Bootstrap-Based Inference for Cube Root Consistent Estimators

Matias D. Cattaneo‡ Michael Jansson† Kenichi Nagasawa§

May 31, 2018

Abstract

This note proposes a consistent bootstrap-based distributional approximation for cube root consistent estimators such as the maximum score estimator of Manski (1975) and the conditional maximum score estimator of Honoré and Kyriazidou (2000). For estimators of this kind, the standard nonparametric bootstrap is inconsistent. Our method restores consistency of the nonparametric bootstrap by altering the shape of the criterion function defining the estimator whose distribution we seek to approximate. This modification leads to a generic and easy-to-implement resampling method for inference that is conceptually distinct from other available distributional approximations, and can also be used in other related settings such as for the isotonic density estimator of Grenander (1956).

Keywords: cube root asymptotics, bootstrapping, maximum score estimation.

JEL: C12, C14, C21.

*For comments and suggestions, we are grateful to the handling coeditor, three referees, Mehmet Caner, Kei Hirano, Bo Honoré, Guido Kuersteiner, Andres Santos, and seminar participants at University of Aarhus, the 2017 Tinbergen Institute Conference, the 2017 California Econometrics Conference, the 2018 Workshop on Advanced Econometrics at University of Kansas, and the 2018 DAEiNA meeting. Cattaneo gratefully acknowledges financial support from the National Science Foundation through grant SES-1459931, and Jansson gratefully acknowledges financial support from the National Science Foundation through grant SES-1459967 and the research support of CREATE (funded by the Danish National Research Foundation under grant no. DNRF78).

‡Department of Economics and Department of Statistics, University of Michigan.

†Department of Economics, University of California at Berkeley and CREATE.

§Department of Economics, University of Michigan.
1 Introduction

In a seminal paper, Kim and Pollard (1990) studied a class of $M$-estimators exhibiting cube root asymptotics. These estimators not only have a non-standard rate of convergence, but also have the property that rather than being Gaussian their limiting distributions are of Chernoff (1964) type; that is, the limiting distribution is that of the maximizer of a Gaussian process. In fact, in leading examples of cube root consistent estimators such as the maximum score estimator of Manski (1975), the covariance kernel of the Gaussian process characterizing the limiting distribution depends on an infinite-dimensional nuisance parameter. From the perspective of inference, this feature of the limiting distribution represents a nontrivial complication relative to the conventional asymptotically normal case, where the limiting distribution is known up to the value of a finite-dimensional nuisance parameter (namely, the covariance matrix of the limiting distribution). In particular, the dependence of the limiting distribution on an infinite-dimensional nuisance parameter implies that resampling-based distributional approximations seem to offer the most attractive approach to inference in estimation problems exhibiting cube root asymptotics. The purpose of this note is to propose an easy-to-implement bootstrap-based distributional approximation applicable in such cases.

As does the familiar nonparametric bootstrap, the method proposed herein employs bootstrap samples of size $n$ from the empirical distribution function. But unlike the nonparametric bootstrap, which is inconsistent in general (e.g., Abrevaya and Huang, 2005; Léger and MacGibbon, 2006), our method offers a consistent distributional approximation for cube root consistent estimators and therefore has the advantage that it can be used to construct asymptotically valid inference procedures. Consistency is achieved by altering the shape of the criterion function defining the estimator whose distribution we seek to approximate. Heuristically, the method is designed to ensure that the bootstrap version of a certain empirical process has a mean resembling the large sample version of its population counterpart. The latter is quadratic in the problems we study, and known up to the value of a certain matrix. As a consequence, the only ingredient needed to implement the proposed “reshapement” of the objective function is a consistent estimator of the unknown matrix entering the quadratic mean of the empirical process. Such estimators turn out to be generically available and easy to compute.
The note proceeds as follows. Section 2 is heuristic in nature and serves the purpose of outlining the main idea underlying our approach in the $M$-estimation setting of Kim and Pollard (1990). Section 3 then makes the heuristics of Section 2 rigorous in a more general setting where the $M$-estimation problem is formed using a possibly $n$-varying (observation specific) objective function, as recently studied by Seo and Otsu (2018). Section 4 discusses two examples covered by our general results, namely the maximum score estimator of Manski (1975) and the conditional maximum score estimator of Honoré and Kyriazidou (2000). Further discussion of our results is provided in Section 5. Finally, all derivations and proofs have been collected in the supplemental appendix, where an extension to the case of the isotonic density estimator of Grenander (1956) is also given.

2 Heuristics

Suppose $\theta_0 \in \Theta \subseteq \mathbb{R}^d$ is an estimand admitting the characterization

$$\theta_0 = \arg\max_{\theta \in \Theta} M_0(\theta), \quad M_0(\theta) = \mathbb{E}[m_0(z, \theta)],$$

where $m_0$ is a known function, and where $z$ is a random vector of which a random sample $z_1, \ldots, z_n$ is available. Studying estimation problems of this kind for non-smooth $m_0$, Kim and Pollard (1990) gave conditions under which the $M$-estimator

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} \hat{M}_n(\theta), \quad \hat{M}_n(\theta) = \frac{1}{n} \sum_{i=1}^n m_0(z_i, \theta),$$

exhibits cube root asymptotics:

$$\sqrt[n]{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow \arg\max_{s \in \mathbb{R}^d} \{Q_0(s) + G_0(s)\},$$

where $\rightsquigarrow$ denotes weak convergence, $Q_0(s)$ is a quadratic form given by

$$Q_0(s) = -\frac{1}{2} s^t V_0 s, \quad V_0 = -\frac{\partial^2}{\partial \theta \partial \theta} M_0(\theta_0),$$

and $G_0$ is a non-degenerate zero-mean Gaussian process with $G_0(0) = 0$.

Whereas the matrix $V_0$ governing the shape of $Q_0$ is finite-dimensional, the covariance kernel of
\( \mathcal{G}_0 \) in (2) typically involves infinite-dimensional unknown quantities. As a consequence, the limiting distribution of \( \hat{\theta}_n \) tends to be more difficult to approximate than Gaussian distributions, implying in turn that basing inference on \( \hat{\theta}_n \) is more challenging under cube root asymptotics than in the more familiar case where \( \hat{\theta}_n \) is (\( \sqrt{\cdot} \)-consistent and) asymptotically normally distributed.

As a candidate method of approximating the distribution of \( \hat{\theta}_n \), consider the nonparametric bootstrap. To describe it, let \( z^*_1, \ldots, z^*_n \) denote a random sample from the empirical distribution of \( z_1, \ldots, z_n \) and let the natural bootstrap analogue of \( \hat{\theta}_n \) be denoted by

\[
\hat{\theta}^*_n = \arg\max_{\theta \in \Theta} \hat{M}_n^*(\theta), \quad \hat{M}_n^*(\theta) = \frac{1}{n} \sum_{i=1}^n m_0(z^*_{i,n}, \theta).
\]

Then the nonparametric bootstrap estimator of \( \mathbb{P}[\hat{\theta}_n - \theta_0 \leq \cdot] \) is given by \( \mathbb{P}^*[\hat{\theta}^*_n - \hat{\theta}_n \leq \cdot] \), where \( \mathbb{P}^* \) denotes a probability computed under the bootstrap distribution conditional on the data. As is well documented, however, this estimator is inconsistent under cube root asymptotics (e.g., Abrevaya and Huang, 2005; Léger and MacGibbon, 2006).

For the purpose of giving a heuristic, yet constructive, explanation of the inconsistency of the nonparametric bootstrap, it is helpful to recall that a proof of (2) can be based on the representation

\[
\sqrt{n}(\hat{\theta}_n - \theta_0) = \arg\max_{s \in \mathbb{R}^d} \{Q_n(s) + \hat{G}_n(s)\}, \quad (3)
\]

where (for \( s \) such that \( \theta_0 + sn^{-1/3} \in \Theta \))

\[
\hat{G}_n(s) = n^{2/3}[\hat{M}_n(\theta_0 + sn^{-1/3}) - \hat{M}_n(\theta_0)] - M_0(\theta_0 + sn^{-1/3}) + M_0(\theta_0)
\]

is a zero-mean random process, while

\[
Q_n(s) = n^{2/3}[M_0(\theta_0 + sn^{-1/3}) - M_0(\theta_0)]
\]

is a non-random function that is correctly centered in the sense that \( \arg\max_{s \in \mathbb{R}^d} Q_n(s) = 0 \). In cases where \( m_0 \) is non-smooth but \( M_0 \) is smooth, \( Q_n \) and \( \hat{G}_n \) are usually asymptotically quadratic.
and asymptotically Gaussian, respectively, in the sense that

$$Q_n(s) \to Q_0(s) \tag{4}$$

and

$$\hat{G}_n(s) \Rightarrow G_0(s). \tag{5}$$

Under regularity conditions ensuring among other things that the convergence in (4) and (5) is suitably uniform in $s$, result (2) then follows from an application of a continuous mapping-type theorem for the argmax functional to the representation in (3).

Similarly to (3), the bootstrap analogue of $\hat{\theta}_n$ admits a representation of the form

$$\frac{1}{\sqrt{n}}(\hat{\theta}^*_n - \hat{\theta}_n) = \arg\max_{s \in \mathbb{R}^d} \{\hat{Q}_n(s) + \hat{G}^*_n(s)\},$$

where (for $s$ such that $\hat{\theta}_n + sn^{-1/3} \in \Theta$)

$$\hat{G}^*_n(s) = n^{2/3}[\hat{M}^*_n(\hat{\theta}_n + sn^{-1/3}) - \hat{M}_n(\hat{\theta}_n)] + M_n(\hat{\theta}_n),$$

and

$$\hat{Q}_n(s) = n^{2/3}[\hat{M}_n(\hat{\theta}_n + sn^{-1/3}) - \hat{M}_n(\hat{\theta}_n)].$$

Under mild conditions, $\hat{G}^*_n$ satisfies the following bootstrap counterpart of (5):

$$\hat{G}^*_n(s) \Rightarrow_P G_0(s), \tag{6}$$

where $\Rightarrow_P$ denotes conditional weak convergence in probability (defined as van der Vaart and Wellner, 1996, Section 2.9). On the other hand, although $\hat{Q}_n$ is non-random under the bootstrap distribution and satisfies argmax$_{s \in \mathbb{R}^d} \hat{Q}_n(s) = 0$, it turns out that $\hat{Q}_n(s) \Rightarrow_P Q_0(s)$ in general. In other words, the natural bootstrap counterpart of (4) typically fails and, as a partial consequence, so does the natural bootstrap counterpart of (2); that is, $\frac{1}{\sqrt{n}}(\hat{\theta}^*_n - \hat{\theta}_n) \not\Rightarrow_P \arg\max_{s \in \mathbb{R}^d} \{Q_0(s) + G_0(s)\}$ in general.

To the extent that the implied inconsistency of the bootstrap can be attributed to the fact that
the shape of $\hat{Q}_n$ fails to replicate that of $Q_n$, it seems plausible that a consistent bootstrap-based distributional approximation can be obtained by basing the approximation on

$$\tilde{\theta}_n^* = \arg\max_{\theta \in \Theta} M_n^*(\theta), \quad \tilde{M}_n^*(\theta) = \frac{1}{n} \sum_{i=1}^{n} \tilde{m}_n(z_{i,n}^*, \theta),$$

where $\tilde{m}_n$ is a suitably “reshaped” version of $m_0$ satisfying two properties. First, $\hat{Q}_n$ should be asymptotically quadratic, where $\hat{Q}_n$ is the counterpart of $\tilde{Q}_n$ associated with $\tilde{m}_n$:

$$\hat{Q}_n(s) = n^{2/3}[\tilde{M}_n(\hat{\theta}_n + sn^{-1/3}) - \tilde{M}_n(\hat{\theta}_n)], \quad \tilde{M}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \tilde{m}_n(z_i, \theta).$$

Second, $\hat{G}_n^*$ should be asymptotically equivalent to $\tilde{G}_n^*$, where

$$\hat{G}_n^*(s) = n^{2/3}[\tilde{M}_n^*(\hat{\theta}_n + sn^{-1/3}) - \tilde{M}_n^*(\hat{\theta}_n) - \tilde{M}_n(\hat{\theta}_n + sn^{-1/3}) + \tilde{M}_n(\hat{\theta}_n)],$$

is the counterpart of $\tilde{G}_n^*$ associated with $\tilde{m}_n$.

Accordingly, let

$$\tilde{m}_n(z, \theta) = m_0(z, \theta) - \tilde{M}_n(\theta) - \frac{1}{2}(\theta - \hat{\theta}_n)'\tilde{V}_n(\theta - \hat{\theta}_n),$$

where $\tilde{V}_n$ is an estimator of $V_0$. Then

$$\sqrt{n}(\tilde{\theta}_n^* - \hat{\theta}_n) = \arg\max_{s \in \mathbb{R}^d}\{\hat{Q}_n(s) + \tilde{G}_n^*(s)\},$$

where, by construction, $\hat{G}_n^*(s) = \tilde{G}_n^*(s)$ and $\hat{Q}_n(s) = -s'\tilde{V}_n s/2$. Because $\tilde{G}_n^* = \tilde{G}_n^*$, $\tilde{G}_n^*(s) \Rightarrow \mathbb{P} G_0(s)$ whenever (6) holds. In addition, $\hat{Q}_n(s) \Rightarrow \mathbb{Q}_0(s)$ provided $\tilde{V}_n \Rightarrow \mathbb{P} V_0$. As a consequence, it seems plausible that $\mathbb{P}^*[\tilde{\theta}_n^* - \theta_0 \leq \cdot]$ is a consistent estimator of $\mathbb{P}[\hat{\theta}_n - \theta_0 \leq \cdot]$ if $\tilde{V}_n \Rightarrow \mathbb{P} V_0$. 

5
3 Main Result

When making the heuristics of Section 2 precise, it is of interest to consider the more general situation where the estimator $\hat{\theta}_n$ is an approximate maximizer (with respect to $\Theta \subseteq \mathbb{R}^d$) of

$$\hat{M}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} m_n(z_i, \theta),$$

where $m_n$ is a known function, and where $z_1, \ldots, z_n$ is a random sample of a random vector $z$. This formulation of $\hat{M}_n$, which reduces to that of Section 2 when $m_n$ does not depend on $n$, is adopted in order to cover certain estimation problems where, rather than admitting a characterization of the form (1), the estimand $\theta_0$ admits the characterization

$$\theta_0 = \arg\max_{\theta \in \Theta} M_0(\theta), \quad M_0(\theta) = \lim_{n \to \infty} M_n(\theta), \quad M_n(\theta) = \mathbb{E}[m_n(z, \theta)].$$

In other words, the setting considered in this section is one where $\hat{\theta}_n$ approximately maximizes a function $\hat{M}_n$ whose population counterpart $M_n$ can be interpreted as a regularization (in the sense of Bickel and Li, 2006) of a function $M_0$ whose maximizer $\theta_0$ is the object of interest. The additional flexibility (relative to the more traditional $M$-estimation setting of Section 2) afforded by the present setting is attractive because it allows us to formulate results that cover local $M$-estimators such as the conditional maximum score estimator of Honoré and Kyriazidou (2000).

Studying this type of setting, Seo and Otsu (2018) gave conditions under which $\hat{\theta}_n$ converges at a rate equal to the cube root of the “effective” sample size and has a limiting distribution of Chernoff (1964) type. Analogous conclusions will be drawn below, albeit under slightly different conditions.

For any $n$ and any $\delta > 0$, define $M_n = \{m_n(\cdot, \theta) : \theta \in \Theta\}$, $m_n(z) = \sup_{m \in M_n} |m(z)|$, $\Theta_0^\delta = \{\theta \in \Theta : ||\theta - \theta_0|| \leq \delta\}$, $D_n^\delta = \{m_n(\cdot, \theta) - m_n(\cdot, \theta_0) : \theta \in \Theta_0^\delta\}$, and $d_n^\delta(z) = \sup_{d \in D_n^\delta} |d(z)|$.

**Condition CRA (Cube Root Asymptotics)** For a positive $q_n$ with $r_n = \sqrt[3]{nq_n} \to \infty$, the following are satisfied:

(i) $\{M_n : n \geq 1\}$ is uniformly manageable for the envelopes $\tilde{m}_n$ and $q_n \mathbb{E}[\tilde{m}_n(z)^2] = O(1)$.

Also, $\sup_{\theta \in \Theta} |M_n(\theta) - M_0(\theta)| = o(1)$ and, for every $\delta > 0$, $\sup_{\theta \in \Theta_0^\delta} M_0(\theta) < M_0(\theta_0)$.

(ii) $\theta_0$ is an interior point of $\Theta$ and, for some $\delta > 0$, $M_0$ and $M_n$ are twice continuously differentiable on $\Theta_0^\delta$ and $\sup_{\theta \in \Theta_0^\delta} \|\partial^2[M_n(\theta) - M_0(\theta)]/\partial \theta \partial \theta'\| = o(1)$.
Also, \( r_n\|\partial M_n(\theta_0)/\partial \theta \| = o(1) \) and \( \mathbf{V}_0 = -\partial^2 M_0(\theta_0)/\partial \theta \partial \theta' \) is positive definite.

(iii) For some \( \delta > 0 \), \( \{D_n^{\delta'} : n \geq 1, 0 < \delta' \leq \delta \} \) is uniformly manageable for the envelopes \( \bar{d}_n^{\delta'} \) and \( q_n \sup_{0 < \delta' \leq \delta} \mathbb{E}[\bar{d}_n^{\delta'}(z)^2/\delta'] = O(1) \).

(iv) For every positive \( \delta_n \) with \( \delta_n = O(r_n^{-1}) \), \( q_n^2 \mathbb{E}[\bar{d}_n^{3}(z)^3] + q_n^3 r_n^{-1} \mathbb{E}[\bar{d}_n^{4}(z)^4] = o(1) \), and, for all \( s, t \in \mathbb{R}^d \) and for some \( C_0 \) with \( C_0(s,s) + C_0(t,t) - 2C_0(s,t) > 0 \) for \( s \neq t \),

\[
\sup_{\theta \in \Theta_n^\delta} q_n \mathbb{E}\{m_n(z, \theta + \delta_n s) - m_n(z, \theta) \{m_n(z, \theta + \delta_n t) - m_n(z, \theta) \}/\delta_n - C_0(s,t)\} = o(1).
\]

(v) For every positive \( \delta_n \) with \( \delta_n = O(r_n^{-1}) \),

\[
\lim_{C \to \infty} \lim_{n \to \infty} \sup_{0 < \delta \leq \delta_n} q_n \mathbb{E}[\mathbb{I}\{q_n \bar{d}_n^{3}(z) > C\} \bar{d}_n^{3}(z)^2/\delta] = 0
\]

and \( \sup_{\theta, \theta' \in \Theta_0^\delta} \mathbb{E}[\|m_n(z, \theta) - m_n(z, \theta')\|/\|\theta - \theta'\|] = O(1) \).

To interpret Condition CRA, consider first the benchmark case where \( m_n = m_0 \) and \( q_n = 1 \). In this case, the condition is similar to, but slightly stronger than, assumptions (ii)-(vii) of the main theorem of Kim and Pollard (1990), to which the reader is referred for a definition of the term (uniformly) manageable. The most notable difference between Condition CRA and the assumptions employed by Kim and Pollard (1990) is probably that part (iv) contains assumptions about moments of orders three and four, that the displayed part of part (iv) is a locally uniform (with respect to \( \theta \) near \( \theta_0 \)) version of its counterpart in Kim and Pollard (1990), and that (i) can be thought of as replacing the high level condition \( \hat{\theta}_n \to \mathbb{P} \theta_0 \) of Kim and Pollard (1990) with more primitive conditions that imply it for approximate \( M \)-estimators. In all three cases, the purpose of strengthening the assumptions of Kim and Pollard (1990) is to be able to analyze the bootstrap.

More generally, Condition CRA can be interpreted as an \( n \)-varying version of a suitably (for the purpose of analyzing the bootstrap) strengthened version of the assumptions of Kim and Pollard (1990). The differences between Condition CRA and the \( i.i.d. \) version of the conditions in Seo and Otsu (2018) seem mostly technical in nature, but for completeness we highlight two differences here. First, to handle dependent data Seo and Otsu (2018) control the complexity of various function classes using the concept of bracketing entropy. In contrast, because we assume random sampling
we can follow Kim and Pollard (1990) and obtain maximal inequalities using bounds on uniform entropy numbers implied by the concept of (uniform) manageability. Second, whereas Seo and Otsu (2018) controls the bias of $\hat{\theta}_n$ through a locally uniform bound on $M_n - M_0$, Condition CRA controls the bias through the first and second derivatives of $M_n - M_0$.

Under Condition CRA, the effective sample size is given by $n q$. In perfect agreement with Seo and Otsu (2018), it turns out that if $\hat{\theta}_n$ is an approximate maximizer of $M_n$,

$$ r_n(\hat{\theta}_n - \theta_0) \rightsquigarrow \underset{s \in \mathbb{R}^d}{\text{argmax}} \{ Q_0(s) + G_0(s) \}, $$

where $Q_0(s) = -s'V_0s/2$, and where $G_0$ is a zero-mean Gaussian process with $G_0(0) = 0$ and covariance kernel $C_0$. The heuristics of the previous section are rate-adaptive, so once again it stands to reason that if $\tilde{V}_n$ is a consistent estimator of $V_0$, then a consistent distributional approximation can be based on an approximate maximizer $\tilde{\theta}_n$ of

$$ \tilde{M}_n^*(\theta) = \frac{1}{n} \sum_{i=1}^{n} \tilde{m}_n(z_{i,n}, \theta), \quad \tilde{m}_n(z, \theta) = m_n(z, \theta) - \tilde{M}_n(\theta) - \frac{1}{2}(\theta - \hat{\theta}_n)'\tilde{V}_n(\theta - \hat{\theta}_n), $$

where $z_{1,n}^*, \ldots, z_{n,n}^*$ is a random sample from the empirical distribution of $z_1, \ldots, z_n$.

Following van der Vaart (1998, Chapter 23), we say that our bootstrap-based estimator of the distribution of $r_n(\hat{\theta}_n - \theta_0)$ is consistent if

$$ \sup_{t \in \mathbb{R}^d} \left| \mathbb{P}[r_n(\tilde{\theta}_n) \leq t] - \mathbb{P}[r_n(\hat{\theta}_n - \theta_0) \leq t] \right| \to_{\mathbb{P}} 0. $$

Because the limiting distribution in (7) is continuous, this consistency property implies consistency of bootstrap-based confidence intervals (e.g., van der Vaart, 1998, Lemma 23.3). Moreover, continuity of the limiting distribution implies that (8) holds provided the estimator $\tilde{\theta}_n^*$ satisfies the following bootstrap counterpart of (7):

$$ r_n(\tilde{\theta}_n^* - \hat{\theta}_n) \rightsquigarrow \mathbb{P} \underset{s \in \mathbb{R}^d}{\text{argmax}} \{ Q_0(s) + G_0(s) \}. $$

Theorem 1, our main result, gives sufficient conditions for this to occur.
Theorem 1 Suppose Condition CRA holds. If \( \tilde{V}_n \to P V_0 \) and if

\[
M_n(\hat{\theta}_n) \geq \sup_{\theta \in \Theta} M_n(\theta) - o_P(r_n^{-2}), \quad M_n^*(\hat{\theta}_n^*) \geq \sup_{\theta \in \Theta} M_n^*(\theta) - o_P(r_n^{-2}),
\]

then (8) holds.

To implement the bootstrap-based approximation to the distribution of \( r_n(\hat{\theta}_n - \theta_0) \), only a consistent estimator of \( V_0 \) is needed. A generic estimator based on numerical derivatives is \( \tilde{V}_n^{\text{ND}} \), the matrix whose element \((k,l)\) is given by

\[
\tilde{V}_{n,kl}^{\text{ND}} = \frac{1}{4\epsilon_n^2}\left[ M_n(\hat{\theta}_n + e_k\epsilon_n + e_l\epsilon_n) - M_n(\hat{\theta}_n + e_k\epsilon_n - e_l\epsilon_n) \\
- M_n(\hat{\theta}_n - e_k\epsilon_n + e_l\epsilon_n) + M_n(\hat{\theta}_n - e_k\epsilon_n - e_l\epsilon_n) \right],
\]

where \( e_k \) is the \( k \)-th unit vector in \( \mathbb{R}^d \) and where \( \epsilon_n \) is a positive tuning parameter. Conditions under which this estimator is consistent are given in the following lemma.

Lemma 1 Suppose Condition CRA holds and that \( r_n(\hat{\theta}_n - \theta_0) = O_P(1) \). If \( \epsilon_n \to 0 \) and if \( r_n \epsilon_n \to \infty \), then \( \tilde{V}_n^{\text{ND}} \to P V_0 \).

Plausibility of the high-level condition \( r_n(\hat{\theta}_n - \theta_0) = O_P(1) \) follows from (7). More generally, if only consistency is assumed on the part of \( \hat{\theta}_n \), then \( \tilde{V}_n^{\text{ND}} \to P V_0 \) holds provided \( \epsilon_n \to 0 \) and \( ||\hat{\theta}_n - \theta_0||/\epsilon_n \to P 0 \). The proof of the lemma goes beyond consistency and develops a Nagar-type mean squared error (MSE) expansion for \( \tilde{V}_n^{\text{ND}} \) under the additional assumption that \( M_0 \) is four times continuously differentiable near \( \theta_0 \). The approximate MSE can be minimized by choosing \( \epsilon_n \) proportional to \( r_n^{-3/7} \), the optimal factor of proportionality being a functional of the covariance kernel \( C_0 \) and the fourth order derivatives of \( M_0 \) evaluated at \( \theta_0 \). For details, see the supplemental appendix (Section A.3, Theorem A.3), which also contains a brief discussion of alternative generic estimators of \( V_0 \).

We close this section by summarizing the algorithm for our proposed bootstrap-based distributional approximation.

**Bootstrap-Based Approximation** Let the notation and conditions in Theorem 1 hold.

**Step 1.** Compute \( \hat{\theta}_n \) and \( \tilde{V}_n \) using the original sample \( z_1, \ldots, z_n \).
Step 2. Compute $\tilde{M}_n^*(\theta)$ and let $\tilde{\theta}_n^*$ be an approximate maximizer thereof, both constructed using the nonparametric bootstrap sample $z_{1,n}^*, \ldots, z_{n,n}^*$. (Note that $\tilde{V}_n$ is not recomputed at this step.)

Step 3. Repeat Step 2 $B$ times, and then compute the empirical distribution of $r_n(\tilde{\theta}_n^* - \hat{\theta}_n)$.

4 Examples

This section briefly discusses two econometric examples covered by our main result, namely the maximum score estimator of Manski (1975) and the conditional maximum score estimator of Honoré and Kyriazidou (2000). From the perspective of this paper, the main difference between these examples is that (only) the latter corresponds to a situation where $m_n$ depends on $n$.

4.1 The Maximum Score Estimator

To describe a version of the maximum score estimator of Manski (1975), suppose $z_1, \ldots, z_n$ is a random sample of $z = (y, x')'$ generated by the binary response model

$$y = \mathbb{I}(\beta_0'x + u \geq 0), \quad F_{u|x}(0|x) = 1/2,$$

where $\beta_0 \in \mathbb{R}^{d+1}$ is an unknown parameter of interest, $x \in \mathbb{R}^{d+1}$ is a vector of covariates, and $F_{u|x}(\cdot|x)$ is the conditional cumulative distribution function of the unobserved error term $u$ given $x$. Following Abrevaya and Huang (2005), we employ the parameterization $\beta_0 = (1, \theta_0')'$, where $\theta_0 \in \mathbb{R}^d$ is unknown. In other words, we assume that the first element of $\beta_0$ is positive and then normalize the (unidentified) scale of $\beta_0$ by setting its first element equal to unity. Partitioning $x$ conformably with $\theta_0$ as $x = (x_1, x_2')'$, a maximum score estimator of $\theta_0 \in \Theta \subseteq \mathbb{R}^d$ is any $\theta_n^{\text{MS}}$ approximately maximizing $\tilde{M}_n$ for

$$m_n(z, \theta) = m^{\text{MS}}(z, \theta) = (2y - 1)\mathbb{I}(x_1 + \theta'x_2 \geq 0).$$

Regarded as a member of the class of $M$-estimators exhibiting cube root asymptotics, the maximum score estimator is representative in a couple of respects. First, under easy-to-interpret primitive conditions the estimator is covered by the results of Section 3. In particular, because
does not depend on \( n \) we can set \( q_n = 1 \) when formulating primitive conditions for Condition CRA; for details, see the supplemental appendix (Section A.4). Second, in addition to the generic estimator \( \tilde{V}_{n}^{\text{ND}} \) discussed above, the maximum score estimator admits an example-specific consistent estimator of the \( V_0 \) associated with it. Let

\[
\tilde{V}_{n}^{\text{MS}} = -\frac{1}{n} \sum_{i=1}^{n} (2y_i - 1) \hat{K}_n(x_{1i} + \theta'x_{2i})x_{2i}x_{2i}' \bigg|_{\theta = \hat{\theta}_{n}^{\text{MS}}},
\]

where, for a kernel function \( K \) and a bandwidth \( h_n \), \( \hat{K}_n(u) = d\hat{K}_n(u)/du \) and \( K_n(u) = K(u/h_n)/h_n \). As defined, \( \tilde{V}_{n}^{\text{MS}} \) is simply minus the second derivative, evaluated at \( \theta = \hat{\theta}_{n}^{\text{MS}} \), of the criterion function associated with the smoothed maximum score estimator of Horowitz (1992). The estimator \( \tilde{V}_{n}^{\text{MS}} \) is consistent under mild conditions on \( h_n \) and \( K \); for details, see the supplemental appendix (Section A.4, Lemma MS), which also reports the results of a Monte Carlo experiment evaluating the performance of our proposed inference procedure.

### 4.2 The Conditional Maximum Score Estimator

Consider the dynamic binary response model

\[
Y_t = I(\beta_0'X_t + \gamma_0Y_{t-1} + \alpha + u_t \geq 0), \quad t = 1, 2, 3,
\]

where \( \beta_0 \in \mathbb{R}^d \) and \( \gamma_0 \in \mathbb{R} \) are unknown parameters of interest, \( \alpha \) is an unobserved (time-invariant) individual-specific effect, and \( u_t \) is an unobserved error term. Honoré and Kyriazidou (2000) analyzed this model and gave conditions under which \( \beta_0 \) and \( \gamma_0 \) are identified up to scale. Assuming these conditions hold and assuming the first element of \( \beta_0 \) is positive, we can normalize that element to unity and employ the parameterization \( (\beta_0', \gamma_0)' = (1, \theta_0)' \), where \( \theta_0 \in \mathbb{R}^d \) is unknown.

To describe a version of the conditional maximum score estimator of Honoré and Kyriazidou (2000), partition \( X_t \) after the first element as \( X_t = (X_{1t}, X_{2t}')' \) and define \( z = (y, x_1, x_2', w')' \), where \( y = Y_2 - Y_1, x_1 = X_{12} - X_{11}, x_2 = (X_{22} - X_{21})', Y_3 - Y_0)' \), and \( w = X_2 - X_3 \). Assuming \( z_1, \ldots, z_n \) is a random sample of \( z \), a conditional maximum score estimator of \( \theta_0 \in \Theta \subseteq \mathbb{R}^d \) is any \( \hat{\theta}_{n}^{\text{CMS}} \)
approximately maximizing $\hat{M}_n$ for

$$m_n(z, \theta) = m^{\text{CMS}}_n(z, \theta) = y\mathbb{I}(x_1 + \theta'x_2 \geq 0)L_n(w),$$

where, for a kernel function $L$ and a bandwidth $b_n$, $L_n(w) = L(w/b_n)/b_n^d$.

Through its dependence on $b_n$, the function $m^{\text{CMS}}_n$ depends on $n$ in a non-negligible way. In particular, the effective sample size is $nb_n^d$ (rather than $n$) in the current setting, so to the extent that they exist one would expect primitive sufficient conditions for Condition CRA to include $q_n = b_n^d$ in this example. Apart from this predictable change, the properties of the conditional maximum score estimator $\hat{\theta}^{\text{CMS}}_n$ turn out to be qualitatively similar to those of $\hat{\theta}^{\text{MS}}_n$. To be specific, under regularity conditions the conditional maximum score estimator is covered by the results of Section 3 and an example-specific alternative (of smoothed maximum score type) to the generic numerical derivative estimator $\tilde{V}^\text{ND}_n$ is available; for details, see the supplemental appendix (Section A.5).

5 Discussion

The applicability of the procedure proposed in this note extends beyond the estimators covered by Theorem 1. For instance, it is not hard to show that our bootstrap-based distributional approximation is consistent also in the more standard case where $m_n(z, \theta)$ is sufficiently smooth in $\theta$ to ensure that an approximate maximizer of $\hat{M}_n$ is asymptotically normal and that the nonparametric bootstrap is consistent. In fact, $\tilde{\theta}^*_n$ is asymptotically equivalent to $\hat{\theta}^*_n$ in that standard case, so our procedure can be interpreted as a modification of the nonparametric bootstrap that is designed to be “robust” to the types of non-smoothness that give rise to cube root asymptotics.

Moreover, and perhaps more importantly, the idea of reshaping can be used to achieve consistency of bootstrap-based approximations to the distributions of certain estimators which are not of $M$-estimator type, yet exhibit cube root asymptotics and have the feature that the standard bootstrap-based approximations to their distribution are known to be inconsistent. In particular, the supplemental appendix (Section A.6) shows how the idea of reshaping a process can be used to achieve consistency on the part of a bootstrap-based approximation to the distribution of the
celebrated ($\sqrt{n}$-consistent) isotonic density estimator of Grenander (1956).\textsuperscript{1}

This note is not the first to propose a consistent resampling-based distributional approximation for cube root consistent estimators. For cube root asymptotic problems, the best known consistent alternative to the nonparametric bootstrap is probably subsampling (Politis and Romano, 1994), whose applicability was pointed out by Delgado, Rodriguez-Poo, and Wolf (2001). A related method is the rescaled bootstrap (Dümbgen, 1993), whose validity in cube root asymptotic $M$-estimation problems was established recently by Hong and Li (2017). In addition, case-specific smooth bootstrap methods have been proposed for leading examples such as maximum score estimation (Patra, Seijo, and Sen, 2015) and isotonic density estimation (Kosorok, 2008; Sen, Banerjee, and Woodroofe, 2010). Like ours, each of these methods can be viewed as offering a “robust” alternative to the nonparametric bootstrap but, unlike ours, they all achieve consistency by modifying the distribution used to generate the bootstrap counterpart of the estimator whose distribution is being approximated. In contrast, our method achieves consistency by means of an analytic modification to the objective function used to construct the bootstrap-based distributional approximation.

As pointed out by two referees and the coeditor, an alternative interpretation of our approach is available. Restating the result in (7) as

$$r_n(\hat{\theta}_n - \theta_0) \sim S_0(\mathcal{G}_0), \quad S_0(\mathcal{G}) = \arg\max_{s \in \mathbb{R}^d} \{Q_0(s) + \mathcal{G}(s)\},$$

our procedure approximates the distribution of $S_0(\mathcal{G}_0)$ by that of $\tilde{S}_n(\hat{G}^*_n)$, where

$$\hat{G}^*_n(s) = r_n^2[\hat{M}^*_n(\hat{\theta}_n + sr_n^{-1}) - \hat{M}^*_n(\hat{\theta}_n) - \hat{M}_n(\hat{\theta}_n + sr_n^{-1}) + \hat{M}_n(\hat{\theta}_n)]$$

is a bootstrap process whose distribution approximates that of $\mathcal{G}_0(s)$ and where

$$\tilde{S}_n(\mathcal{G}) = \arg\max_{s \in \mathbb{R}^d} \{Q_n(s) + \mathcal{G}(s)\}, \quad \hat{Q}_n(s) = -\frac{1}{2}s'\hat{V}_ns,$$

is an estimator of $S_0(\mathcal{G})$. In other words, our procedure replaces the functional $S_0$ with a consistent

\textsuperscript{1}The asymptotic properties of the Grenander estimator have been studied by Prakasa Rao (1969), Groeneboom (1985), and Kim and Pollard (1990), among others. Inconsistency of the standard bootstrap-based approximation to the distribution of the Grenander estimator has been pointed out by Kosorok (2008) and Sen, Banerjee, and Woodroofe (2010), among others.
estimator (namely, $\tilde{S}_n$) and its random argument $G_0$ with a bootstrap approximation (namely, $\hat{G}_n^*$). Similar constructions have been successfully applied in other settings, two relatively recent examples being Andrews and Soares (2010) and Fang and Santos (2016).

Finally, Seo and Otsu (2018) give conditions under which results of the form (7) can be obtained also when the data is weakly dependent. It seems plausible that a version of our procedure, implemented with resampling procedure suitable for dependent data, can be shown to be consistent under similar conditions, but it is beyond the scope of this note to substantiate that conjecture.

References


