \Phi(-\mathbf{X}\boldsymbol{\beta} - D\,\gamma)\}$. Then $h(d, \mathbf{x}, u)$ defines a structural function that is weakly monotonic in u where $P(U \leq u|\mathbf{Z} = \mathbf{z}) = u$ for all $u \in (0, 1)$ and all $\mathbf{z} \in \Omega_{\mathbf{Z}}$ since by assumption U is independent of $\mathbf{Z}$. This implies that the BVP model is equivalent to a structure, $\mathfrak{S}_{\text{BVP}}$ say, where Y can be expressed as in Equation (3) with a threshold function given by $p(d, \mathbf{x}) = 1 - \Phi(\mathbf{x}'\boldsymbol{\beta} + d\,\gamma)$.

[2] Adding the two types of discrepancy together implies they should be equally weighted. In particular applications the loss associated with the six event categories and the different types of discrepancy may be very different, and different weights can be assigned to discrepancies from under-estimated and over-estimated ATE values if desired. For example, in program evaluation it may be more important to estimate the minimum program benefit with a one-sided confidence region.

To give some guidance on likely empirical performance, the experimental DGP was designed to give control over features such as the degree of treatment endogeneity and instrument strength, thereby facilitating the consideration of different scenarios likely to be encountered in practice. The linear index of exogenous covariates $\mathbf{X}\boldsymbol{\beta} = \sum_{k=1}^{v} X_k\beta_k$ was summarised as $X\beta$ where X possesses a hypergeometric distribution

$$Pr(X = x) = \binom{M}{x}\binom{N-M}{K-x}/\binom{N}{K}, x = 0, 1, \ldots, M,$$

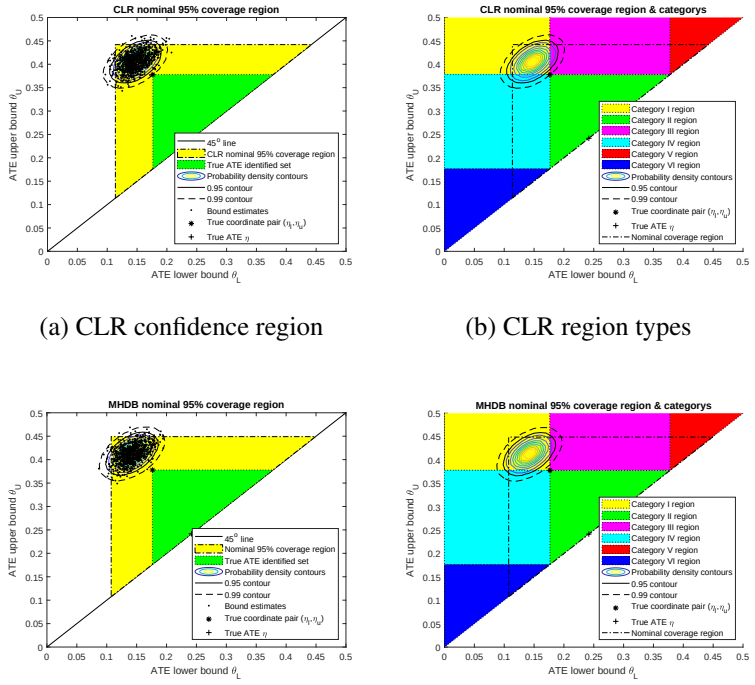
with parameters $K = 8$, $N = 20$ and $M = 4$. The instrumental variable $\mathbf{Z} = (Z_1, Z_2)'$ where $Z_1 \in \{0, 1\}$ with $\Pr(Z_1 = 0) = 0.5$ and $Z_2 \in \{-3, -2, -1, 0, 1, 2, 3\}$ with probabilities $(0.1, 0.1, 0.2, 0.2, 0.2, 0.1, 0.1)$. The variables Z_1, Z_2 and X are mutually independent. In addition $(X, Z_1, Z_2) \perp (\varepsilon_Y, \varepsilon_D)$.

The BVP model parameter vector $\boldsymbol{\theta} = (\beta, \gamma, \alpha, \boldsymbol{\delta}', \rho)'$ was set so that $\gamma = 1$ and $\alpha = 0$ across all other experimental settings. For this specification variations in X exert their effects on the ATE bounds through the value of β relative to γ. Since γ is held fixed we selected β from the set $\{0.05, 0.25, 0.45, 0.65\}$ so that, given the distribution of X, changes in β capture variations in the exogenous covariate signal-noise ratio in the outcome equation. Given the distribution of Z_1 and Z_2, there is a direct mapping from the coefficients of the instruments δ_1 and δ_2 to the instrument strength and identification power and we generated changes in these using the parameter grid $(\delta_1, \delta_2) = \delta(0.5, 0.2)$, $\delta = -4 : 0.2 : 4$. Different levels of endogeneity were explored using the grid $\rho = -0.9 : 0.1 : 0.9$. We have conducted a range of experiments with different parameter value and sample size combinations, but due to space limitations and for purposes of clarity we only present detailed results for settings that demonstrate particular features.

To provide a basis for comparison, in the illustrations that follow we have also constructed the pth-quantile precision-corrected CLR interval $[\eta_{ln}^{(\widehat{\boldsymbol{\theta}}_n)}(p), \eta_{un}^{(\widehat{\boldsymbol{\theta}}_n)}(p)]$ with $p = (1-\alpha)/2$. This was computed according to the "preferred" simulation-based method described in Algorithm 1 of Chernozhukov et al. (2013, p.708). The main input into this algorithm is a standardized statistic that can be approximated by a standard normal random variable and simulated for inference. This provides a confidence region for the ATE intersection bounds with significance level $1 - \alpha$ that, subject to the same regularity conditions as apply to MHD, satisfies Equation (2) (see Chernozhukov et al., 2013, Section 4). We will label this confidence region $CLR_{(1-\alpha)}^{(\widehat{\boldsymbol{\theta}}_n)}$. The experimental results reported here are based on $CLR_{0.95}^{(\widehat{\boldsymbol{\theta}}_n)}$ and $MHD_{0.95}^{(\widehat{\boldsymbol{\theta}}_n)}$, which henceforth we will refer to simply as CLR and MHD to aid legibility.

Figure 2 presents results from $R = 1000$ replications of a BVP DGP with parameter

values of $\beta = 0.25$, $(\delta_1, \delta_2) = (0.5, 0.2)$ and $\rho = 0.3$ when $n = 2750$ and $B = 1000$. This represents a setting involving coefficient values that yield moderate levels of the exogenous covariate signal-noise ratio, instrument strength, and degree of endogeneity. Figure 2 depicts the outcomes obtained when $X = 2$, it's modal value. The left hand



(a) CLR confidence region

(b) CLR region types

(c) MHD confidence region

(d) MHD region types

Figure 2. Confidence region nominal coverage regions (left hand panels) and event categories (right hand panels). Scatter plots of $[\widehat{\eta}_{ln}, \widehat{\eta}_{un}]$ in $R = 1000$ replications of samples of size $n = 2750$ from BVP DGP and BVP MLE, with $\beta = 0.25$, $(\delta_1, \delta_2) = (0.5, 0.2)$ and $\rho = 0.3$. Distribution of $[\widehat{\eta}_{ln}, \widehat{\eta}_{un}]$ for CLR and MHD with $B = 1000$ bootstrap redraws across event categories with $\eta = 0.2417$ denoted by an asterisk and $(\eta_l, \eta_u) = (0.1764, 0.3779)$ by a plus sign.

panels of Figure 2 plot the simulated coordinate pairs $[\widehat{\eta}_{ln}, \widehat{\eta}_{un}]$ of CLR and MHD in the subset of the ATE identified set parameter space given by the upper triangle of the quarter square $[0, 0.5] \times [0, 0.5]$. The parameter space of the ATE identified set corresponds to the set of coordinate pairs (η_l, η_u) that lie within the triangle $\{\eta_l : -1 \leq \eta_l \leq 1\} \times \{\eta_u : \eta_l \leq \eta_u \leq 1\}$ in $\mathbb{R}^2$. The coordinate pairs $(\widehat{\eta}_{ln}, \widehat{\eta}_{un})$ overlay the contours of a bi-variate normal distribution with mean vector and variance-covariance matrix set equal to the maximum

likelihood estimates $\left[\overline{\widehat{\eta}}_{ln}, \overline{\widehat{\eta}}_{un}\right]$ and

$$\begin{bmatrix} \widehat{\sigma}^2_{ln} & \widehat{\rho}_{lun}\widehat{\sigma}_{ln}\widehat{\sigma}_{un} \\ \widehat{\rho}_{lun}\widehat{\sigma}_{un}\widehat{\sigma}_{ln} & \widehat{\sigma}^2_{un} \end{bmatrix}$$

calculated from the simulated coordinate pairs $(\widehat{\eta}_{ln}, \widehat{\eta}_{un})$. These in turn overlay the nominal coverage regions of CLR and MHD shaded in yellow. The nominal $(1-\alpha)100\%$ coverage region of each confidence region was calculated as those coordinate pairs such that the interval $(\eta_l, \eta_u) \subseteq [c_{ln(1-\alpha)}, c_{un(1-\alpha)}]$ where $c_{ln(1-\alpha)} = \overline{\widehat{\eta}_{ln}} - \Phi^{-1}(1-\alpha/2)\widehat{\sigma}_{ln}$ and $c_{un(1-\alpha)} = \overline{\widehat{\eta}_{un}} + \Phi^{-1}(1-\alpha/2)\widehat{\sigma}_{un}$. The nominal coverage region represents those coordinate pairs (η_l, η_u) closest to the line $\eta_l = \eta_u$ that are covered by the distribution of $[\widehat{\eta}_{ln}, \widehat{\eta}_{un}]$ with probability $(1-\alpha)$. The plus sign and asterisk denote the 'true' ATE BVP DGP point η and identified set bound coordinates (η_l, η_u) respectively. The right hand panels of Figure 2 graph the contours of the maximum likelihood bi-variate normal distributions of CLR and MHD in the same space as in the left hand panels, with areas corresponding to the six events $\mathcal{E}_a$, $a = I, II, II, IV, V, VI$, highlighted in colour, and with the CLR and MHD nominal coverage regions outlined.

A casual perusal of Figure 2 would suggest that the distribution of the CLR and MHD confidence regions are very similar, an impression that is reinforced by noting that the nominal coverage region encompasses the true identified set. From Figure 2(a) and Figure 2(c) it is apparent that, bar event categories $\mathcal{E}_V$ and $\mathcal{E}_{VI}$ where $\Pr(\mathcal{E}_V) \approx \Pr(\mathcal{E}_{VI}) \approx 0$, for any $\eta \in (\eta_{l0}, \eta_{u0})$ there exist confidence intervals $[\widehat{\eta}_{ln}, \widehat{\eta}_{un}]$ in the nominal coverage region such that $\eta \in [\widehat{\eta}_{ln}, \widehat{\eta}_{un}]$. But $(\eta_{l0}, \eta_{u0}) \subseteq [\widehat{\eta}_{ln}, \widehat{\eta}_{un}]$ only for those $[\widehat{\eta}_{ln}, \widehat{\eta}_{un}]$ that fall in event category $\mathcal{E}_I$, as is made clear in Figure 2(b) and Figure 2(d).

The visual similarities seen in Figure 2 disguise some important numerical differences. Noting that $[\eta_l, \eta_u] = [0.1764, 0.3779]$, for CLR $\left[\overline{\widehat{\eta}}_{ln}, \overline{\widehat{\eta}}_{un}\right] = [0.1478, 0.4056]$ and

$$\begin{bmatrix} \widehat{\sigma}^2_{ln} & \widehat{\rho}_{lun}(\widehat{\sigma}_{ln}\widehat{\sigma}_{un}) \\ \widehat{\rho}_{lun}(\widehat{\sigma}_{un}\widehat{\sigma}_{ln}) & \widehat{\sigma}^2_{un} \end{bmatrix} = \begin{bmatrix} 0.3228 \cdot 10^{-3} & 0.4726(0.3236 \cdot 10^{-3}) \\ 0.4726(0.3236 \cdot 10^{-3}) & 0.3459 \cdot 10^{-3} \end{bmatrix}$$

whereas for MHD $\left[\overline{\widehat{\eta}}_{ln}, \overline{\widehat{\eta}}_{un}\right] = [0.1419, 0.4127]$ and

$$\begin{bmatrix} \widehat{\sigma}^2_{ln} & \widehat{\rho}_{lun}(\widehat{\sigma}_{ln}\widehat{\sigma}_{un}) \\ \widehat{\rho}_{lun}(\widehat{\sigma}_{un}\widehat{\sigma}_{ln}) & \widehat{\sigma}^2_{un} \end{bmatrix} = \begin{bmatrix} 0.3098 \cdot 10^{-3} & 0.4746(0.3274 \cdot 10^{-3}) \\ 0.4746(0.3274 \cdot 10^{-3}) & 0.3460 \cdot 10^{-3} \end{bmatrix}.$$

These differences generate a distribution of $[\widehat{\eta}_{ln}, \widehat{\eta}_{un}]$ for MHD that is more concentrated on the covering event $\mathcal{E}_I$ than is the distribution of $[\widehat{\eta}_{ln}, \widehat{\eta}_{un}]$ for CLR, as is apparent from a closer inspection of Figure 2(b) and Figure 2(d). It can be seen that CLR is less focused on

$\mathcal{E}_I$ events than is MHD, with more values dispersed over categories $\mathcal{E}_{III}$ and $\mathcal{E}_{IV}$. These features manifest numerically in 93.6% of MHD values $[\widehat{\eta}_{ln}, \widehat{\eta}_{un}]$ falling in category $\mathcal{E}_I$ whilst only 86.6% of CLR values $[\widehat{\eta}_{ln}, \widehat{\eta}_{un}]$ fall in category $\mathcal{E}_I$, with 6.2% in category $\mathcal{E}_{III}$ and 7.2% in category $\mathcal{E}_{IV}$.

The upshot is that the expected discrepancy of CLR equals $58.31280 \cdot 10^{-3}$ compared to $70.2243 \cdot 10^{-3}$ for MHD. The discrepancy is defined as in Section 4 and was calculated in the simulations as the average per replication for each of the different event types. In line with the development in Section 4, we denote this by $E_R[\Delta] = E_R[\Delta_I] + E_R[\Delta_{II}]$. The overall value $E_R[\Delta]$ is made up of $E_R[\Delta_I] = 1.0036 \cdot 10^{-3}$ and $E_R[\Delta_{II}] = 57.3092 \cdot 10^{-3}$ for CLR, and $E_R[\Delta_I] = 0.4576 \cdot 10^{-3}$ and $E_R[\Delta_{II}] = 69.7667 \cdot 10^{-3}$ for MHD. These figures indicate that the Type I discrepancy of CLR is just over twice that of MHD, whereas the Type II discrepancy of CLR is a little under 83% of the Type II discrepancy of MHD, discrepancy values that reflect the subtle differences in the distributions of $[\widehat{\eta}_{ln}, \widehat{\eta}_{un}]$ over the six events $\mathcal{E}_a$, $a = I, II, II, IV, V, VI$.

Similar features to those seen in Figure 2 were observed for other values of the covariate, and the performance characteristics observed across the distribution of X are summarised numerically in Table 1. This table presents the observed coverage probability

Table 1. Coverage probability and discrepancy measures across $R = 1000$ replications of samples of size $n = 2750$ from BVP DGP and BVP MLE with $\beta = 0.25$, $(\delta_1, \delta_2) = (0.5, 0.2)$ and $\rho = 0.3$. Confidence region performance of CLR and MHD with $B = 1000$ bootstrap redraws.

Covariate Values		Coverage Probability		Omitted Coverage $100 * E_R[\Delta_I]/(\eta_u - \eta_l)$		False Coverage $100 * E_R[\Delta_{II}]/(\eta_u - \eta_l)$	
X	Pr	CLR	MHD	CLR(%)	MHD(%)	CLR(%)	MHD(%)
0	0.102	0.905	0.929	0.326	0.203	29.018	30.010
1	0.363	0.872	0.934	0.440	0.200	26.463	30.275
2	0.381	0.908	0.936	0.326	0.227	31.807	34.633
3	0.139	0.964	0.945	0.180	0.262	43.369	41.054
4	0.014	0.984	0.954	0.063	0.237	60.290	46.578
Average		0.903	0.936	0.343	0.220	31.596	33.641

and the observed Type I and Type II discrepancies. To maintain compatibility across the different X values and help gauge variations in performance the Type I and Type II discrepancies are expressed as a percentage of the true identified set, $E_R[\Delta_I]/(\eta_u - \eta_l)$ and $E_R[\Delta_{II}]/(\eta_u - \eta_l)$ times 100% respectively. The column averages were evaluated as weighted averages using the covariate probabilities as weights. The outcomes sum-

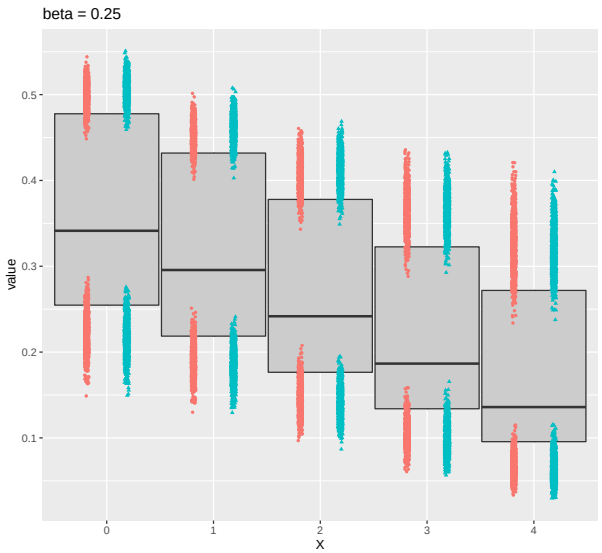


Figure 3. ATE and identified set for BVP DGP with $\beta = 0.25$, $(\delta_1, \delta_2) = (0.5, 0.2)$ and $\rho = 0.3$. Scatter plots of confidence region limits $\widehat{\eta}_{ln}$ and $\widehat{\eta}_{un}$ for CLR (Orange) and MHD (Cyan) with $B = 1000$ across the support of X, with $R = 1000$ replications of samples of size $n = 2750$.

marised in Table 1 are displayed in Figure 3. Figure 3 graphs η_l and η_u and the identified set $[\eta_l, \eta_u]$ on the vertical axis, depicted as a grey rectangle for each X, with the DGP ATE η and the confidence region lower and upper limits $\widehat{\eta}_{ln}$ and $\widehat{\eta}_{un}$ superimposed, the former as a horizontal line and the latter as scatter plots.

The first obvious feature apparent from Table 1 is that the coverage probability of MHD is more uniform across the different covariate values than that of CLR, and on average the coverage probability of MHD exceeds the coverage probability of CLR with an attendent Type I discrepancy that is about two-thirds of that of CLR. The second most notable feature is that for both confidence regions Type II discrepancies are the predominant component of $E[\Delta]$, which is perhaps not too surprising given that both MHD and CLR are designed to satisfy Equation (2). The average Type II discrepancy of CLR is roughly 94% of that of MHD, but this reduction in false coverage relative to MHD appears to have been bought at the cost of a significant reduction in coverage probability for the smaller values of the covariate. These features are also reflected in Figure 3.

Table 2 presents the average coverage probability and discrepancy measures from the BVP DGP with $\beta = 0.25$ and $(\delta_1, \delta_2) = (0.5, 0.2)$, as in Table 1, but with different combinations of sample size (n), degree of endogeneity (ρ) and number of bootstrap resamples (B). Across all 24 scenarios listed in Table 2 the coverage probability of CLR exceeded that of MHD on only 7 occasions. Virtually all of those occasions happened

Table 2. Average coverage probability and discrepancy types from BVP DGP with $\beta = 0.25$ and $(\delta_1, \delta_2) = (0.5, 0.2)$ across $R = 1000$ replications. Confidence region performance of CLR and MHD with combinations of sample size n, latent error correlation ρ and bootstrap redraws B.

			Coverage Probability		Type I Discrepancy Omitted Coverage		Type II Discrepancy False Coverage	
n	ρ	B	CLR	MHD	CLR(%)	MHD(%)	CLR(%)	MHD(%)
1000	-0.8	1000	0.986	0.966	0.153	0.343	379.467	173.867
		3000	0.986	0.982	0.153	0.176	379.467	186.800
	-0.3	1000	0.961	0.958	0.292	0.329	96.076	90.560
		3000	0.961	0.971	0.292	0.205	96.076	97.624
	0.3	1000	0.957	0.965	0.210	0.195	64.195	63.342
		3000	0.957	0.977	0.210	0.124	64.195	68.482
	0.8	1000	0.953	0.939	0.210	0.295	51.264	46.038
		3000	0.953	0.957	0.210	0.204	51.264	50.105
2750	-0.8	1000	0.976	0.963	0.143	0.259	217.749	88.383
		3000	0.976	0.978	0.143	0.150	217.749	95.526
	-0.3	1000	0.941	0.965	0.286	0.169	50.529	53.633
		3000	0.941	0.975	0.286	0.111	50.529	57.870
	0.3	1000	0.903	0.936	0.343	0.220	31.596	33.641
		3000	0.903	0.962	0.343	0.128	31.596	36.661
	0.8	1000	0.910	0.931	0.236	0.177	26.949	25.765
		3000	0.910	0.954	0.236	0.103	26.949	28.055
5000	-0.8	1000	0.975	0.959	0.089	0.262	162.170	65.971
		3000	0.975	0.971	0.089	0.193	162.170	71.200
	-0.3	1000	0.937	0.964	0.234	0.140	34.103	38.374
		3000	0.937	0.976	0.234	0.086	34.103	41.492
	0.3	1000	0.886	0.927	0.339	0.182	21.490	23.993
		3000	0.886	0.948	0.339	0.113	21.490	26.111
	0.8	1000	0.886	0.908	0.258	0.200	18.697	18.150
		3000	0.886	0.934	0.258	0.130	18.697	19.821

when B was small and $\rho = -0.8$, or when n was small, and were accompanied by a Type II discrepancy for CLR that was more than twice that of MHD, whether or not the coverage probability fell short of the nominal confidence coefficient. By way of contrast, when the coverage probability of MHD exceeded that of CLR, the Type II discrepancy of MHD was never more than 2% larger that of CLR. For both MHD and CLR the Type

I discrepancy was always less than 0.35% and the ratio of Type I discrepancy to Type II discrepancy (the relative size of omitted coverage to false coverage) never exceeded 1.5%.

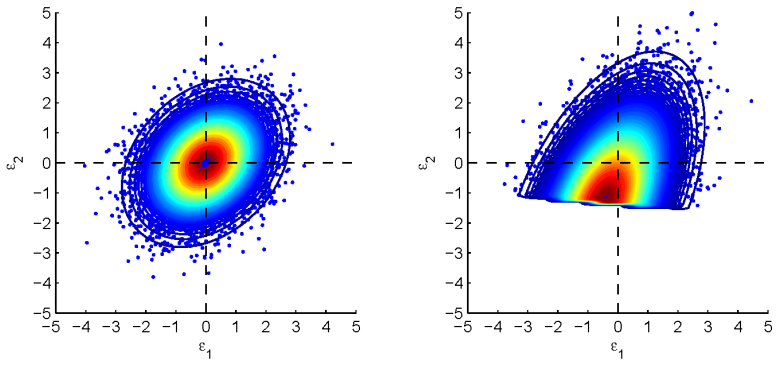
The general conclusion from Table 2 is that, all else being equal, increases in sample size improve Type II discrepancy, as would be expected, although improvements in coverage probability and Type I discrepancy are not so clear cut. Any difference in the performance of CLR and MHD appears to become smaller as n increases. Since for this DGP design the ATE is positive, performance is also better when ρ is negative (when true bounds are narrower) and improves as the degree of endogeneity increases as $\rho \to -1$. Increases in B has no impact on CLR of course, but increasing B increases the precision of the MHD confidence region and provides a seemingly costless practical way to improve MHD performance for a given sample size. In fact, MHD has better coverage probability than CLR for virtually all cases when $B = 3000$.

6 ATE Inference Using the BVP QMLE

The previous results are based on confidence regions constructed using correctly specified bounding functions, namely, those based on the BVP MLE calculated from data generated from a BVP DGP. As such they provide a guide to the type of outcomes likely to be encountered in the most favourable circumstances where the conditions for the validity of the confidence regions, as outlined in Section 3, are satisfied. In recognition that when an M-estimator (or parametric model) is used to estimate the bounds the structure of the DGP will not be known, we now investigate the properties of CLR and MHD confidence regions using a misspecified bound estimator.[3] Specifically, we construct bounds using a BVP QMLE fitted to data generated from bivariate skewed-normal (BVSN) random variables.

We first generate data from the model in Equation (7) when (following Li et al., 2019a) the latent error term $(\varepsilon_Y, \varepsilon_D)'$ is a BVSN random variable, standardized to have mean zero and variance one with correlation ρ. The BVSN is asymmetric and allows for heavy tails, and offers greater flexibility for creating misspecification in a BVP model and criterion function than considering alternative members of the spherically symmetric family such as Student-t. Figure 4 presents scatter plots of $n = 1000$ observations from $\Phi_2(\cdot, \cdot; \rho)$ and the standardised BVSN distribution when $\rho = 0.3$. Each scatter plot is superimposed over the probability density contours of their respective distribution. This illustrates a

[3]Recall that Li et al. (2019b) found that data scarcity issues preventing the use of non-parametric estimates and necessitating the use of parametric models were present in a real world sample of over 19,500 records.



(a) Bivariate standard normal distribution

(b) Bivariate standardised skew-normal distribution

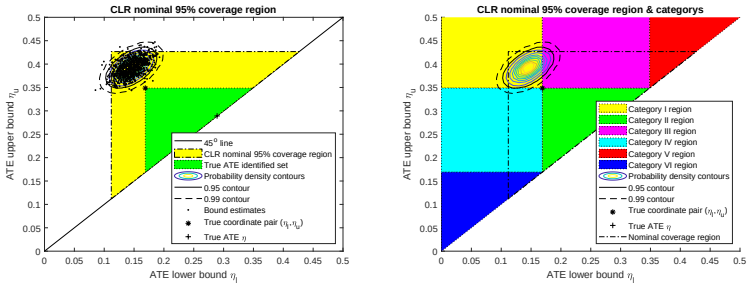
Figure 4. Theoretical contours and scatter plots from bivariate normal and skew-normal latent errors with zero means, unit variances and correlation coefficient $\rho = 0.3$.

setting in which the skewness and excess kurtosis of the BVSN marginal distributions is not excessive, 0.0392 and 0.0116 respectively, yet the truncation and distortion in the BVSN joint distribution relative to the bivariate standard normal of the BVP is extreme.

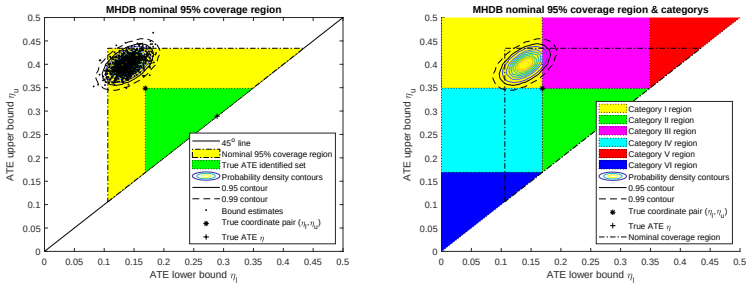
Figure 5 presents simulation results obtained from BVP QMLE based CLR and MHD confidence regions constructed using data derived from a BVSN DGP with parameter values of $\beta = 0.25$, $(\delta_1, \delta_2) = (0.5, 0.2)$ and $\rho = 0.3$ when $X = 2$. This represents a scenario with the same coefficient and variable values as employed previously in Section 5 but with misspecification of the BVP model used for estimating the bounds induced by the skewed-normal latent errors of the BVSN DGP. The outstanding feature of Figure 5 is its similarity to Figure 2, with all four panels of Figure 5 being visually very similar to those of Figure 2. The correspondence between the patterns of behaviour seen in Figure 5 and Figure 2 is striking, particularly as the difference in the underlying latent error distributions in the two settings is far from trivial, as seen in Figure 4.

That ATE identified set confidence regions calculated using the BVP QMLE can exhibit acceptable performance characteristics, as seen here, accords with the results of Li et al. (2018, 2019a), who found that the BVP QMLE could yield acceptable mean squared error performance across a range of DGPs. The observation of these features provides motivation for examining properties of the BVP QMLE that might provide an explanation for such outcomes.

Set $t_1 = (2\mathrm{y}-1)(\mathbf{x}'\boldsymbol{\beta}+\mathrm{d}\,\gamma)$, $t_2 = (2\mathrm{d}-1)(\mathbf{x}'\boldsymbol{\alpha}+\mathbf{z}'\boldsymbol{\delta})$ and $\varrho = (2\mathrm{y}-1)(2\mathrm{d}-1)\rho$. Then the conditional distribution of (Y, D) given $\mathbf{X}$ and $\mathbf{Z}$ for the BVP model in Equation (7) can



(a) CLR confidence region (b) CLR region types



(c) MHD confidence region (d) MHD region types

Figure 5. Confidence region nominal coverage regions (left hand panels) and event categories (right hand panels). Scatter plots of $[\widehat{\eta}_{ln}, \widehat{\eta}_{un}]$ in $R = 1000$ replications of samples of size $n = 2750$ from BVSN DGP with $\beta = 0.25$, $(\delta_1, \delta_2) = (0.5, 0.2)$ and $\rho = 0.3$. Distribution of $[\widehat{\eta}_{ln}, \widehat{\eta}_{un}]$ for CLR and MHD with $B = 3000$ bootstrap redraws across event categories with $\eta = 0.2891$ denoted by an asterisk and $(\eta_l, \eta_u) = (0.1689, 0.3486)$ by a plus sign.

be expressed as $\Phi_2(t_1, t_2; \varrho)$. If we now let $(t_{1i}, t_{2i}, \varrho_i)$ denote the ith value generated by $\mathbf{w}_i$, $i = 1, \ldots, n$, the BVP log-likelihood function is $\log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \Phi_2(t_{1i}, t_{2i}; \varrho_i)$. Let $\mathbb{L} = \exp\{\sum_{i=1}^{N} \log \mathbb{P}_2(t_{1i}, t_{2i})\}$, where $\mathbb{P}_2(t_1, t_2)$ denotes the probability distribution determined by the DGP. Representing $p(t_1, t_2) = \partial^2 \mathbb{P}_2(t_1, t_2)/\partial t_1 \partial t_2$ in terms of a Gram-Charlier Type A expansion about

$$\phi_2(t_1, t_2; \varrho) = \frac{1}{2\pi(1-\varrho^2)^{\frac{1}{2}}} \exp\left\{-\frac{(t_1^2 - 2\varrho t_i t_2 + t_2^2)}{2(1-\varrho^2)}\right\},$$

and substituting into the Kullback-Leibler divergence $K_n(\boldsymbol{\theta}) = \mathbb{E}[\log \mathbb{L} - \log L(\boldsymbol{\theta})]$ yields the expressions

$$p(t_1, t_2) = \{1 + \sum_{r+s \geq 3} c_{rs} H_{rs}(t_1, t_2, \varrho)\} \phi_2(t_1, t_2; \varrho) \tag{8}$$

and

$$K_n\left(\boldsymbol{\theta}\right) = \mathbb{E}\left[\sum_{i=1}^{n} \log\left\{1 - \frac{\phi_2\left(t_{1i}, t_{2i}; \varrho_i\right)}{\Phi_2(t_{1i}, t_{2i}; \varrho_i)} \sum_{r+s\geq 3} c_{rs} H_{(r-1)(s-1)}(t_{1i}, t_{2i}, \varrho_i)\right\}\right] \tag{9}$$

where the covariant Hermite polynomials are defined by the Rodriguez formula

$$H_{rs}(t_1, t_2, \varrho)\phi_2\left(t_1, t_2; \varrho\right) = (-1)^{r+s} \frac{\partial^r}{\partial t_1^r}\left\{\frac{\partial^s \phi_2\left(t_1, t_2; \varrho\right)}{\partial t_2^s}\right\},$$

and the coefficients are given by multiplying Equation (8) through by $H_{rs}(t_1, t_2, \varrho)$, integrating over $\mathbb{R}^2$, and exploiting the orthogonality of the Hermite polynomials, to give

$$c_{rs} = \frac{1}{(r+s)!} \int_{\mathbb{R}^2} H_{rs}(t_1, t_2, \varrho) p(t_1, t_2) dt_1 dt_2 \,.$$

From Equation (9) it is apparent that the ability of a BVP model to match the probabilities of the DGP depends on higher order moments than the first and second, and ρ of course. For detailed particulars of covariant Hermite polynomials and the Gram-Charlier expansion in the bivariate case see Barndorff-Nielsen and Pedersen (1979) and Stuart and Ord (1993, Exercise 7.21) respectively.

The two leading terms in the Gram-Charlier expansion in Equation (8) which depend on the third and fourth cumulant adjust for skewness and kurtosis. These terms are particularly sensitive to tail behaviour and this sensitivity is carried over into $K_n\left(\boldsymbol{\theta}\right)$ in Equation (9) via the term

$$\frac{\phi_2\left(t_{1i}, t_{2i}; \varrho_i\right)}{\Phi_2(t_{1i}, t_{2i}; \varrho_i)} \sum_{r+s\geq 3} c_{rs} H_{(r-1)(s-1)}(t_{1i}, t_{2i}, \varrho_i) \,.$$

Deviations of this term from zero obviously govern the extent to which $K_n\left(\boldsymbol{\theta}\right) \geq 0$ will depart from it's lower bound. From the Fréchet inequalities

$$\phi(t_1) + \phi(t_2) - 1 \leq \Phi_2(t_1, t_2; \varrho) \leq \min\{\phi(t_1), \phi(t_2)\} \,,$$

and the inequality $1 - \Phi(t) < \phi(t)/t$ for $t > 0$, it follows that the inverse Mill's ratio $\phi_2\left(t_1, t_2; \varrho\right)/\Phi_2(t_1, t_2; \varrho)$ converges to zero as $t_1 \to \infty$ or $t_2 \to \infty$ but diverges as $t_1 \to -\infty$ or $t_2 \to -\infty$. This ratio therefore either amplifies or attenuates the sensitivity of $K_n\left(\boldsymbol{\theta}\right)$ to moments greater than the first and second. Consequently, a fitted BVP QMLE model may be able to capture the properties of a DGP reasonably well whenever the DGP does not have levels of skewness and kurtosis, and higher order moments, that are too excessive and the inverse Mill's ratio is small.

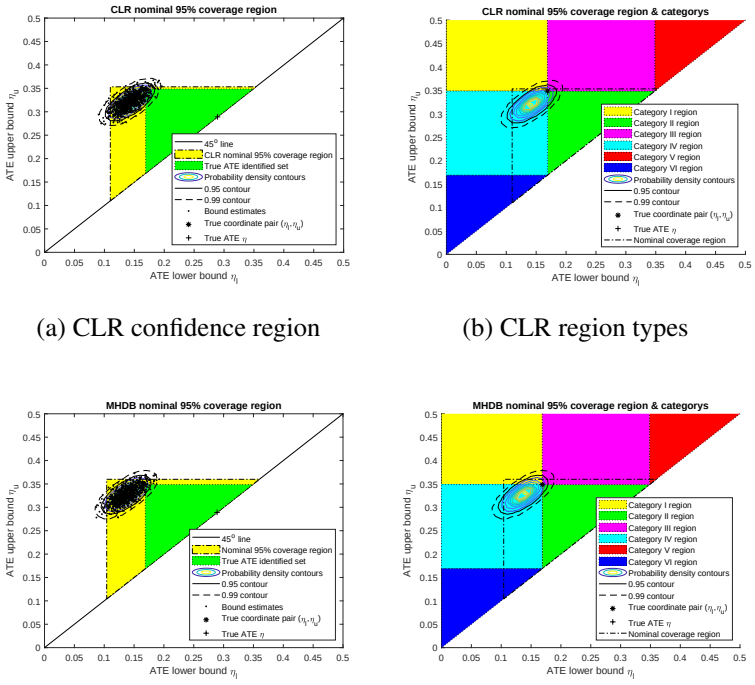
For example, the observation of behaviour in Figure 2 and Figure 5 that is not too dissimilar suggests that the nature of the current misspecification is not detrimental to the performance of the BVP QMLE despite the distributions of the latent errors in the BVP model and the BVSN DGP being very different. Evidence for this is provided in Table 3. Table 3 presents the average coverage probability and Type I and Type II discrepancies observed across the support of X in samples of size $n = 2750$ from the BVSN DGP when $\beta = 0.25$, $(\delta_1, \delta_2) = (0.5, 0.2)$, $\rho = \{-0.8, -0.3, 0.3, 0.8\}$ and $B = \{1000, 3000\}$. A comparison of the values given in Table 3 with those given previously in Table 2 for

Table 3. Average coverage probability and discrepancy types from BVSN DGP and BVP QMLE, with $\beta = 0.25$ and $(\delta_1, \delta_2) = (0.5, 0.2)$ across $R = 1000$ replications. Confidence region performance of CLR and MHD with combinations of sample size n, latent error correlation ρ and bootstrap redraws B.

			Coverage Probability		Type I Discrepancy Omitted Coverage		Type II Discrepancy False Coverage	
n	ρ	B	CLR	MHD	CLR(%)	MHD(%)	CLR(%)	MHD(%)
2750	-0.8	1000	0.000	0.000	58.555	57.103	52.362	20.769
		3000	0.000	0.000	58.555	55.717	52.362	22.014
	-0.3	1000	0.073	0.069	15.695	13.978	19.069	19.231
		3000	0.073	0.101	15.695	12.591	19.069	20.968
	0.3	1000	0.938	0.948	0.234	0.197	42.137	44.715
		3000	0.938	0.967	0.234	0.108	42.137	47.967
	0.8	1000	0.946	0.942	0.295	0.317	84.568	84.810
		3000	0.946	0.955	0.295	0.219	84.568	88.099

$n = 2750$ and $\rho = 0.3$ shows that although the Type II discrepancy (false coverage) has increased, as might have been anticipated, the coverage probability and the Type I discrepancy of the BVP QMLE in this setting are roughly on a par with those obtained in the correctly specified case.

A notable feature of Table 3, however, is that changing ρ from positive to negative has a serious deleterious effect on the performance characteristics of the BVP QMLE and produces an effect that is the opposite to that seen in the correctly specified case. For example, in the BVP DGP case the average coverage probability of CLR and MHD increased, respectively, from 0.861 and 0.954 when $\rho = 0.3$ to 0.915 and 0.975 when $\rho = -0.3$, whereas for the BVSN DGP the average coverage probability of CLR and MHD decreases from 0.9 and 0.967 when $\rho = 0.3$ to 0.038 and 0.101 when $\rho = -0.3$.



(a) CLR confidence region

(b) CLR region types

(c) MHD confidence region

(d) MHD region types

Figure 6. Confidence region nominal coverage regions (left hand panels) and event categories (right hand panels). Scatter plots of $[\widehat{\eta}_{ln}, \widehat{\eta}_{un}]$ in $R = 1000$ replications of samples of size $n = 2750$ from BVSN DGP and BVP QMLE estimates, with $\beta = 0.25$, $(\delta_1, \delta_2) = (0.5, 0.2)$ and $\rho = -0.3$. Distribution of $[\widehat{\eta}_{ln}, \widehat{\eta}_{un}]$ for CLR and MHD with $B = 3000$ bootstrap redraws across event categories with $\eta = 0.2891$ denoted by an asterisk and $(\eta_l, \eta_u) = (0.1689, 0.3486)$ by a plus sign.

To examine the latter phenomenon further, Figure 6 depicts the results obtained from the BVSN DGP when $\beta = 0.25$, $(\delta_1, \delta_2) = (0.5, 0.2)$, $\rho = -0.3$ and $X = 2$. The configuration seen in Figure 6 is very different from that seen previously in Figure 5. In Figure 6 the event categories $\mathcal{E}_I$ and $\mathcal{E}_{IV}$ occurred 2.2% and 91.8% of the time for CLR and 5.7% and 89.3% of the time for MHD, and the shortfall in coverage probability for the true identified set fairly obviously arises from the prevalence of $\mathcal{E}_{IV}$ events. Since the event category $\mathcal{E}_{VI}$ does not occur, it follows that the lower bound of the true identified set falls in the CLR confidence region 94% of the time and in the MHD confidence region 95% of the time. The upper bound of the true identified set fell in the confidence regions only 4.7% and 10% respectively. This leads to Type I discrepancies (omitted coverage) for both confidence regions that are an order of magnitude larger than seen previously when $\beta = 0.25$, $(\delta_1, \delta_2) = (0.5, 0.2)$ and $\rho = 0.3$, but Type II discrepancies (false coverage), on

the other hand, that are less than half than those seen previously.

Although the coverage probability and Type I and Type II discrepancies obtained across the support of X when $\beta = 0.25$, $(\delta_1, \delta_2) = (0.5, 0.2)$ and $\rho = 0.3$, and $\beta = 0.25$, $(\delta_1, \delta_2) = (0.5, 0.2)$ and $\rho = -0.3$, were not uniform, the qualitative characteristics outlined here for $X = 2$ were observed with other values of the covariate. The upshot is that although the average coverage probability when $\rho = -0.3$ has collapsed, the average discrepancy of CLR has decreased from 42.37% when $\rho = 0.3$ to 34.76% when $\rho = -0.3$, and the average discrepancy of MHD has decreased from 44.91% to 33.20%. Interestingly enough, similar comparisons were observed when $\beta = 0.25$, $(\delta_1, \delta_2) = (0.5, 0.2)$ and $\rho = 0.8$ and $\beta = 0.25$, $(\delta_1, \delta_2) = (0.5, 0.2)$ and $\rho = -0.8$, only now the coverage probability fails in the negative case having achieved the nominal confidence coefficient value in the positive case, and the average discrepancy increases from 84.86% when $\rho = 0.8$ to 110.92% when $\rho = -0.8$ for CLR, and decreases from 88.32% when $\rho = 0.8$ to 77.73% when $\rho = -0.8$ for MHD. For both CLR and MHD the balance between Type I and Type II discrepancies is reversed, with Type II discrepancies dominating for positive ρ and Type I discrepancies dominating for negative ρ.

The conclusions from the discussion surrounding the results in Table 2 and Table 3 and the illustrations in Figure 2, Figure 5 and Figure 6 are that the condition that $\mathfrak{S}_{\text{BVP}}$ should approximate $\mathfrak{S}$ required for the results in Section 3.2 to hold can be satisfied even in situations where the misspecification in the BVP model underlying the BVP QMLE is quite extreme. In such cases the behaviour of CLR and MHD confidence regions based on the BVP QMLE conforms with Equation (2). Nevertheless, practitioners will need to be wary of presupposing that the process giving rise to their data satisfies the requirements for the BVP QMLE to work well since the conditions under which the performance of BVP QMLE based confidence regions are acceptable appear to be fragile and their breakdown can have dire consequences. The latter problem might be addressed by considering a robust version of the BVP QMLE, following Huber (1967), but an exploration of that possibility would take us too far afield here and how best to estimate a confidence region in practice when non-parametric estimation is impracticable due to data scarcity remains an unresolved issue.

7 Summary

In this paper we have developed an empirical bootstrap approach to constructing a minimum Hausdorff distance bootstrap confidence region for the ATE identified set (labelled MHD). The MHD confidence region was shown to possess a prescribed asymptotic cov-

erage probability, under suitable regularity. A characterisation of the discrepancy between a confidence region and its target identified set was also introduced. The characterisation was designed to elucidate features of a confidence region for an identified set that are not catered for by the conventional confidence interval considerations of accuracy and precision. Six possible outcomes were delineated and associated measures of discrepancy were introduced, including a Type I discrepancy for omitted coverage and a Type II discrepancy for false coverage, and an overall discrepancy.

The finite sample properties of MHD confidence regions was illustrated using Monte Carlo experiments. The MHD confidence region was compared to what might be regarded as the current state of the art method, the CLR confidence regions proposed in Chernozhukov et al. (2013). Coverage probability and the discrepancy measures were used to explore the finite sample behaviour of the two confidence regions. Overall the performance of the MHD confidence regions compared favourably with that of the CLR confidence regions.

In light of the issues raised in empirical studies where non-parametric estimates for bounding functions are not feasible due to data paucity, the use of the BVP QMLE for estimating ATE bounds was also examined. It was shown that both CLR and MHD confidence regions obtained using misspecified BVP QMLE estimates can perform very well or very poorly. As part of this exercise a Gram-Charlier expansion of the DGP distribution about a BVP model was explored, revealing hitherto undocumented features that underlie the performance characteristics of the BVP QMLE. These features provide an explanation of the experimental outcomes that were observed here, not only those that accord with the positive findings previously documented in Li et al. (2018, 2019a), but also those indicating that ATE inference based on the BVP QMLE can exhibit extremely poor finite sample performance characteristics.

Finally, there are three obvious extensions of the ideas presented in this paper that warrant investigation. First, we have structured our analysis around the ATE intersection bounds of Chesher, but MHD confidence regions can be used to conduct inference on other treatment effect bounds that have appeared in the literature. Second, validity of MHD confidence regions was established under sufficient generality to allow for implementation using estimators other than the empirical distribution function, such as nonparametric, semi-parametric or M-estimators (other than the BVP QMLE examined here). Third, the use of minimum Hausdorff distance in the construction of the MHD confidence region implicitly weights Type I and Type II discrepancies evenly; different weights for the two types of discrepancy could be used to allow for asymmetric loss functions, or to construct one-sided confidence regions. These extensions offer interesting avenues for further re-

search.

A Supplementary Appendix: Proofs

Proof of Theorem 1 The empirical distribution $\mathbb{F}_n$ as a real-valued (measurable) function with overall domain $\Omega_{(Y,D)}$ times $\Omega_{\mathbf{X}} \times \Omega_{\mathbf{Z}}$ and $\mathbb{G}_n = \sqrt{n}(\mathbb{F}_n - F)$ constitutes an empirical process indexed by the Vapnik-Červonenkis class of functions $\{\mathbf{1}\{(-\infty, y]\}\mathbf{1}\{(-\infty, d]\}\delta_{\mathbf{x}}\delta_{\mathbf{z}}\}$. It follows that $\mathbb{G}_n$ converges weakly to $\mathbb{G} \circ F$, an F-Brownian bridge in $\ell^\infty[0,1]$ whose induced probability distribution is tight. By Theorem 3.9.11 of Van der Vaart and Wellner (1996), for every Hadamard-differentiable function ψ the conditional distribution of $\sqrt{n}(\psi(\mathbb{F}_n^*) - \psi(\mathbb{F}_n))$ converges to the same limit as $\sqrt{n}(\psi(\mathbb{F}_n) - \psi(F))$ and yields an asymptotically weakly consistent estimate of the latter's limiting law.

Now let Ψ denote the probability distribution function of the common limiting random variable T of $T_n = \sqrt{n}(\psi(\mathbb{F}_n) - \psi(F))$ and $T_n^* = \sqrt{n}(\psi(F_n^*) - \psi(\mathbb{F}_n))$, and for $0 < p < 1$ set the inverse map $\Psi^{-1}(p) = \inf\{\|T\| : \Psi(T) \geq p\}$. If Ψ_n^* is the probability distribution function of T_n^* then by Lemma 3.9.20 of Van der Vaart and Wellner (1996) $\Psi_n^{*-1}(p)$ converges to $\Psi^{-1}(p)$, and hence $\Pr(\|T_n\| \leq \Psi_n^{*-1}(1-\alpha))$ converges to $\Pr(\|T\| \leq \Psi^{-1}(1-\alpha)) = 1 - \alpha$.

To complete the proof, rewrite $\psi(F)$ as the composition $\pi \circ \varphi(F) = \pi(\varphi(F))$ where $\varphi(F)$ denotes the mapping from F to the coordinate pair (η_l, η_u) corresponding to the ATE lower and upper bounds, and π is the projection that maps members of $\{(\eta_l, \eta_u) : -1 \leq \eta_l \leq 1, \eta_l \leq \eta_u \leq 1\} \subset \mathbb{R}^2$ into intervals $[\eta_l, \eta_u] \subseteq [-1, 1]$. The function $\varphi(\cdot)$ yields a co-ordinate pair constructed from the composition of the supremum and infimum operators applied to linear functions of F. The norm $\|\varphi(F)\| = \sqrt{|\eta_l|^2 + |\eta_u|^2}$ is uniformly bounded by $\sqrt{2}$, and straightforward if somewhat tedious manipulations using the properties of the supremum and infimum operators shows that

$$\|\varphi(F_1) - \varphi(F_2)\| \leq 4\sqrt{2} \sum_{y=0}^{1} \sum_{d=0}^{1} \sup_{\mathbf{z} \in \Omega_{\mathbf{Z}}} |F_1(y, d|\mathbf{x}, \mathbf{z}) - F_2(y, d|\mathbf{x}, \mathbf{z})|$$

for all pairs of conditional distributions F_1 and F_2 in $\ell^\infty[0,1]$. By Rademacher's theorem on metric measure spaces a Lipshitz function is Fréchet-differentiable almost everywhere in the interior of its domain, and we can therefore conclude that $\varphi(F)$ is Fréchet-differentiable with respect to F. The projection π is a uniformly bounded map that is Fréchet-differentiable with respect to the Hausdorff metric on $[-1, 1]$. It follows from the chain rule (Van der Vaart and Wellner, 1996, Lemma 3.9.3) that $\psi = \pi \circ \varphi$ is Fréchet-

differentiable, and hence Hadamard-differentiable at F. The theorem now follows.

Proof of Corollary 1 The assumption that $\sqrt{n}\|F_n^{(a)} - \mathbb{F}\|_\infty = o((\log\log n)^{\frac{1}{2}})$ implies the uniform norm of the first and third term on the right hand side of Equation (A.1), and the first term on the right hand side of Equation (A.2), below, are dominated by $\sqrt{n}\|\mathbb{F}_n^* - \mathbb{F}_n\|_\infty$ and $\sqrt{n}\|\mathbb{F}_n - F\|_\infty$ respectively.

$$\sqrt{n}(F_n^{*(a)} - F_n^{(a)}) = \sqrt{n}(F_n^{*(a)} - \mathbb{F}_n^*) + \sqrt{n}(\mathbb{F}_n^* - \mathbb{F}_n) - \sqrt{n}(F_n^{(a)} - \mathbb{F}_n) \tag{A.1}$$

$$\sqrt{n}(F_n^{(a)} - F) = \sqrt{n}(F_n^{(a)} - \mathbb{F}_n) + \sqrt{n}(\mathbb{F}_n - F) \tag{A.2}$$

We can therefore conclude that the conditional distribution of $\sqrt{n}(F_n^{*(a)} - F_n^{(a)})$ converges to the same asymptotic limit law as $\sqrt{n}(F_n^{(a)} - F)$ since the conditional distribution of $\sqrt{n}(\mathbb{F}_n^* - \mathbb{F}_n)$ converges to the same asymptotic limit law as $\sqrt{n}(\mathbb{F}_n - F)$ (Van der Vaart and Wellner, 1996, Theorem 3.9.11). An application of the chain rule to $\psi(F_n^{(\boldsymbol{a})}) = \psi(\lambda^{(\boldsymbol{a})}(\mathbb{F}_n))$ implies that $\psi(F_n^{(\boldsymbol{a})})$ is Fréchet-differentiable and the proof that $MHD_{(1-\alpha)}^{(a)}$ has the correct asymptotic coverage probability is now completed as in the proof of Theorem 1.

Proof of Theorem 2 From Assumption M1 and Assumption M2 and Skorohod's represenation theorem it follows that there exists a parameter value $\boldsymbol{\theta}_{0n}$ and an associated $\boldsymbol{V}(\boldsymbol{\theta}_{0n})$ such that the M-estimator $\widehat{\boldsymbol{\theta}}_n$ satisfies plim$\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{0n}\| = 0$ and $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{0n}) = \boldsymbol{V}(\boldsymbol{\theta}_{0n})^{\frac{1}{2}}\boldsymbol{\zeta}_n + o_p(1)$ where $\boldsymbol{\zeta}_n$ converges in distribution to $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Employing the delta method and the chain rule using the differentiability of ψ and $F_n^{(\boldsymbol{\theta})}$ implies that $\sqrt{n}(\psi(F_n^{(\widehat{\boldsymbol{\theta}}_n)}) - \psi(F_n^{(\boldsymbol{\theta}_{0n})}))$ converges weakly to a Gaussian process. Similary, $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n^* - \widehat{\boldsymbol{\theta}}_n)) = \boldsymbol{V}(\widehat{\boldsymbol{\theta}}_n)^{\frac{1}{2}}\boldsymbol{\zeta}_n^* + o_p(1)$ where $\boldsymbol{\zeta}_n^* \Rightarrow \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\|\boldsymbol{V}(\boldsymbol{\theta}_{0n}) - \boldsymbol{V}(\widehat{\boldsymbol{\theta}}_n)\| \to 0$ as $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{0n}\| \to 0$. From Theorem 23.5 of Van der Vaart (1998) (See also Van der Vaart and Wellner, 1996, Chapter 3.9.3) we can conclude that the conditional distribution of $\sqrt{n}(\psi(F_n^{(\widehat{\boldsymbol{\theta}}_n^*)}) - \psi(F_n^{(\widehat{\boldsymbol{\theta}}_n)}))$ yields an asymptotically weakly consistent estimate of the limit law of $\sqrt{n}(\psi(F_n^{(\widehat{\boldsymbol{\theta}}_n)}) - \psi(F_n^{(\boldsymbol{\theta}_{0n})}))$, establishing the first part of the theorem.

That $MHD_{(1-\alpha)}^{(a)}$ has the correct asymptotic coverage probability when $F_n^{(\boldsymbol{\theta}_{0n})}$ approximates the observational equivalence class of the structure $\mathfrak{S}$ and $\sqrt{n}(\psi(F_n^{(\boldsymbol{\theta}_{0n})}) - \psi(F)) = o(1)$ follows from observing that

$$\begin{aligned}\sqrt{n}(\psi(F_n^{(\widehat{\boldsymbol{\theta}}_n)}) - \psi(F_n^{(\boldsymbol{\theta}_{0n})})) &= \sqrt{n}(\psi(F_n^{(\widehat{\boldsymbol{\theta}}_n)}) - \psi(F)) + \sqrt{n}(\psi(F) - \psi(F_n^{(\boldsymbol{\theta}_{0n})})) \\ &= \sqrt{n}(\psi(F_n^{(\widehat{\boldsymbol{\theta}}_n)}) - \psi(F)) + o(1)\,. \end{aligned} \tag{A.3}$$

Combining Equation (A.3) with the convergence in distribution of $\sqrt{n}(\psi(F_n^{(\widehat{\boldsymbol{\theta}}_n^*)}) - \psi(F_n^{(\widehat{\boldsymbol{\theta}}_n)}))$ and $\sqrt{n}(\psi(F_n^{(\widehat{\boldsymbol{\theta}}_n)}) - \psi(F_n^{(\boldsymbol{\theta}_{0n})}))$ yields the result that $\sqrt{n}(\psi(F_n^{(\widehat{\boldsymbol{\theta}}_n^*)}) - \psi(F_n^{(\widehat{\boldsymbol{\theta}}_n)}))$ converges in distribution to $\sqrt{n}(\psi(F_n^{(\widehat{\boldsymbol{\theta}}_n)}) - \psi(F))$. The second result stated in the theorem now follows as in the proof of Theorem 1 and its corollary.

Acknowledgement: We are grateful to Ze-Yu Zhong for research assistance and for running the simulation results discussed in this paper on the Monash University GPU-based supercomputer dubbed MASSIVE-3. We alone are responsible for collating, analysing and presenting the results, and are culpable for any outstanding errors or misconceptions.

Funding: Financial support under Australian Research Council grant DP210103094 is gratefully acknowledged.

References

Andrews, D. W. and Barwick, P. J. (2012). Inference for parameters defined by moment inequalities: a recommended moment selection procedure. *Econometrica* **80** 2805–2826.

Andrews, D. W. and Soares, G. (2010). Inference for parameters defined by moment inequalities using generalized moment selection. *Econometrica* **78** 119–157.

Barndorff-Nielsen, O. and Pedersen, B. V. (1979). The bivariate hermite polynomials up to order six. *Scandinavian Journal of Statistics* 127–128.

Beresteanu, A. and Molinari, F. (2008). Asymptotic properties for a class of partially identified models. *Econometrica* **76** 763–814.

Bugni, F. A. (2010). Bootstrap inference in partially identified models defined by moment inequalities: coverage of the identified set. *Econometrica* **78** 735–753.

Canay, I. A. (2010). EL inference for partially identified models: large deviations optimality and bootstrap validity. *Journal of Econometrics* **156** 408–425.

Chernozhukov, V., Hong, H. and Tamer, E. (2007). Estimation and confidence regions for parameter sets in econometric models. *Econometrica* **75** 1243–1284.

Chernozhukov, V., Lee, S. and Rosen, A. (2013). Intersection bounds: Estimation and inference. *Econometrica* **81** 667–737.

Chesher, A. (2005). Nonparametric identification under discrete variation. *Econometrica* **73** 1525–1550.

Chesher, A. (2010). Instrumental variable models for discrete outcomes. *Econometrica* **78** 575–601.

Flores, C. A. and Chen, X. (2018). *Average treatment effect bounds with an instrumental variable: Theory and practice*. Springer: Singapore.

Ho, K. and Rosen, A. M. (2013). Partial identification in applied research: Benefits and challenges.

Horowitz, J. L. and Manski, C. F. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association* **95** 77–84.

Huber, P. J. (1967). *The behaviour of maximum likelihodd estimates under non-standard conditions*, chap. Proceedings Berkeley Symposium Mathematical Statistics and Probability, Vol. 1. 221–233.

Imbens, G. W. and Manski, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica* **72** 1845–1857.

Kitagawa, T. (2009). Identification region of the potential outcome distributions under instrument independence. *CeMMAP working papers CWP 30/09* .

Li, C., Poskitt, D. and Zhao, X. (2018). Bounds for average treatment effect: A comparison of nonparametric and quasi maximum likelihood estimators. Tech. rep., Working Paper, Monash University.

Li, C., Poskitt, D. S. and Zhao, X. (2019a). The bivariate probit model, maximum likelihood estimation, pseudo true parameters and partial identification. *Journal of Econometrics* **209** 94–113.

Li, C., Poskitt, D. S. and Zhao, X. (2019b). Bounding the effect of private health insurance on dental care utilisation. Tech. rep.

Manski, C. F. (1988). Identification of binary response models. *Journal of the American statistical Association* **83** 729–738.

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

Menzel, K. (2014). Consistent estimation with many moment inequalities. *Journal of Econometrics* **182** 329–350.

Neyman, J. S. (1923). On the application of probability theory to agricultural experiments. essays on principles, section 9. *English translation in Statistical Science* **5** 465 – 480. 1990.

Romano, J. P. and Shaikh, A. M. (2010). Inference for the identified set in partially identified econometric models. *Econometrica* **78** 169–211.

Rosen, A. M. (2008). Confidence sets for partially identified parameters that satisfy a finite number of moment inequalities. *Journal of Econometrics* **146** 107–117.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66** 688.

Stuart, A. and Ord, J. K. (1993). *Kendall's Advanced Theory of Statistics*. Edward Arnold.

Swanson, S. A., Hernán, M. A., Miller, M., Robins, J. M. and Richardson, T. S. (2018). Partial identification of the average treatment effect using instrumental variables: review of methods for binary instruments, treatments, and outcomes. *Journal of the American Statistical Association* **113** 933–947.

Tamer, E. (2010). Partial Identification in Econometrics. *Annual Review of Economics* **2** 167–195.

Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.

Vytlacil, E. (2006). A note on additive separability and latent index models of binary choice: representation results. *Oxford Bulletin of Economics and Statistics* **68** 515–518.