More on Confidence Intervals for Partially Identified Parameters

Jörg Stoye*

New York University

(First draft: January 2007.)

April 8, 2008

Abstract

This paper extends Imbens and Manski's (2004) analysis of confidence intervals for interval identified parameters. For their final result, Imbens and Manski implicitly assume superefficient estimation of a nuisance parameter. This appears to have gone unnoticed before, and it limits the result's applicability.

I re-analyze the problem both with assumptions that merely weaken the superefficiency condition and with assumptions that remove it altogether. Imbens and Manski's confidence region is found to be valid under weaker assumptions than theirs, yet superefficiency is required. I also provide a different confidence interval that is valid under superefficiency but can be adapted to the general case, in which case it embeds a specification test for nonemptiness of the identified set. A methodological contribution is to notice that the difficulty of inference comes from a boundary problem regarding a nuisance parameter, clarifying the connection to other work on partial identification.

Keywords: Bounds, identification regions, confidence intervals, uniform convergence, superefficiency.

JEL classification codes: C10, C14.

*I am indebted to Adam Rosen, whose insightful comments helped to much improve this paper. I also thank Don Andrews, Xiaohong Chen, Aureo de Paula, Yanqin Fan, Guido Imbens, Thierry Magnac, Chuck Manski, Francesca Molinari, Hyungsik Roger Moon, and Sang Soo Park for feedback and encouragement and seminar audiences at Brown, Cologne, Mannheim, Penn, Vanderbilt, Yale, as well as conference audiences in Budapest (Econometric Society 2007 European Summer Meeting), Durham (Econometric Society 2007 North American Summer Meeting), London (CEMMAP conference on inference in partially identified models), Munich (Verein für Socialpolitik 2007 annual meeting), and Philadelphia (2007 Greater New York Metropolitan Area Econometrics Colloquium), for comments. Of course, any and all errors are mine. Address: Department of Economics, New York University, 19 W. 4^{th} Street, New York, NY 10012, j.stoye@nyu.edu.

1 Introduction

Analysis of partial identification, that is, of models where only bounds on parameters are identified, has become an active field of econometrics.[1] Within this field, attention has only recently turned to general treatments of estimation and inference. An important contribution in this direction is due to Imbens and Manski (2004, IM henceforth). Their major innovation is to point out that in constructing confidence regions for partially identified parameters, one might be interested in coverage probabilities for the parameter rather than its "identified set." The intuitively most obvious, and previously used, confidence regions have nominal coverage probabilities defined for the latter, which means that they are conservative with respect to the former. IM go on to propose a number of confidence regions designed to cover real-valued parameters that can be asymptotically concluded to lie in an interval.

This paper refines and extends IM's technical analysis, specifically their last result, a confidence interval that exhibits uniform coverage of partially identified parameters if the length of the identified interval is a nuisance parameter. IM's proof of coverage for that confidence set relies on a high-level assumption that turns out to imply superefficient estimation of this nuisance parameter and that will fail in many applications. I take this discovery as point of departure for a new analysis of the problem, providing different confidence intervals that are valid both with and without superefficiency.

A brief summary and overview of results goes as follows. In section 2, I describe a simplified, and somewhat generalized, version of IM's model, briefly summarize the relevant aspects of their contribution, and explain the aforementioned issue. Section 3 provides a re-analysis of the problem. I first show how to construct a confidence region if the length of the identified interval is known. This case is a simple but instructive benchmark; subsequent complications stem from the fact that the interval's length is generally a nuisance parameter. Section 3.2 analyses inference given superefficient estimation of this nuisance parameter. It reconstructs IM's result from weaker and, as will be shown, significantly more generic assumptions, but also proposes a different confidence region. In section 3.3, superefficiency is dropped altogether. This case requires a quite different analysis, and I propose a confidence region that adapts the last of the previous ones and embeds a specification test for emptiness of the identified set. Section 4 concludes and highlights connections to current research on partially identified models. The appendix contains all proofs.

[1] See Manski (2003) for a survey and Haile and Tamer (2003) as well as Honoré and Tamer (2006) for recent contributions.

2 Background

Following Woutersen (2006), I consider a simplification and generalization of IM's setup that removes some nuisance parameters. The object of interest is the real-valued parameter $\theta_0(P)$ of a probability distribution $P(X)$; P must lie in a set $\mathcal{P}$ that is characterized by ex ante constraints (maintained assumptions). The random variable X is not completely observable, so that θ_0 may not be identified. Assume, however, that the observable aspects of $P(X)$ identify bounds $\theta_l(P)$ and $\theta_u(P)$ s.t. $\theta_0 \in [\theta_l, \theta_u]$ a.s. See the aforecited references for examples. The interval $\Theta_0 \equiv [\theta_l, \theta_u]$ will also be called *identified set*. Let $\Delta(P) \equiv \theta_u - \theta_l$ denote its length; obviously, Δ is identified as well. Assume that estimators $\widehat{\theta}_l$, $\widehat{\theta}_u$, and $\widehat{\Delta}$ exist and are connected by the identity $\widehat{\Delta} \equiv \widehat{\theta}_u - \widehat{\theta}_l$.

Confidence regions for identified sets of this type are conventionally formed as

$$CI_\alpha = \left[\widehat{\theta}_l - \frac{c_\alpha \widehat{\sigma}_l}{\sqrt{N}}, \widehat{\theta}_u + \frac{c_\alpha \widehat{\sigma}_u}{\sqrt{N}}\right],$$

where $\widehat{\sigma}_l$ $[\widehat{\sigma}_u]$ is a standard error for $\widehat{\theta}_l$ $[\widehat{\theta}_u]$, and where c_α is chosen s.t.

$$\Phi(c_\alpha) - \Phi(-c_\alpha) = 1 - \alpha. \tag{1}$$

For example, $c_\alpha = \Phi^{-1}(0.975) \approx 1.96$ for a 95%-confidence interval. Under regularity conditions, $\Pr(\Theta_0 \subseteq CI_\alpha) \rightarrow 1 - \alpha$; see Horowitz and Manski (2000). IM's contribution is motivated by the observations that (i) one might be interested in coverage of θ_0 rather than Θ_0, (ii) whenever $\Delta > 0$, then $\Pr(\theta_0 \in CI_\alpha) \rightarrow 1 - \alpha/2$. In words, a 90% C.I. for Θ_0 is a 95% C.I. for θ_0. The reason is that asymptotically, Δ is large relative to sampling error, so that noncoverage risk is effectively one-sided at $\{\theta_l, \theta_u\}$ and vanishes otherwise. One would, therefore, be tempted to construct a level α C.I. for θ as $CI_{2\alpha}$.[2]

Unfortunately, this intuition works pointwise but not uniformly over interesting specifications of $\mathcal{P}$. Specifically, $\Pr(\theta_0 \in CI_\alpha) = 1 - \alpha$ if $\Delta = 0$ and also $\Pr(\theta_0 \in CI_\alpha) \rightarrow 1 - \alpha$ along any local parameter sequence where $\Delta_N \leq O(N^{-1/2})$, i.e. Δ fails to diverge relative to sampling error. While uniformity failures are standard in econometrics, this one is unpalatable because it concerns a very salient region of the parameter space; were it neglected, one would be led to construct confidence intervals that *shrink* as a parameter moves from point identification to slight underidentification.[3]

[2] To avoid uninstructive complications, I presume $\alpha \leq .5$ throughout.

[3] The problem would be avoided if $\mathcal{P}$ were restricted s.t. Δ is bounded away from 0. But such a restriction will frequently be inappropriate. For example, one cannot a priori bound from below the degree of item nonresponse in a survey or of attrition in a panel.

Even in cases where Δ is known a priori, e.g. interval data, the problem arguably disappears only in a superficial sense. Were it ignored, one would construct confidence intervals that work uniformly given any model but whose performance deteriorates across models as point identification is approached.

IM therefore conclude by proposing an intermediate confidence region that takes the uniformity problem into account. It is defined as

$$CI_{\alpha}^{1} \equiv \left[\widehat{\theta}_l - \frac{c_{\alpha}^{1}\widehat{\sigma}_l}{\sqrt{N}}, \widehat{\theta}_u + \frac{c_{\alpha}^{1}\widehat{\sigma}_u}{\sqrt{N}}\right], \tag{2}$$

where c_α solves

$$\Phi\left(c_{\alpha}^{1} + \frac{\sqrt{N}\widehat{\Delta}}{\max\left\{\widehat{\sigma}_l, \widehat{\sigma}_u\right\}}\right) - \Phi\left(-c_{\alpha}^{1}\right) = 1 - \alpha. \tag{3}$$

Comparison with (1) reveals that calibration of c_{α}^{1} takes into account the estimated length of the identified set. For a 95% confidence set, c_{α}^{1} will be $\Phi^{-1}(0.975) \approx 1.96$ if $\widehat{\Delta} = 0$, that is if point identification must be presumed, and will approach $\Phi^{-1}(0.95) \approx 1.64$ as $\widehat{\Delta}$ grows large relative to sampling error. IM show uniform validity of CI_{α}^{1} under the following assumption.

Assumption 1 ***(i)*** *There exist estimators $\widehat{\theta}_l$ and $\widehat{\theta}_u$ that satisfy:*

$$\sqrt{N}\left[\begin{array}{c} \widehat{\theta}_l - \theta_l \\ \widehat{\theta}_u - \theta_u \end{array}\right] \xrightarrow{d} N\left(\left[\begin{array}{c} 0 \\ 0 \end{array}\right], \left[\begin{array}{cc} \sigma_l^2 & \rho\sigma_l\sigma_u \\ \rho\sigma_l\sigma_u & \sigma_u^2 \end{array}\right]\right)$$

uniformly in $P \in \mathcal{P}$, and there are estimators $\left(\widehat{\sigma}_l^2, \widehat{\sigma}_l^2, \widehat{\rho}\right)$ that converge to their population values uniformly in $P \in \mathcal{P}$.

(ii) *For all $P \in \mathcal{P}$, $\underline{\sigma}^2 \leq \sigma_l^2, \sigma_u^2 \leq \overline{\sigma}^2$ for some positive and finite $\underline{\sigma}^2$ and $\overline{\sigma}^2$, and $\theta_u - \theta_l \leq \overline{\Delta} < \infty$.*

(iii) *For all $\epsilon > 0$, there are $v > 0$, K, and N_0 s.t. $N \geq N_0$ implies $Pr\left(\sqrt{N}\left|\widehat{\Delta} - \Delta\right| > K\Delta^{v}\right) < \epsilon$ uniformly in $P \in \mathcal{P}$.*

While it is clear that uniformity can obtain only under restrictions on $\mathcal{P}$, it is important to note that Δ is not bounded from below, thus the specific uniformity problem that arises near point identification is not assumed away. Having said that, conditions (i) and (ii) are fairly standard, but (iii) deserves some explanation. It implies that $\widehat{\Delta}$ approaches its population counterpart Δ in a specific way. If $\Delta = 0$, then $\widehat{\Delta} = 0$ with probability approaching 1 in finite samples, i.e. if point identification obtains, then this will be learned exactly, and the limiting distribution of $\widehat{\Delta}$ must be degenerate. What's more, degenerate limiting distributions occur along any local parameter sequence that converges to zero, as is formally stated in the following lemma.[4]

Lemma 1 *Assumption 1(iii) implies that $\sqrt{N}\left|\widehat{\Delta} - \Delta_N\right| \xrightarrow{p} 0$ for all sequences of distributions $\{P_N\} \subseteq \mathcal{P}$ s.t. $\Delta_N \equiv \Delta(P_N) \to 0$.*

[4] This paper makes heavy use of local parameters, and to minimize confusion, I reserve the subscript $(\cdot)_N$ for deterministic functions of N, including local parameters; hence the use of c_α where IM used C_N. Estimators are denoted by $\widehat{(\cdot)}$ throughout.

In words, assumption 1(iii) requires $\widehat{\Delta}$ to be superefficient at $\Delta = 0$. This feature appears to not have been previously recognized; it is certainly nonstandard and might even seem undesirable.[5] This judgment is moderated by the fact that, as will be shown below, it obtains in numerous applications. Nonetheless, some issues remain. First, superefficiency of $\widehat{\Delta}$ is not given in other leading applications, notably when $\widehat{\theta}_l$ and $\widehat{\theta}_u$ come from moment conditions, and CI_α^1 is not valid without it. Second, the superefficiency condition can be substantially weakened and then turns out to obtain whenever $\widehat{\theta}_u$ and $\widehat{\theta}_l$ are jointly normal and $\widehat{\theta} \geq \widehat{\theta}_l$ by construction. Finally, whenevr superefficiency holds, one can formulate an interesting alternative to CI_α^1 that is also easily adapted to a setting without superefficiency. All in all, there is ample reason to take a second look at the inference problem.

3 Re-analysis of the Inference Problem

3.1 Inference with Known Δ

I will now re-analyze the problem and provide several results that circumvent the aforementioned issues. To begin, assume that Δ is known:

Assumption 2 ***(i)*** *There exists an estimator $\widehat{\theta}_l$ that satisfies:*

$$\sqrt{N}\left[\widehat{\theta}_l - \theta_l\right] \xrightarrow{d} N\left(0, \sigma_l^2\right)$$

uniformly in $P \in \mathcal{P}$, and there is an estimator $\widehat{\sigma}_l^2$ that converges to σ_l^2 uniformly in $P \in \mathcal{P}$.

(ii) $\Delta \geq 0$ *is known.*

(iii) *For all $P \in \mathcal{P}$, $\underline{\sigma}^2 \leq \sigma_l^2, \sigma_u^2 \leq \overline{\sigma}^2$ for some positive and finite $\underline{\sigma}^2$ and $\overline{\sigma}^2$.*

By symmetry, it could of course be θ_u that can be estimated. A natural application for this scenario would be inference about the mean from interval data, where the length of intervals (e.g., income brackets) does not vary on the support of θ. Define

$$\widetilde{CI}_\alpha = \left[\widehat{\theta}_l - \frac{\widetilde{c}_\alpha \widehat{\sigma}_l}{\sqrt{N}}, \widehat{\theta}_u + \frac{\widetilde{c}_\alpha \widehat{\sigma}_l}{\sqrt{N}}\right],$$

where

$$\Phi\left(\widetilde{c}_\alpha + \frac{\sqrt{N}\Delta}{\widehat{\sigma}_l}\right) - \Phi\left(-\widetilde{c}_\alpha\right) = 1 - \alpha. \tag{4}$$

Lemma 2 establishes that this confidence interval is uniformly valid.

[5] When Hodges originally defined a superefficient estimator, his intent was not, of course, to propose its use. For cautionary tales regarding the implicit, and sometimes inadvertent, use of superefficient estimators, see Leeb and Pötscher (2005).

Lemma 2 *Let assumption 2 hold. Then*

$$\lim_{N\to\infty}\inf_{\theta\in\Theta}\inf_{P:\theta_0(P)=\theta}\Pr\left(\theta_0\in\widetilde{CI}_\alpha\right)=1-\alpha.$$

Lemma 2 generalizes IM's lemma 3. It is technically new but easy to prove: The normal approximation to $\Pr\left(\theta_0\in\widetilde{CI}_\alpha\right)$ is concave in θ_0 and equals $(1-\alpha)$ if $\theta_0\in\{\theta_l,\theta_u\}$. The main purpose of lemma 2 is as a backdrop for the case with unknown Δ, when $\widetilde{CI}_\alpha$ is not feasible. As will be seen, the impossibility of estimating $\sqrt{N}\Delta$, and by implication $\widetilde{c}_\alpha$, is the root cause of most complications.

3.2 Inference with Superefficiency

In this section, I assume that Δ is unknown but maintain superefficiency. I begin by showcasing the weakest (to my knowledge) assumption under which CI_α^1 is valid.

Assumption 3 ***(i)*** *There exists an estimator $\widehat{\theta}_l$ that satisfies:*

$$\sqrt{N}\left[\widehat{\theta}_l-\theta_l\right]\xrightarrow{d}N\left(0,\sigma_l^2\right)$$

uniformly in $P\in\mathcal{P}$, and there is an estimator $\widehat{\sigma}_l^2$ that converges to σ_l^2 uniformly in $P\in\mathcal{P}$.

(ii) *There exists an estimator $\widehat{\Delta}$ that satisfies:*

$$\sqrt{N}\left[\left(\widehat{\theta}_l+\widehat{\Delta}\right)-(\theta_l+\Delta)\right]\xrightarrow{d}N\left(0,\sigma_u^2\right)$$

uniformly in $P\in\mathcal{P}$, and there is an estimator $\widehat{\sigma}_u^2$ that converges to σ_u^2 uniformly in $P\in\mathcal{P}$.

(iii) *There exists a sequence $\{a_N\}$ s.t. $a_N\to 0$, $a_N\sqrt{N}\to\infty$, and $\sqrt{N}\left|\widehat{\Delta}-\Delta_N\right|\xrightarrow{p}0$ for all sequences of distributions $\{P_N\}\subseteq\mathcal{P}$ with $\Delta_N\le a_N$.*

(iv) *For all $P\in\mathcal{P}$, $\underline{\sigma}^2\le\sigma_l^2,\sigma_u^2\le\overline{\sigma}^2$ for some positive and finite $\underline{\sigma}^2$ and $\overline{\sigma}^2$.*

Assumption 3 models a situation where θ_u is estimated only indirectly by $\widehat{\theta}_u\equiv\widehat{\theta}_l+\widehat{\Delta}$. (By symmetry, the case of directly estimating (θ_u,Δ) is covered as well.) Importantly, uniform joint asymptotic normality of $\left(\widehat{\theta}_l,\widehat{\theta}_l+\widehat{\Delta}\right)$ is not imposed. Furthermore, condition (iii) has been replaced with a requirement that is strictly weaker and arguably more transparent about what is really being required.[6]

Of course, assumption 3(iii) is again a superefficiency condition, but it is fulfilled in more cases than one might have thought. One example is IM's motivating application, namely estimation of a

[6] By lemma 1, assumption 1(iii) implies assumption 3(iii), where the quantifier is strengthened to "for all sequences $\{a_N\}$..." But assumption 1(iii) is even stronger: If $N\left[\widehat{\Delta}-\Delta\right]\xrightarrow{d}N(0,1)$ uniformly, then assumption 3(iii) is fulfilled with the stronger quantifier but assumption 1(iii) is violated at $\Delta=0$.

mean with missing data. Let $\theta = \mathbb{E}X$, where $X \in [0,1]$, and assume that one observes realizations of $(D, D \cdot X)$, where $D \in \{0,1\}$ indicates whether a data point is present ($D=1$) or missing ($D=0$). Then the identified set for θ_0 is

$$[\theta_l, \theta_u] = [(1-\Delta)\,\mathbb{E}(X|D=1), (1-\Delta)\,\mathbb{E}(X|D=1)+\Delta],$$

where $\Delta \equiv \Pr(D=0) = 1 - \mathbb{E}D$ (the definition as "one minus propensity score" insures consistency with previous use). The obvious estimator for Θ_0 is its sample analog

$$\widehat{\Theta}_0 \equiv \left[\underbrace{\frac{1}{N}\sum_{i=1}^{N} D_i X_i}_{\widehat{\theta}_l}, \underbrace{\frac{1}{N}\sum_{i=1}^{N} D_i X_i}_{\widehat{\theta}_l} + \underbrace{1 - \frac{1}{N}\sum_{i=1}^{N} D_i}_{\widehat{\Delta}}\right]. \tag{5}$$

In this application, indirect estimation of $\widehat{\theta}_u$ as $\widehat{\theta}_l + \widehat{\Delta}$ is, therefore, natural. Under regularity conditions, uniform convergence of $\left(\widehat{\theta}_l, \widehat{\theta}_l + \widehat{\Delta}\right)$ to individually normal distributions follows from a uniform central limit theorem. What's more, $\widehat{\Delta}$ fulfils part (iii) of both assumptions 1 and 3, making it a natural example of a superefficient estimator. Other examples where assumption 3 holds will be elaborated later.

This section's first result is as follows.

Proposition 1 *Let assumption 3 hold. Then*

$$\lim_{N\to\infty} \inf_{\theta\in\Theta} \inf_{P:\theta_0(P)=\theta} \Pr\left(\theta_0 \in CI_\alpha^1\right) = 1 - \alpha.$$

In words, assumption 3 suffices for validity of IM's interval. To understand the use of superefficiency, it is helpful to think of CI_α^1 as feasible version of $\widetilde{CI}_\alpha$, with c_α^1 being an estimator of $\widetilde{c}_\alpha$. Validity of CI_α^1 would easily follow from consistency of c_α^1, but such consistency does not obtain under standard assumptions: $\left(\widehat{\Delta} - \Delta\right)$ is usually of order $O(N^{-1/2})$, so that $\left(\sqrt{N}\widehat{\Delta} - \sqrt{N}\Delta\right)$ does not vanish.

This is where superefficiency comes into play. Think in terms of sequences of distributions P_N that give rise to local parameters Δ_N, and distinguish between sequences where Δ_N vanishes fast enough for condition (iii) to apply and sequences where this fails. In the former case, $\left(\sqrt{N}\widehat{\Delta} - \sqrt{N}\Delta\right)$ does vanish, and consistency of c_α^1 for $\widetilde{c}_\alpha$ is recovered. In the latter case, Δ_N grows uniformly large relative to sampling error, so that the uniformity problem does not arise to begin with. The "naive" $CI_{2\alpha}$ is then a valid construction, and CI_α^1 (as well as $\widetilde{CI}_\alpha$) is asymptotically equivalent to it.

Proposition 1 shows that CI_α^1 is valid under conditions that weaken assumption 1 and remove some hidden restrictions. However, further investigation reveals a nonstandard property of CI_α^1. Its simplicity stems in part from the fact that expression (3) simultaneously calibrates $\Pr\left(\theta_l \in CI_\alpha^1\right)$ and

$\Pr\left(\theta_u \in CI_\alpha^1\right)$. This is possible because $\max\{\widehat{\sigma}_l, \widehat{\sigma}_u\}$ is substituted where one would otherwise have to differentiate between $\widehat{\sigma}_l$ and $\widehat{\sigma}_u$. In effect, c_α^1 is calibrated under the presumption that $\max\{\widehat{\sigma}_l, \widehat{\sigma}_u\}$ will be used as standard error at both ends of the confidence interval. Of course, this presumption is not correct – $\widehat{\sigma}_l$ is used near $\widehat{\theta}_l$ and $\widehat{\sigma}_u$ near $\widehat{\theta}_u$. As a result, the nominal size of CI_α^1 is not $1-\alpha$ in finite samples.[7] If $\widehat{\sigma}_l > \widehat{\sigma}_u$, the interval will be nominally conservative at θ_u, nominally invalid at θ_l, and therefore nominally invalid for θ_0. It also follows that CI_α^1 is the inversion of a hypothesis test for $H_0 : \theta_0 \in \Theta_0$ that is nominally biased, that is, its nominal power is larger for some points inside of Θ_0 than for some points outside of it.

To be sure, this feature is a finite sample phenomenon. Under assumption 3, nominal size of CI_α^1 will approach $1-\alpha$ at both θ_l and θ_u as $N \to \infty$. (For future reference, note the reason for this when Δ_N is small: Superefficiency then implies that $\sigma_l \to \sigma_u$.) Nonetheless, it is of interest to notice that it can be avoided at the price of a mild strengthening of assumptions. Specifically, impose:

Assumption 4 ***(i)*** *There exist estimators $\widehat{\theta}_l$ and $\widehat{\theta}_u$ that satisfy:*

$$\sqrt{N}\begin{bmatrix} \widehat{\theta}_l - \theta_l \\ \widehat{\theta}_u - \theta_u \end{bmatrix} \xrightarrow{d} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_l^2 & \rho\sigma_l\sigma_u \\ \rho\sigma_l\sigma_u & \sigma_u^2 \end{bmatrix}\right)$$

uniformly in $P \in \mathcal{P}$, and there are estimators $\left(\widehat{\sigma}_l^2, \widehat{\sigma}_l^2, \widehat{\rho}\right)$ that converge to their population values uniformly in $P \in \mathcal{P}$.

(ii) *For all $P \in \mathcal{P}$, $\underline{\sigma}^2 \leq \sigma_l^2, \sigma_u^2 \leq \overline{\sigma}^2$ for some positive and finite $\underline{\sigma}^2$ and $\overline{\sigma}^2$, and $\theta_u - \theta_l \leq \overline{\Delta} < \infty$.*

(iii) *There exists a sequence $\{a_N\}$ s.t. $a_N \to 0$, $a_N\sqrt{N} \to \infty$, and $\sqrt{N}\left|\widehat{\Delta} - \Delta_N\right| \xrightarrow{p} 0$ for any sequence of distributions $\{P_N\} \subseteq \mathcal{P}$ with $\Delta_N \leq a_N$.*

Assumption 4 re-introduces joint normality and differs from assumption 1 merely by the modification of part (iii). Of course, this modification continues to be a weakening. Significantly, it can now be derived from a sufficient condition that is easy to check and illustrates the subsequent result's reach.[8]

Lemma 3 *Let assumption 4(i) hold and assume that* $\Pr(\widehat{\theta}_u \geq \widehat{\theta}_l) = 1$ *for all $\mathcal{P}$ (e.g., $\widehat{\theta}_u \geq \widehat{\theta}_l$ holds by construction). Then assumption 4(iii) is implied.*

Thus, assumption 4(iii) will hold whenever $\left(\widehat{\theta}_l, \widehat{\theta}_u\right)$ are jointly asymptotically normal and $\widehat{\theta}_u \geq \widehat{\theta}_l$ holds by construction. The weakened superefficiency condition therefore turns out to be reasonably generic. Examples where it can be easily verified from lemma 3 include the obvious estimators for

[7] By the nominal size of CI_α^1 at θ_l, say, I mean $\int_{CI_\alpha^1} \frac{1}{\widehat{\sigma}_l}\phi((x-\hat{\theta}_l)/\hat{\sigma}_l)dx$, i.e. its size at θ_l as predicted from sample data. Confidence regions are typically constructed by setting nominal size equal to $1-\alpha$.

[8] I thank Thierry Magnac for suggesting this result.

bounds on the mean and any smooth functions of it with missing data, on the mean of interval data when the length of intervals is unknown, and on interior quantiles when there are sufficiently few missing data (i.e. bounds are interior to the support of the data).[9] Of course, because assumption 4 is stronger than assumption 3, it also follows that CI_α^1 is valid under these same conditions.

Some observations regarding tightness of assumptions are as follows.[10] Lemma 3 fails if assumption 4(i) is replaced with assumption 1(i), i.e. joint normality of estimators is required. Also, lemma 3 fails if in the conclusion, assumption 4(iii) is quantified over all sequences $\{a_N\}$ that vanish slowly in the sense of assumption 4(iii). Recalling that by lemma 1, assumption 1(iii) implies the latter strengthening of assumption 4(iii), it also follows that $\Pr(\widehat{\theta}_u \geq \widehat{\theta}_l) = 1$ for all $\mathcal{P}$ is sufficient to establish assumption 4 but not assumption 1. Therefore, the strengthened justification of CI_α^1 is not just a matter of seeing lemma 3, but also requires proposition 1 as opposed to the similar result in IM.

Given assumption 4, one can construct a confidence region that reflects the bivariate nature of the estimation problem by taking into account the correlation between $\widehat{\theta}_l$ and $\widehat{\theta}_u$. Specifically, let (c_l^2, c_u^2) minimize (c_l, c_u) subject to the constraint that

$$\Pr\left(-\frac{c_l}{\widehat{\sigma}_l} \leq z_1, \widehat{\rho} z_1 \leq \frac{c_u + \sqrt{N}\widehat{\Delta}}{\widehat{\sigma}_u} + \sqrt{1-\widehat{\rho}^2} z_2\right) \quad \geq \quad 1-\alpha \tag{6}$$

$$\Pr\left(-\frac{c_l + \sqrt{N}\widehat{\Delta}}{\widehat{\sigma}_l} - \sqrt{1-\widehat{\rho}^2} z_2 \leq \widehat{\rho} z_1, z_1 \leq \frac{c_u}{\widehat{\sigma}_u}\right) \quad \geq \quad 1-\alpha, \tag{7}$$

where z_1 and z_2 are independent standard normal random variables.[11] In typical cases, (c_l^2, c_u^2) will be uniquely characterized by the fact that both of (6,7) hold with equality, but it is conceivable that one of the conditions is slack at the solution. Let

$$CI_\alpha^2 \equiv \left[\widehat{\theta}_l - \frac{c_l^2}{\sqrt{N}}, \widehat{\theta}_u + \frac{c_u^2}{\sqrt{N}}\right].$$

Then:

Proposition 2 *Let assumption 4 hold. Then*

$$\lim_{N\to\infty} \inf_{\theta\in\Theta} \inf_{P:\theta_0(P)=\theta} \Pr\left(\theta_0 \in CI_\alpha^2\right) = 1-\alpha.$$

---

[9] Retunring to the notation of the mean with missing data example, the α-quantile of X is bounded by $F_X^{-1}(\alpha) \in [F_x^{-1}((\alpha-\Delta)/(1-\Delta)), F_X^{-1}(\alpha/1-\Delta)]$. Any reasonable estimators of these quantiles are ordered by construction, and many are uniformly asymptotically bivariate normal (for the case of sample quantiles, see Ikeda and Nonaka 1983).

With some relief, I report that the observation regarding smooth functions of the mean justifies the use of IM in Stoye (2007), although in the light of this paper's results, the argument given there is not complete.

[10] Examples are available from the author.

[11] Appendix B exhibits closed-from expressions for (6,7), illustrating that they can be evaluated without simulation.

Observe that if Δ were known, CI_α^2 would simplify to $\widetilde{CI}_\alpha$: Knowledge of Δ would imply that $\rho = 1$, $\widehat{\rho} = 1$, and $\widehat{\sigma}_l = \widehat{\sigma}_u$, which can be substituted into (6,7) to get

$$\begin{aligned}
\Pr\left(-\frac{c_l}{\widehat{\sigma}_l} \le z \le \frac{c_l + \sqrt{N}\Delta}{\widehat{\sigma}_l}\right) &\ge 1-\alpha \\
\Pr\left(-\frac{c_u + \sqrt{N}\Delta}{\widehat{\sigma}_l} \le z \le \frac{c_u}{\widehat{\sigma}_l}\right) &\ge 1-\alpha,
\end{aligned}$$

where z is standard normal. The program is then solved by setting $c_l^3 = c_u^3 = \widehat{\sigma}_l \widetilde{c}_\alpha$, yielding $\widetilde{CI}_\alpha$. By the same token, CI_α^2 is asymptotically equivalent to CI_α^1 along any parameter sequence where superefficiency applies. For parameter sequences where Δ does not vanish, all of these intervals are asymptotically equivalent anyway because they converge to $CI_{2\alpha}$.

In the regular case where both of (6,7) bind, CI_α^3 has nominal size of exactly $1-\alpha$ at both endpoints of Θ_0, and accordingly corresponds to a nominally unbiased hypothesis test. This might be considered a refinement, although (i) given asymptotic equivalence, it will only matter in small samples, and (ii) nominal size must be taken with a large grain of salt due to nonvanishing estimation error in $\sqrt{N}\Delta$. Perhaps the more important difference is that CI_α^2, unlike CI_α^1, is readily adapted to the more general case.[12]

3.3 Inference with Joint Normality

While the superefficiency assumption was seen to have applications, it is of obvious interest to consider inference about θ_0 without it. For a potential application, imagine that $\widehat{\theta}_u$ and $\widehat{\theta}_l$ derive from separate inequality moment conditions, as in Pakes et al.'s (2006) analysis of investment in automated teller machines. The situation would also occur in the preceding examples whenever researchers refine bounds by means of partially identifying assumptions that are testable; such assumptions can turn identified sets into singletons or even empty sets and will not, in general, come with superefficient indicators of when this happens. I therefore turn to the following assumption.

Assumption 5 ***(i)*** *There are estimators $\widehat{\theta}_l$ and $\widehat{\theta}_u$ that satisfy:*

$$\sqrt{N}\begin{bmatrix} \widehat{\theta}_l - \theta_l \\ \widehat{\theta}_u - \theta_u \end{bmatrix} \stackrel{d}{\to} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_l^2 & \rho\sigma_l\sigma_u \\ \rho\sigma_l\sigma_u & \sigma_u^2 \end{bmatrix}\right)$$

uniformly in $P \in \mathcal{P}$, and there are estimators $(\widehat{\sigma}_l, \widehat{\sigma}_u, \widehat{\rho})$ that converge to their population values uniformly in $P \in \mathcal{P}$.

[12] As an aside, this section's findings resolve questions posed to me by Adam Rosen and other readers of IM, namely, (i) why CI_α^1 is valid even though ρ is not estimated and (ii) whether estimating ρ can lead to a refinement. The brief answers are: (i) Superefficiency implies $\rho = 1$ in the critical case, eliminating the need to estimate it; indeed, it is now seen that mention of ρ can be removed from the assumptions. (ii) Estimating ρ allows for inference that is different in finite samples but not under first-order asymptotics.

(ii) *For all* $P \in \mathcal{P}$, $\underline{\sigma}^2 \leq \sigma_l^2, \sigma_u^2 \leq \overline{\sigma}^2$ *for some positive and finite* $\underline{\sigma}^2$ *and* $\overline{\sigma}^2$, *and* $\Delta \leq \overline{\Delta} < \infty$.

Relative to previous assumptions, assumption 5 simply removes superefficiency. This leads to numerous difficulties. At the core of these lies the fact that sample variation in $\widehat{\Delta}$ need not vanish as $\Delta \to 0$, inducing boundary problems in the implicit estimation of Δ. Indeed, Δ is the exact example for inconsistency of the bootstrap given by Andrews (2000), and it is not possible to consistently estimate a local parameter $\Delta_N = O(N^{-1/2})$.

To circumvent this issue, I use a shrinkage estimator

$$\Delta^* \equiv \begin{cases} \widehat{\Delta}, & \widehat{\Delta} > b_N \\ 0 & \text{otherwise} \end{cases},$$

where b_N is some pre-assigned sequence s.t. $b_N \to 0$ and $b_N\sqrt{N} \to \infty$. Δ^* will replace $\widehat{\Delta}$ in the calibration of c_α but not in the subsequent construction of a confidence region. This will insure uniform validity, intuitively because superefficiency at $\Delta = 0$ is artificially restored. Of course, there is some price to be paid: The confidence region presented below will be uniformly valid and pointwise exact, but asymptotically dissimilar, i.e. conservative along local parameter sequences.

A second modification relative to IM is that I propose to generalize not CI_α^1 but CI_α^2. The reason is that without superefficiency, the distortion of nominal size of CI_α^1 will persist for large N as Δ vanishes, and the interval is accordingly expected to be invalid. (Going back to the discussion that motivated CI_α^2, the problem is that $\Delta_N \to 0$ does not any more imply $\sigma_l \to \sigma_u$.) Hence, let (c_l^3, c_u^3) minimize $(c_l + c_u)$ subject to the constraint that

$$\Pr\left(-\frac{c_l}{\widehat{\sigma}_l} \leq z_1, \widehat{\rho} z_1 \leq \frac{c_u + \sqrt{N}\Delta^*}{\widehat{\sigma}_u} + \sqrt{1-\widehat{\rho}^2} z_2\right) \quad \geq \quad 1-\alpha \tag{8}$$

$$\Pr\left(-\frac{c_l + \sqrt{N}\Delta^*}{\widehat{\sigma}_l} + \sqrt{1-\widehat{\rho}^2} z_2 \leq \widehat{\rho} z_1, z_1 \leq \frac{c_u}{\widehat{\sigma}_u}\right) \quad \geq \quad 1-\alpha, \tag{9}$$

where z_1 and z_2 are independent standard normal random variables. As before, it will typically but not necessarily be the case that both of (8,9) bind at the solution. Finally, define

$$CI_\alpha^3 \equiv \begin{cases} \left[\widehat{\theta}_l - \frac{c_l^3}{\sqrt{N}}, \widehat{\theta}_u + \frac{c_u^3}{\sqrt{N}}\right], & \widehat{\theta}_l - \frac{c_l^3}{\sqrt{N}} \leq \widehat{\theta}_u + \frac{c_u^3}{\sqrt{N}} \\ \varnothing & \text{otherwise} \end{cases}. \tag{10}$$

The definition reveals a third modification: If $\widehat{\theta}_u$ is too far below $\widehat{\theta}_l$, then CI_α^3 is empty, which can be interpreted as rejection of the maintained assumption that $\theta_u \geq \theta_l$. In other words, CI_α^3 embeds a specification test. IM do not consider such a test, presumably for two reasons: It does not arise in their leading application, and it is trivial in their framework because superefficiency implies fast learning about Δ in the critical region where $\Delta \approx 0$. But the issue is substantively interesting in other applications, and is nontrivial when, as in this section's examples, $\widehat{\Delta} < 0$ is a generic possibility. Of course, one could construct a version of CI_α^3 that is never empty; one example would be the convex hull of $\left\{\widehat{\theta}_l - c_l^3/\sqrt{N}, \widehat{\theta}_u + c_u^3/\sqrt{N}\right\}$. But realistically, samples where $\widehat{\theta}_u$ is much below $\widehat{\theta}_l$ would lead one to question whether $\theta_u \geq \theta_l$ holds. This motivates the specification test, which does not affect the interval's asymptotic validity.

This section's result is the following.

Proposition 3 *Let assumption 5 hold. Then*

$$\lim_{N\to\infty} \inf_{\theta\in\Theta} \inf_{P:\theta_0(P)=\theta} \Pr\left(\theta_0 \in CI_\alpha^3\right) = 1-\alpha.$$

An intriguing aspect of CI_α^3 is that it is analogous to CI_α^2, except that it uses Δ^* and accommodates the resulting possibility that $\widehat{\theta}_l - \frac{c_l^3}{\sqrt{N}} > \widehat{\theta}_u + \frac{c_u^3}{\sqrt{N}}$. Together, CI_α^2 and CI_α^3 therefore provide a unified approach to inference for interval identified parameters – one can switch between the setting with and without superefficiency by substituting Δ^* for $\widehat{\Delta}$.

Some further remarks on CI_α^3 are in order.

- The construction of Δ^* can be refined in two ways. First, I defined a soft thresholding estimator for simplicity, but making Δ^* a smooth function of $\widehat{\Delta}$ would also insure validity and might improve performance for Δ close to b_N. Second, the sequence b_N is left to adjustment by the user. This adjustment is subject to the following trade-off: The slower b_N vanishes, the less conservative CI_α^3 is along local parameter sequences, but the quality of the uniform approximation to $\lim_{N\to\infty} \inf_{\theta\in\Theta} \inf_{P:\theta_0(P)=\theta} \Pr\left(\theta_0 \in CI_\alpha^3\right)$ deteriorates, and uniformity breaks down for $b_N = O(N^{-1/2})$. Fine-tuning this trade-off is a possible subject of further research. Importantly, a more sophisticated choice of Δ^* will not make the interval asymptotically exact along local sequences. A simple way to see this is to note the interval is valid if $\Delta = 0$ only if Δ^* is sparse in the sense of Leeb and Pötscher (2008), implying that Δ^* cannot be uniformly consistent for Δ.

- The event that $\Delta^* = 0$ can be interpreted as failure of a pre-test to reject $H_0 : \theta_u = \theta_l$, where the size of the pre-test approaches 1 as $N \to \infty$. In this sense, the present approach is similar to the "conservative pre-test" solution to the parameter-on-the-boundary problem given by Andrews (2000, section 4). However, one should not interpret CI_α^3 as being based on model selection. If $\theta_u = \theta_l$, then it is efficient to estimate both from the same variance-weighted average of $\widehat{\theta}_u$ and $\widehat{\theta}_l$ and to construct an according Wald confidence region, and a post-model selection confidence region would do just that. Unfortunately, the method would be invalid if $\Delta_N = O(N^{-1/2})$. In contrast, CI_α^3 employs a shrinkage estimator of Δ to calibrate cutoff values for implicit hypothesis tests, but not in the subsequent construction of the interval. What's more, even this first step is not easily interpreted as model selection once Δ^* is smoothed.

Having said that, there is a tight connection between the parameter-on-the-boundary issues encountered here and issues with post-model selection estimators (Leeb and Pötscher 2005), the underlying problem being discontinuity of pointwise limit distributions. See Andrews and Guggenberger (2007) for a more elaborate discussion.

- CI_α^3 could be simplified by letting $c_l^3 = \widehat{\sigma}_l \Phi^{-1}(1-2\alpha)$ and $c_u^3 = \widehat{\sigma}_u \Phi^{-1}(1-2\alpha)$, implying that $CI_\alpha^3 = CI_{2\alpha}$, whenever $\Delta^* = \widehat{\Delta}$. This would render the interval shorter without affecting its first-order asymptotics, because the transformation of $\widehat{\Delta}$ suffices to insure uniformity. But it would imply that whenever $\Delta^* = \widehat{\Delta}$, the confidence region ignores the two-sided nature of noncoverage risk and hence has nominal size below $1-\alpha$ (although approaching $1-\alpha$ as N grows large). The improvement in interval length is due to this failure of nominal size and therefore spurious.

- Finally, CI_α^3 holds under more general conditions than the preceding confidence intervals, but these conditions are still with substantial loss of generality for partially identified models. To be sure, uniform validity of normal approximations could be replaced by uniform validity of bootstrap approximations after minor adjustments. But neither of these will obtain if upper and lower bounds are characterized as minima respectively maxima over a number of moment conditions. Also, while the present setting is clearly a special case of moment conditions, it is not obvious how to generalize the construction of CI_α^3 to the case of many moment conditions. Other methods that were developed independently of, and sometimes subsequently to, the present approach (Andrews and Guggenberger 2007, Andrews and Soares 2007, Fan and Park 2007) will work more generally. When CI_α^3 applies, however, it is attractive for a numer of reasons: It is trivial to compute, it is by construction the shortest interval whose nominal size is exact at both ends (this feature has also received Monte Carlo validation in Fan and Park 2007), and it is is easily adjusted to the presence or absence of superefficiency by switching between $\widehat{\Delta}$ and Δ^* in the calibration of cutoff values.

4 Conclusion

This paper extended Imbens and Manski's (2004) analysis of confidence regions for partially identified parameters. A brief summary of its findings goes as follows. First, I establish that one assumption used for IM's final result boils down to superefficient estimation of a nuisance parameter Δ. This nature of their assumption appears to have gone unnoticed before. The inference problem is then re-analyzed with and without superefficiency. IM's confidence region is found to be valid under conditions that are substantially weaker than theirs, an interesting sufficient condition being the case where estimators of upper and lower bounds are jointly asymptotically normal and ordered by construction. Furthermore, valid inference can be achieved by a different confidence region that is easily adapted to the case without superefficiency, in which case it also embeds a specification test.

A conceptual contribution beyond these findings is to recognize that the gist of the inference problem lies in estimation of Δ, specifically when it is small. This insight allows for rather brief and transparent proofs. More importantly, it connects the present, very specific setting to more general models of partial identification. For example, once the boundary problem has been recognized, analogy to Andrews (2000) suggests that a straightforward normal approximation, as well as a bootstrap, will fail, whereas subsampling might work. Indeed, carefully specified subsampling techniques are known to yield valid inference for parameters identified by moment inequalities, of which the present scenario is a special case (Andrews and Guggenberger 2007, Chernozhukov, Hong, and Tamer 2007, Romano and Shaikh 2008). The bootstrap, on the other hand, does not work in the same setting, unless it is modified in several ways, one of which resembles the trick employed here (Bugni 2007). Against the backdrop of these (subsequent to IM) results, validity of simple normal approximations in IM appears as a puzzle that is now resolved. At the same time, the updated version of these normal approximations has practical value because it provides closed-form and otherwise attractive inference for important, if relatively simple, applications.

A Proofs

Lemma 1 The aim is to show that if $\Delta_N \to 0$, then

$$\forall \delta, \varepsilon > 0, \exists N^* : N \geq N^* \Longrightarrow \Pr\left(\sqrt{N}\left|\widehat{\Delta} - \Delta_N\right| > \delta\right) < \varepsilon.$$

Fix δ and ε. By assumption 1(iii), there exist N^{**}, $v > 0$, and k s.t.

$$N \geq N^{**} \Longrightarrow \Pr\left(\sqrt{N}\left|\widehat{\Delta} - \Delta\right| > K\Delta^v\right) < \varepsilon$$

uniformly over $\mathcal{P}$. Specifically, the preceding inequality will obtain if Δ is chosen in $(0, \delta^{1/v}K^{-1/v}]$, in which case $K\Delta^v \leq \delta$. Because $\Delta_N \to 0$, N^{***} can be chosen s.t. $N \geq N^{***} \Rightarrow \Delta_N \leq \delta^{1/v}K^{-1/v}$. Hence, the conclusion obtains by choosing $N^* = \max\{N^{**}, N^{***}\}$.

Lemma 2 Parameterize θ_0 as $\theta_0 = \theta_l + a\Delta$ for some $a \in [0,1]$. Then

$$\begin{aligned}
& \Pr\left(\theta_0 \in \widetilde{CI}_\alpha\right) \\
= \;& \Pr\left(\widehat{\theta}_l - \frac{\widetilde{c}_\alpha \widehat{\sigma}_l}{\sqrt{N}} \leq \theta_l + a\Delta \leq \widehat{\theta}_l + \Delta + \frac{\widetilde{c}_\alpha \widehat{\sigma}_l}{\sqrt{N}}\right) \\
= \;& \Pr\left(-\widetilde{c}_\alpha \frac{\widehat{\sigma}_l}{\sigma_l} - \frac{\sqrt{N}}{\sigma_l} a\Delta \leq \frac{\sqrt{N}\left(\theta_l - \widehat{\theta}_l\right)}{\sigma_l} \leq \frac{\sqrt{N}}{\sigma_l}(1-a)\Delta + \widetilde{c}_\alpha \frac{\widehat{\sigma}_l}{\sigma_l}\right) \\
\to \;& \Phi\left(\widetilde{c}_\alpha + \frac{\sqrt{N}}{\sigma_l}(1-a)\Delta\right) - \Phi\left(-\widetilde{c}_\alpha - \frac{\sqrt{N}}{\sigma_l} a\Delta\right)
\end{aligned}$$

uniformly over $\mathcal{P}$. Besides uniform asymptotic normality of $\widehat{\theta}_l$, this convergence statement uses that by uniform consistency of $\widehat{\sigma}_l$ in conjunction with the lower bound on σ_l, $\widehat{\sigma}_l/\sigma_l \to 1$ uniformly, and also that the derivative of the standard normal c.d.f. is uniformly bounded.

Evaluation of derivatives straightforwardly establishes that the last expression in the preceding display is strictly concave in a, hence it is minimized at $a \in \{0,1\} \Leftrightarrow \theta_0 \in \{\theta_l, \theta_u\}$. But in those cases, the preceding algebra simplifies to

$$\Pr\left(\theta_l \in \widetilde{CI}_\alpha\right) \to \Phi\left(\frac{\sqrt{N}}{\sigma_l}\Delta + \widetilde{c}_\alpha\right) - \Phi\left(-\widetilde{c}_\alpha\right) = 1 - \alpha$$

and similarly for θ_u.

Preliminaries to Subsequent Proofs The following proofs mostly consider sequences $\{P_N\}$ that will be identified with the implied sequences $\{\Delta_N, \theta_N\} \equiv \{\Delta(P_N), \theta_0(P_N)\}$. For ease of notation, I will generally suppress the N subscript on $(\theta_l, \sigma_l, \sigma_u)$ and on estimators. Some algebraic steps treat $(\theta_l, \sigma_l, \sigma_u)$ as constant; this is w.l.o.g. because by compactness implied in part (ii) of every assumption, any sequence $\{P_N\}$ induces a sequence of values $(\theta_l, \sigma_l, \sigma_u)$ with finitely many accumulation points, and the argument can be conducted separately for the according subsequences.

I will show that $\inf_{\{\theta_N\}\subseteq\Theta} \inf_{\{P_N\}:\theta_N\in\Theta_0(P_N)} \lim_{N\to\infty} \Pr\left(\theta_N \in CI_\alpha^i\right) \to 1-\alpha, i = 1,2,3$. These are pointwise limits, but because they are taken over sequences of distributions, the propositions are implied. In particular, the limits apply to sequences s.t. (θ_N, P_N) is least favorable given N. Proofs present two arguments, one for the case that $\{\Delta_N\}$ is small enough and one for the case that $\{\Delta_N\}$ is large. "Small" and "large" is delimited by a_N in propositions 1 and 2 and by c_N, to be defined later, in proposition 3. In either case, any sequence $\{P_N\}$ can be decomposed into two subsequences such that either subsequence is covered by one of the cases.

Proposition 1 Let $\Delta_N \leq a_N$, then $\sqrt{N}\left|\widehat{\Delta} - \Delta_N\right| \xrightarrow{p} 0$ by condition (iii). This furthermore implies that

$$\sqrt{N}\left(\widehat{\theta}_u - \theta_u\right) = \sqrt{N}\left(\widehat{\theta}_l + \widehat{\Delta} - \theta_l - \Delta_N\right) \xrightarrow{p} \sqrt{N}\left(\widehat{\theta}_l - \theta_l\right),$$

which in conjunction with conditions (i)-(ii) implies that $\sigma_u = \sigma_l$, hence by (iv) that $\widehat{\sigma}_u - \widehat{\sigma}_l \xrightarrow{p} 0$. It follows that

$$\Phi\left(c_\alpha^1 + \sqrt{N}\frac{\widehat{\Delta}}{\max\left\{\widehat{\sigma}_l, \widehat{\sigma}_u\right\}}\right) \xrightarrow{p} \Phi\left(c_\alpha^1 + \sqrt{N}\frac{\Delta_N}{\widehat{\sigma}_l}\right),$$

but then the argument can be completed as in lemma 2.

Let $\Delta_N > a_N$, then $\sqrt{N}\Delta_N \to \infty$, hence $\limsup_{N\to\infty} \sqrt{N}\left(\theta_N - \theta_l\right) = \infty$ or $\limsup_{N\to\infty} \sqrt{N}\left(\theta_u - \theta_N\right) = \infty$ or both. Write

$$\begin{aligned}
& \Pr\left(\theta_N \in CI_\alpha^1\right) \\
= \; & \Pr\left(\widehat{\theta}_l - \frac{c_\alpha^1\widehat{\sigma}_l}{\sqrt{N}} \leq \theta_N \leq \widehat{\theta}_l + \widehat{\Delta} + \frac{c_\alpha^1\widehat{\sigma}_u}{\sqrt{N}}\right) \\
= \; & \Pr\left(-c_\alpha^1\widehat{\sigma}_l \leq \sqrt{N}\left(\theta_N - \theta_l\right) + \sqrt{N}\left(\theta_l - \widehat{\theta}_l\right) \leq \sqrt{N}\widehat{\Delta} + c_\alpha^1\widehat{\sigma}_u\right) \\
= \; & \Pr\left(-c_\alpha^1\widehat{\sigma}_l \leq \sqrt{N}\left(\theta_N - \theta_l\right) + \sqrt{N}\left(\theta_l - \widehat{\theta}_l\right)\right) \\
& - \Pr\left(\sqrt{N}\left(\theta_N - \theta_l\right) + \sqrt{N}\left(\theta_l - \widehat{\theta}_l\right) > \sqrt{N}\widehat{\Delta} + c_\alpha^1\widehat{\sigma}_u\right).
\end{aligned}$$

Assume $\limsup_{N\to\infty} \sqrt{N}\left(\theta_N - \theta_l\right) < \infty$. By consistency of $\widehat{\Delta}$, divergence of $\sqrt{N}\Delta_N$ implies divergence in probability of $\sqrt{N}\widehat{\Delta}$. Thus

$$\begin{aligned}
& \Pr\left(\sqrt{N}\left(\theta_N - \theta_l\right) + \sqrt{N}\left(\theta_l - \widehat{\theta}_l\right) > \sqrt{N}\widehat{\Delta} + c_\alpha^1\widehat{\sigma}_u\right) \\
\leq \; & \Pr\left(\sqrt{N}\left(\theta_l - \widehat{\theta}_l\right) > \sqrt{N}\widehat{\Delta} - \sqrt{N}\left(\theta_N - \theta_l\right)\right) \\
\to \; & 0,
\end{aligned}$$

where the convergence statement uses that $c_\alpha^1\widehat{\sigma}_u \geq 0$ by construction and that $\sqrt{N}\left(\theta_l - \widehat{\theta}_l\right)$ converges to a random variable by assumption. It follows that

$$\begin{aligned}
& \lim_{N\to\infty} \Pr\left(\theta_N \in CI_\alpha^1\right) \\
= \; & \lim_{N\to\infty} \Pr\left(-c_\alpha^1\widehat{\sigma}_l \leq \sqrt{N}\left(\theta_N - \theta_l\right) + \sqrt{N}\left(\theta_l - \widehat{\theta}_l\right)\right) \\
\geq \; & \lim_{N\to\infty} \Pr\left(-c_\alpha^1\widehat{\sigma}_l \leq \sqrt{N}\left(\theta_l - \widehat{\theta}_l\right)\right) \\
= \; & 1 - \Phi(c_\alpha^1) \\
\geq \; & 1 - \alpha,
\end{aligned}$$

where the first inequality uses that $\sqrt{N}\left(\theta_N - \theta_l\right) \geq 0$, and the second inequality uses the definition of c_α^1, as well as convergence of $\widehat{\sigma}_l$ and $\sqrt{N}\left(\theta_l - \widehat{\theta}_l\right)/\sigma_l$.

For any subsequence of $\{P_N\}$ s.t. $\sqrt{N}\left(\theta_N - \theta_u\right)$ fails to diverge, the argument is entirely symmetric. If both diverge, coverage probability trivially converges to 1. To see that a coverage probability of $1 - \alpha$ can be attained, consider the case of $\Delta = 0$.

Lemma 3 By assumption 4(i), $\sqrt{N}(\widehat{\Delta}-\Delta) \to N(0,\sigma_{\Delta}^2)$ uniformly in $\mathcal{P}$, where $\sigma_{\Delta}^2 \equiv \sigma_l^2+\sigma_u^2-2\rho\sigma_l\sigma_u$ and where $N(0,0)$ is understood to denote a point mass at zero. Hence, one can fix a sequence $\varepsilon_N \to 0$ s.t. $\sup_{P\in\mathcal{P}:\sigma_{\Delta}^2>0,d\in\mathbb{R}}\left|\Pr(\sqrt{N}(\widehat{\Delta}-\Delta)/\sigma_{\Delta}^2 \leq d)-\Phi(d)\right| \leq \varepsilon_N$ for all N. Fix a nonpositive sequence $\delta_N \to -\infty$ s.t. $\Phi(\gamma\cdot\delta_N) > O(\varepsilon_N)$ for any fixed $\gamma > 0$. This is possible because of well known uniform bounds on the standard normal c.d.f., e.g. $\Phi(\gamma\cdot\delta_N) > -(\gamma\delta_N)^{-1}(2\pi)^{-1/2}\exp(-(\gamma\delta_N)^2/2)$ as $\delta_N \to -\infty$; using this bound, one can verify that $\delta_N = -(\log(-\log\varepsilon_N))^{1/2}$ will do.

Consider sequences P_N s.t. $\Delta_N \leq a_N \equiv -\delta_N N^{-1/2}$. I will show that $\sigma_{\Delta N}^2 \to 0$ uniformly over such sequences, implying assumption 4(iii) with a_N as just defined. Assume this fails, then one can diagonalize across sequences to generate a sequence P_N s.t. $\sigma_{\Delta N}^2 \to 0$ fails. The resulting sequence $\{\sigma_{\Delta N}^2\}$ must have an accumulation point $\sigma_{\Delta\infty}^2 > 0$. Along any subsequence converging to $\sigma_{\Delta\infty}^2$, one would have

$$\begin{aligned}
\Pr\left(\widehat{\theta}_u \leq \widehat{\theta}_l\right) &= \Pr\left(\widehat{\Delta} \leq 0\right) = \Pr(\sqrt{N}(\widehat{\Delta}-\Delta_N)/\sigma_{\Delta\infty}^2 \leq -\sqrt{N}\Delta_N/\sigma_{\Delta\infty}^2) \\
&\geq \Pr(\sqrt{N}(\widehat{\Delta}-\Delta_N)/\sigma_{\Delta\infty}^2 \leq \delta_N/\sigma_{\Delta\infty}^2) \geq \Phi\left(\delta_N/\sigma_{\Delta\infty}^2\right) - \varepsilon_N > 0
\end{aligned}$$

for N large enough, a contradiction.

Proposition 2 A short proof uses asymptotic equivalence to CI_{α}^1, which was essentially shown in the text. The longer argument below shows why CI_{α}^2 will generally have exact nominal size and will also be needed for proposition 3. To begin, let $(\widetilde{c}_l,\widetilde{c}_u)$ fulfil

$$\begin{aligned}
\Pr\left(-\frac{\widetilde{c}_l}{\sigma_l} \leq z_1, \rho z_1 \leq \frac{\widetilde{c}_u+\sqrt{N}\Delta}{\sigma_u}+\sqrt{1-\rho^2}z_2\right) &\geq 1-\alpha \\
\Pr\left(-\frac{\widetilde{c}_l+\sqrt{N}\Delta}{\sigma_l}+\sqrt{1-\rho^2}z_2 \leq \rho z_1, z_1 \leq \frac{\widetilde{c}_u}{\sigma_u}\right) &\geq 1-\alpha
\end{aligned}$$

and write

$$
\begin{aligned}
& \Pr\left(\theta_l \in \left[\widehat{\theta}_l - \frac{\widetilde{c}_l}{\sqrt{N}}, \widehat{\theta}_u + \frac{\widetilde{c}_u}{\sqrt{N}}\right]\right) \\
= \; & \Pr\left(\widehat{\theta}_l - \frac{\widetilde{c}_l}{\sqrt{N}} \leq \theta_l \leq \widehat{\theta}_u + \frac{\widetilde{c}_u}{\sqrt{N}}\right) \\
= \; & \Pr\left(-\frac{\widetilde{c}_l}{\sigma_l} \leq \frac{\sqrt{N}}{\sigma_l}\left(\theta_l - \widehat{\theta}_l\right) \leq \frac{\sqrt{N}}{\sigma_l}\left(\widehat{\theta}_u - \theta_l\right) + \frac{\widetilde{c}_u}{\sigma_l}\right) \\
= \; & \Pr\left(-\frac{\widetilde{c}_l}{\sigma_l} \leq \frac{\sqrt{N}}{\sigma_l}\left(\theta_l - \widehat{\theta}_l\right) \leq \frac{\sqrt{N}}{\sigma_l}\left(\widehat{\theta}_u - \theta_u + \theta_u - \theta_l + \theta_l - \widehat{\theta}_l\right) + \frac{\widetilde{c}_u}{\sigma_l}\right) \\
\to \; & \Pr\left(-\frac{\widetilde{c}_l}{\sigma_l} \leq \frac{\sqrt{N}}{\sigma_l}\left(\theta_l - \widehat{\theta}_l\right) \leq \frac{\sqrt{N}}{\sigma_l}\Delta + \frac{\sqrt{N}}{\sigma_l}\left(1 - \rho\frac{\sigma_u}{\sigma_l}\right)\left(\theta_l - \widehat{\theta}_l\right) + \frac{\sigma_u}{\sigma_l}\sqrt{1-\rho^2}z_2 + \frac{\widetilde{c}_u}{\sigma_l}\right) \\
= \; & \Pr\left(-\frac{\widetilde{c}_l}{\sigma_l} \leq \frac{\sqrt{N}}{\sigma_l}\left(\theta_l - \widehat{\theta}_l\right), \rho\sigma_u\frac{\sqrt{N}}{\sigma_l}\left(\theta_l - \widehat{\theta}_l\right) \leq \frac{\widetilde{c}_u}{\sigma_l} + \frac{\sqrt{N}}{\sigma_l}\Delta + \frac{\sigma_u}{\sigma_l}\sqrt{1-\rho^2}z_2\right) \\
\to \; & \Pr\left(-\frac{\widetilde{c}_l}{\sigma_l} \leq z_1, \rho z_1 \leq \frac{\widetilde{c}_u}{\sigma_u} + \frac{\sqrt{N}}{\sigma_u}\Delta + \sqrt{1-\rho^2}z_2\right) && (11) \\
\geq \; & 1 - \alpha. && (12)
\end{aligned}
$$

Here, the first convergence statement uses that by assumption,

$$
\left(\sqrt{N}\left(\widehat{\theta}_u - \theta_u\right) | \sqrt{N}\left(\theta_l - \widehat{\theta}_l\right)\right) \xrightarrow{d} N\left(-\rho\frac{\sigma_u}{\sigma_l}\sqrt{N}\left(\theta_l - \widehat{\theta}_l\right), \sigma_u^2(1-\rho^2)\right)
$$

uniformly; the algebra also uses that $\sigma_u, \sigma_l \geq \underline{\sigma}$, so that neither can vanish.

As before, for any sequence $\{\Delta_N\}$ s.t. $\Delta_N < a_N$, superefficiency implies that (c_l^2, c_u^2) is consistent for $(\widetilde{c}_l, \widetilde{c}_u)$, so that validity of CI_α^2 at $\{\theta_l, \theta_u\}$ follows. Convexity of the power function over $[\theta_l, \theta_u]$ follows as before. For $\Delta_N \geq a_N$, the argument entirely resembles proposition 1. Notice finally that (12) will bind if (6) binds, implying that CI_α^2 will then have exact nominal size at θ_l. A similar argument applies for θ_u. But (c_l^2, c_u^2) can minimize $(c_l + c_u)$ subject to (6,7) only if at least one of (6,7) binds, implying that CI_α^2 is nominally exact.

Proposition 3 Let $c_N \equiv \left(N^{-1/2}b_N\right)^{1/2}$, thus $N^{1/2}c_N = \left(N^{1/2}b_N\right)^{1/2} \to \infty$, and for parameter sequences s.t. $\Delta_N \geq c_N$, the proof is again as before. For the other case, consider $(\widetilde{c}_l, \widetilde{c}_u)$ as defined in the previous proof. Convergence of (c_l^3, c_u^3) to $(\widetilde{c}_l, \widetilde{c}_u)$ cannot be claimed. However, by uniform convergence of estimators and uniform bounds on (σ_l, σ_u), $\Pr(\widehat{\Delta} \leq b_N)$ is uniformly asymptotically bounded below by $\Phi\left(\sqrt{N}\left(b_N - c_N\right)/2\overline{\sigma}\right) = \Phi\left(\left(N^{1/2}b_N - \left(N^{1/2}b_N\right)^{1/2}\right)/2\overline{\sigma}\right) \to 1$. Hence, $\Delta^* = 0 \leq \Delta$ with probability approaching 1. Expression (11) is easily seen to increase in Δ for every $(\widetilde{c}_l, \widetilde{c}_u)$, hence CI_α^3 is valid (if potentially conservative) at θ_l. The argument for θ_u is similar. (Regarding pointwise exactness of the interval, notice that $c_N \to 0$, so the conservative distortion vanishes under pointwise asymptotics.)

Now consider $\theta_0 \equiv a\theta_l + (1-a)\theta_u$, some $a \in [0,1]$. By assumption,

$$\sqrt{N}\left(a\widehat{\theta}_l + (1-a)\widehat{\theta}_u - \theta_0\right) \xrightarrow{d} N(0, \sigma_a),$$

where $\sigma_a^2 \equiv a^2\sigma_l^2 + (1-a)^2\sigma_u^2 - 2a(1-a)\rho\sigma_l\sigma_u$. Let $\sigma_a > 0$, then

$$\begin{aligned}
& \Pr\left(\theta_0 \in \left[\widehat{\theta}_l - \frac{c_l^3}{\sqrt{N}}, \widehat{\theta}_u + \frac{c_u^3}{\sqrt{N}}\right]\right) \\
= & \Pr\left(\widehat{\theta}_l - \frac{c_l^3}{\sqrt{N}} \leq \theta_0 \leq \widehat{\theta}_u + \frac{c_u^3}{\sqrt{N}}\right) \\
= & \Pr\left(\frac{\sqrt{N}}{\sigma_a}(1-a)\left(\widehat{\theta}_l - \widehat{\theta}_u\right) - \frac{c_l^3}{\sigma_a} \leq \frac{\sqrt{N}}{\sigma_a}(\theta_0 - a\widehat{\theta}_l - (1-a)\widehat{\theta}_u) \leq \frac{\sqrt{N}}{\sigma_a}a\left(\widehat{\theta}_u - \widehat{\theta}_l\right) + \frac{c_u^3}{\sigma_a}\right) \\
= & \Pr\left(\frac{\sqrt{N}}{\sigma_a}(1-a)\left(\Delta - \widehat{\Delta}\right) - \frac{\sqrt{N}}{\sigma_a}(1-a)\Delta - \frac{c_l^3}{\sigma_a}\right. \\
& \left. \leq \frac{\sqrt{N}}{\sigma_a}(\theta_0 - a\widehat{\theta}_l - (1-a)\widehat{\theta}_u) \leq \frac{\sqrt{N}}{\sigma_a}a\left(\widehat{\Delta} - \Delta\right) + \frac{\sqrt{N}}{\sigma_a}a\Delta + \frac{c_u^3}{\sigma_a}\right).
\end{aligned}$$

Consider varying Δ, holding $(\theta_l, \sigma_l, \sigma_u, \rho)$ constant. The cutoff values c_l and c_u depend on Δ only through Δ^*, but recall that $\Pr(\Delta^* = 0) \to 1$. Also, $\frac{\sqrt{N}}{\sigma_a}\left(\widehat{\Delta} - \Delta\right)$ is asymptotically pivotal. Hence, the preceding probability's limit depends on Δ only through $\frac{\sqrt{N}}{\sigma_a}(1-a)\Delta$ and $\frac{\sqrt{N}}{\sigma_a}a\Delta$. As $\frac{\sqrt{N}}{\sigma_a} > 0$, the probability is minimized by setting $\Delta = 0$. In this case, however, $\theta_0 = \theta_l$, for which coverage has already been established. Finally, $\sigma_a = 0$ only if $\sigma_l = \sigma_u$ and $\rho = 1$, in which case $\widehat{\Delta} = \Delta_N$ for large enough N, and the conclusion follows from lemma 2.

B Closed-Form Expressions for (c_l, c_u)

This appendix provides a closed-form equivalent of (6,7). Specifically, these expressions can be written as

$$\begin{aligned}
\int_{-\infty}^{c_l/\widehat{\sigma}_l} \Phi\left(\frac{\widehat{\rho}}{\sqrt{1-\widehat{\rho}^2}}z + \frac{c_u + \sqrt{N}\widehat{\Delta}}{\widehat{\sigma}_u\sqrt{1-\widehat{\rho}^2}}\right) d\Phi(z) \quad &\geq \quad 1-\alpha \\
\int_{-\infty}^{c_u/\widehat{\sigma}_u} \Phi\left(\frac{\widehat{\rho}}{\sqrt{1-\widehat{\rho}^2}}z + \frac{c_l + \sqrt{N}\widehat{\Delta}}{\widehat{\sigma}_l\sqrt{1-\widehat{\rho}^2}}\right) d\Phi(z) \quad &\geq \quad 1-\alpha
\end{aligned}$$

if $-1 < \widehat{\rho} < 1$,

$$\begin{aligned}
\Phi\left(\frac{c_l}{\widehat{\sigma}_l}\right) - \Phi\left(-\frac{c_u + \sqrt{N}\widehat{\Delta}}{\widehat{\sigma}_u}\right) \quad &\geq \quad 1-\alpha \\
\Phi\left(\frac{c_u}{\widehat{\sigma}_u}\right) - \Phi\left(-\frac{c_l + \sqrt{N}\widehat{\Delta}}{\widehat{\sigma}_l}\right) \quad &\geq \quad 1-\alpha
\end{aligned}$$

if $\widehat{\rho} = 1$, and

$$\begin{aligned} \Phi\left(\min\left\{\frac{c_l}{\widehat{\sigma}_l}, \frac{c_u + \sqrt{N}\widehat{\Delta}}{\widehat{\sigma}_u}\right\}\right) &\geq 1-\alpha \\ \Phi\left(\min\left\{\frac{c_u}{\widehat{\sigma}_u}, \frac{c_l + \sqrt{N}\widehat{\Delta}}{\widehat{\sigma}_l}\right\}\right) &\geq 1-\alpha, \end{aligned}$$

implying that $\Phi\left(\frac{c_l}{\widehat{\sigma}_l}\right) = \Phi\left(\frac{c_u}{\widehat{\sigma}_u}\right) = 1-\alpha$ at the minimization problem's solution, if $\widehat{\rho} = -1$. There is no discontinuity at the limit because $\Phi\left(\frac{\widehat{\rho}}{\sqrt{1-\widehat{\rho}^2}} z + \frac{c_u + \sqrt{N}\widehat{\Delta}}{\widehat{\sigma}_u \sqrt{1-\widehat{\rho}^2}}\right) \to \mathbb{I}\left\{z \geq [\leq] \frac{c_u + \sqrt{N}\widehat{\Delta}}{\widehat{\sigma}_u}\right\}$ as $\widehat{\rho} \to 1[-1]$. It follows that (6,7), and similarly (8,9), can be evaluated without simulation.

References

[1] Andrews, D.W.K. (2000): "Inconsistency of the Bootstrap when a Parameter is on the Boundary of the Parameter Space," *Econometrica* 68: 399-405.

[2] Andrews, D.W.K. and P. Guggenberger (2007): "Validity of Subsampling and 'Plug-in Asymptotic' Inference for Parameters Defined by Moment Inequalities," mimeo, Cowles Foundation.

[3] Andrews, D. and G. Soares (2007): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," mimeo, Cowles Foundation.

[4] Bugni, F.A. (2007): "Bootstrap Inference in Partially Identified Models," mimeo, Northwestern University.

[5] Chernozhukov, V., H. Hong, and E.T. Tamer (2007): "Parameter Set Inference in a Class of Econometric Models," *Econometrica* 75: 1243-1284.

[6] Fan, Y. and S.S. Park (2007): "Confidence Intervals for Some Partially Identified Parameters," mimeo, Vanderbilt University.

[7] Haile, P. and E.T. Tamer (2003): "Inference with an Incomplete Model of English Auctions," *Journal of Political Economy* 111: 1-50.

[8] Honoré, B.E. and E.T. Tamer (2006): "Bounds on Parameters in Panel Dynamic Discrete Choice Models," *Econometrica* 74: 611–629.

[9] Horowitz, J. and C.F. Manski (2000): "Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data," *Journal of the American Statistical Association* 95: 77–84.

[10] Ikeda, S. and Y. Nonaka (1983): "Uniform Asymptotic Normality of a Set of Increasing Number of Sample Quantiles," *Annals of the Institute of Statistical Mathematics* 35: 329-341.

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

[12] Leeb, H. and B. Pötscher (2005): "Model Selection and Inference: Facts and Fiction," *Econometric Theory* 21: 21-59.

[13] — (2008): "Sparse Estimators and the Oracle Property, or the Return of Hodges' Estimator," *Journal of Econometrics* 142: 201-211.

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

[15] Pakes, A., J. Porter, K. Ho, and J. Ishii (2006): "Moment Inequalities and their Application," mimeo, Harvard University.

[16] Romano, J.P. and A.M. Shaikh (2008): "Inference for Identifiable Parameters in Partially Identified Econometric Models," forthcoming in the *Journal of Statistical Planning and Inference.*

[17] Stoye, J. (2007): "Bounds on Generalized Linear Predictors with Incomplete Outcome Data," *Reliable Computing* 13: 293-302.

[18] Woutersen, T. (2006): "A Simple Way to Calculate Confidence Intervals for Partially Identified Parameters," mimeo, Johns Hopkins University.