Bootstrap-Based Inference for Cube Root Asymptotics*

Matias D. Cattaneo†  Michael Jansson‡  Kenichi Nagasawa§

June 26, 2019

Abstract

This paper proposes a consistent bootstrap-based distributional approximation for cube root consistent and related estimators exhibiting a Chernoff (1964)-type limiting distribution. 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. We illustrate the applicability of our core idea with six canonical examples in statistics, machine learning, econometrics, and biostatistics. Simulation evidence is also provided.

Keywords: cube root asymptotics, bootstrapping, maximum score, empirical risk minimization, monotone density, monotone regression, current status model, shape restrictions.

*A previous version of this paper circulated under the title “Bootstrap-Based Inference for Cube Root Consistent Estimators”. For comments and suggestions, we are grateful to Mehmet Caner, Andreas Hagemann, Kei Hirano, Bo Honoré, Guido Kuersteiner, Whitney Newey, and participants at various conferences, workshops, and seminars. 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 Operations Research and Financial Engineering, Princeton University.
‡Department of Economics, University of California at Berkeley and CREATE.
§Department of Economics, University of Warwick.
1 Introduction

In a seminal paper, Kim and Pollard (1990) studied 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; i.e., the limiting distribution is that of the maximizer of a Gaussian process. Estimators covered by Kim and Pollard’s results include not only celebrated estimators such as the isotonic density estimator of Grenander (1956) in statistics and the maximum score estimator of Manski (1975) in econometrics, but also more contemporary estimators arising in examples related to classification problems in machine learning (Mohammadi and van de Geer, 2005), nonparametric inference under shape restrictions in statistics and biostatistics (Groeneboom and Jongbloed, 2018), massive data $M$-estimation framework (Shi, Lu, and Song, 2018), and maximum score estimation in high-dimensional settings (Mukherjee, Banerjee, and Ritov, 2019). Moreover, Seo and Otsu (2018) recently generalized Kim and Pollard (1990) to allow for $n$-varying objective functions ($n$ denotes the sample size), further widening the applicability of cube-root-type asymptotics in statistics, machine learning, econometrics, biostatistics, and other disciplines. For example, their results covered the conditional maximum score estimator of Honoré and Kyriazidou (2000) in panel data econometrics.

An important feature of Chernoff-type asymptotic distributional approximations is that the covariance kernel of the Gaussian process characterizing the limiting distribution often 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). 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. Unfortunately, however, the standard nonparametric bootstrap is well known to be invalid in this setting (e.g., Abrevaya and Huang, 2005; Léger and MacGibbon, 2006; Kosorok, 2008; Sen, Banerjee, and Woodroofe, 2010). The purpose of this paper is to propose a generic and easy-to-implement bootstrap-based distributional approximation applicable in the context of cube root asymptotics.
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, our method offers a consistent distributional approximation for estimators exhibiting cube root asymptotics, and therefore can be used to construct asymptotically valid inference procedures in such settings. 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.

This paper is not the first to propose a consistent resampling-based distributional approximation for cube-root-type estimators. For canonical 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). Related methods are the rescaled bootstrap (Dümbgen, 1993) and the numerical bootstrap (Hong and Li, 2019), which can also be applied to this type of estimators. In addition, case-specific (smooth or non-standard) bootstrap methods have been proposed for leading examples such as monotone density estimation (Kosorok, 2008; Sen, Banerjee, and Woodroofe, 2010), maximum score estimation (Patra, Seijo, and Sen, 2018), and the current status model (Groeneboom and Hendrickx, 2018). For the more generic cube-root-type estimators analyzed in Seo and Otsu (2018), subsampling appears to be the only alternative available, and indeed the authors discuss in their concluding remarks the need for (and importance of) developing resampling methods based on the standard nonparametric bootstrap. Our paper appears to be the first to provide one such method.

Conceptually, each of the other resampling methods mentioned above (including ours) can be viewed as offering a “robust” alternative to the standard nonparametric bootstrap but, unlike ours, all other methods achieve consistency by modifying the distribution used to generate the bootstrap counterpart of the estimator whose distribution is being approximated. In contrast, our bootstrap-based method achieves consistency by means of an analytic modification to the objective
function used to construct the bootstrap-based distributional approximation. Furthermore, this modification only requires estimating a finite dimensional matrix, which can be interpreted as an analogue of a “standard error” estimator for the cube-root-type estimator. We discuss in detail how this matrix can be easily estimated in general using numerical derivatives, including consistency and an approximate mean square error expansion of our proposed estimator. In addition, we illustrate how exploiting example-specific features other estimators can be constructed.

The paper 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 objective function, as recently studied by Seo and Otsu (2018). We then illustrate our novel bootstrap-based inference methods with six distinct examples in statistics, machine learning, econometrics, and biostatistics settings. Specifically, Section 4 discusses three examples directly covered by our general results: the maximum score estimator of Manski (1975), the empirical risk minimization problem of Mohammadi and van de Geer (2005), and the conditional maximum score panel data estimator of Honoré and Kyriazidou (2000). Then, Section 5 presents three other examples that, while not in $M$-estimation form, can nonetheless be analyzed using our core ideas via the well-known “switching technique” (see Groeneboom and Jongbloed, 2014, 2018, for overviews): the monotone density estimator of Grenander (1956), the monotone regression estimator of Brunk (1958), and the distribution estimator proposed by Ayer, Brunk, Ewing, Reid, and Silverman (1955) for the current status model. Section 6 offers numerical evidence of our inference methods for two canonical examples: maximum score and monotone density estimation. To conclude, a methodological discussion of our results, including some potential avenues for future research, is provided in Section 7. Detailed derivations and proofs are collected in the online supplemental appendix.

2 Heuristics

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

$$\theta_0 = \underset{\theta \in \Theta}{\text{argmax}} \; M_0(\theta), \quad M_0(\theta) = \mathbb{E}[m_0(z, \theta)], \quad (1)$$
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:

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

where \( \sim \) denotes weak convergence, \( G_0 \) is a non-degenerate zero-mean Gaussian process with \( G_0(0) = 0 \), and \( Q_0(s) \) is a quadratic form given by

\[
Q_0(s) = -\frac{1}{2} s^T V_0 s, \quad V_0 = -\frac{\partial^2}{\partial \theta \partial \theta^T} M_0(\theta_0).
\]

Whereas the matrix \( V_0 \) governing the shape of \( Q_0 \) is finite-dimensional, the covariance kernel of \( G_0 \) 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 (\( \sqrt{n} \)-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 \( z_{1,n}^*, \ldots, z_{n,n}^* \) 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 \( \mathbb{P}_n^*[\hat{\theta}_n^* - \hat{\theta}_n \leq \cdot] \), where \( \mathbb{P}_n^* \) 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 (Abrevaya and Huang, 2005; Léger and MacGibbon, 2006; Kosorok, 2008; Sen, Banerjee, and Woodroofe, 2010).

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

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

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, $\hat{G}_n$ and $Q_n$ are usually asymptotically Gaussian and asymptotically quadratic, respectively, in the sense that

$$
\hat{G}_n(s) \rightsquigarrow G_0(s)
$$

and

$$
Q_n(s) \rightarrow Q_0(s).
$$

Under regularity conditions ensuring among other things that the convergence in (6) and (7) is suitably uniform in s, (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

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

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

$$
\hat{G}_n(s) = n^{2/3}[\hat{M}_n^*(\hat{\theta}_n + sn^{-1/3}) - \hat{M}_n^*(\hat{\theta}_n) - \hat{M}_n(\hat{\theta}_n + sn^{-1/3}) + \hat{M}_n(\hat{\theta}_n)]
$$
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 (6):

\[ \hat{G}_n^*(s) \sim_P G_0(s), \quad (8) \]

where \( \sim_P \) 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 \( \arg\max_{s \in \mathbb{R}^d} \hat{Q}_n(s) = 0 \), it turns out that \( \hat{Q}_n(s) \sim_P Q_0(s) \) in general. In other words, the natural bootstrap counterpart of (7) typically fails and, as a partial consequence, so does the natural bootstrap counterpart of (2); that is, \( \sqrt{n}(\hat{\theta}_n - \hat{\theta}_n) \not\sim_P \arg\max_{s \in \mathbb{R}^d}\{Q_0(s) + G_0(s)\} \)

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

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

where \( \hat{m}_n \) is a suitably “reshaped” version of \( m_0 \) satisfying two properties. First, \( \hat{G}_n^* \) should be asymptotically equivalent to \( \hat{G}_n^* \), where \( \hat{G}_n^* \) is the counterpart of \( \hat{G}_n^* \) associated with \( \hat{m}_n \):

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

Second, and most importantly, \( \tilde{Q}_n \) should be asymptotically quadratic, where \( \tilde{Q}_n \) is the counterpart of \( \hat{Q}_n \) associated with \( \tilde{m}_n :\)

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

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

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

where, by construction, $\hat{G}_n^*(s) = \hat{G}_n(s)$ and $\hat{Q}_n(s) = -s^T \hat{V}_n s / 2$. Because $\hat{G}_n^* = \hat{G}_n$, $\hat{G}_n^*(s) \to_P G_0(s)$ whenever (8) holds. In addition, $\hat{Q}_n(s) \to_P Q_0(s)$ provided $\tilde{V}_n \to_P V_0$. As a consequence, it seems plausible that if $\tilde{V}_n \to_P V_0$, then

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

implying in particular that $P_n[\hat{\theta}_n^* - \hat{\theta}_n \leq \cdot]$ is a consistent estimator of $P[\hat{\theta}_n - \theta_0 \leq \cdot]$.

### 3 Main Result

When making the heuristics of Section 2 precise, we consider the more general situation where the estimator $\hat{\theta}_n$ is an approximate maximizer (with respect to $\theta \in \Theta \subseteq \mathbb{R}^d$) 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 $\mathcal{M}_n = \{m_n(\cdot, \theta) : \theta \in \Theta\}$, $\bar{m}_n(z) = \sup_{m \in \mathcal{M}_n} |m(z)|$, $\Theta^\delta_0 = \{\theta \in \Theta : ||\theta - \theta_0|| \leq \delta\}$, $\mathcal{D}^\delta_0 = \{m_n(\cdot, \theta) - m_n(\cdot, \theta_0) : \theta \in \Theta^\delta_0\}$, and $d^\delta_n(z) = \sup_{d \in \mathcal{D}^\delta_0} |d(z)|$.

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

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

Also, $\sup_{\theta \in \Theta} |M_n(\theta) - M_0(\theta)| = o(1)$ and, for every $\delta > 0$, $\sup_{\theta \in \Theta^\delta_0} M_0(\theta) \leq M_0(\theta_0)$.

(ii) $\theta_0$ is an interior point of $\Theta$ and, for some $\delta > 0$, $M_0$ and $M_n$ are twice continuously differentiable on $\Theta^\delta_0$ and $\sup_{\theta \in \Theta^\delta_0} ||\partial^2 [M_n(\theta) - M_0(\theta)]/\partial \theta \partial \theta'|| = o(1)$.

Also, $r_n ||\partial M_n(\theta_0)/\partial \theta|| = o(1)$ and $V_0 = -\partial^2 M_0(\theta_0)/\partial \theta \partial \theta'$ is positive definite.

(iii) For some $\delta > 0$, $\{\mathcal{D}^\delta_0 : n \geq 1, 0 < \delta' \leq \delta\}$ is uniformly manageable for the envelopes $d^\delta_n$ and $q_n \sup_{0 < \delta' \leq \delta} \mathbb{E}[d^\delta_n(z)^2/\delta'] = O(1)$.

(iv) For every positive $\delta_n$ with $\delta_n = O(r_n^{-1})$, $q_n^2 \mathbb{E}[d^\delta_n(z)^3] + q_n^3 r_n^{-1} \mathbb{E}[d^\delta_n(z)^4] = o(1)$, 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 $s \neq t$,

$$\sup_{\theta \in \Theta^\delta_0} |q_n \mathbb{E}[\{m_n(z, \theta + \delta_n s) - m_n(z, \theta)\} \{m_n(z, \theta + \delta_n t) - m_n(z, \theta)\}] / \delta_n] - C_0(s, t)| = o(1).$$

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

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

and $\sup_{\theta, \theta' \in \Theta^\delta_0} \mathbb{E}[||m_n(z, \theta) - m_n(z, \theta')||/||\theta - \theta'||] = O(1)$.

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 $\hat{\theta}_n \to_P \theta_0$ 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 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 $nq_n = r_n^3$. In perfect agreement with Seo and Otsu (2018), if $\hat{\theta}_n$ is an approximate maximizer of $\tilde{M}_n$, then

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

where $G_0$ is a zero-mean Gaussian process with $G_0(\theta) = 0$ and covariance kernel $C_0$ and where $Q_0(s) = -s'V_0s/2$. 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

\[
\sup_{t \in \mathbb{R}^d} \left| \mathbb{P}_n[r_n(\check{\theta}_n - \hat{\theta}_n) \leq t] - \mathbb{P}[r_n(\check{\theta}_n - \theta_0) \leq t] \right| \to_{\mathbb{P}} 0. \tag{10}
\]

Because the limiting distribution in (9) 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 (10) holds provided the estimator \( \check{\theta}_n \) satisfies the following bootstrap counterpart of (9):

\[
r_n(\check{\theta}_n - \hat{\theta}_n) \sim_{\mathbb{P}} \arg\max_{s \in \mathbb{R}^d} \mathcal{G}_0(s) + \mathcal{Q}_0(s).
\]

Theorem 1, our main result, gives sufficient conditions for this to occur.

**Theorem 1** Suppose Condition CRA holds. If \( \tilde{V}_n \to_{\mathbb{P}} V_0 \) and if

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

then (10) holds.

Using the notation and conditions in Theorem 1, the algorithm for our proposed bootstrap-based distributional approximation is as follows:

**Step 1.** Using the sample \( z_1, \ldots, z_n \), compute \( \hat{M}_n(\theta) \) and let \( \hat{\theta}_n \) be an approximate maximizer thereof.

**Step 2.** Using \( \hat{\theta}_n \) and \( z_1, \ldots, z_n \), compute \( \tilde{V}_n \). (A generic estimator \( \tilde{V}_n \) is given in the next subsection, and example-specific estimators are discussed in Section 4.)

**Step 3.** Using \( \hat{\theta}_n, \tilde{V}_n, \) and the bootstrap sample \( z_{1,n}^*, \ldots, z_{n,n}^* \), compute \( \hat{M}_n^*(\theta) \) and let \( \hat{\theta}_n^* \) be an approximate maximizer thereof. (\( \hat{\theta}_n \) and \( \tilde{V}_n \) are not recomputed at this step.)

**Step 4.** Repeat Step 3 \( B \) times and compute the empirical distribution of \( r_n(\hat{\theta}_n^* - \hat{\theta}_n) \) using the \( B \) bootstrap replications.
3.1 Generic Estimation of $V_0$

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 the matrix $\tilde{V}_n^{\text{ND}}$ with element $(k,l)$ given by

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

where $e_k$ is the $k$th unit vector in $\mathbb{R}^d$ and where $\epsilon_n$ is a positive (vanishing) 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 $\tilde{V}_n^{\text{ND}} \to \mathbb{P} V_0$.

Plausibility of the high-level condition $r_n(\hat{\theta}_n - \theta_0) = O_P(1)$ follows from (9). More generally, if only consistency is assumed on the part of $\hat{\theta}_n$, then $\tilde{V}_n^{\text{ND}} \to \mathbb{P} V_0$ holds provided $\epsilon_n \to 0$ and $||\hat{\theta}_n - \theta_0||/\epsilon_n \to 0$.

For practical implementation, we also develop a Nagar-type mean squared error (MSE) expansion for $\tilde{V}_n^{\text{ND}}$ under additional, mild regularity conditions.

**Lemma 2** Suppose the conditions of Lemma 1 hold and that, for some $\delta > 0$, $M_0$ and $M_n$ are four times continuously differentiable on $[0,\delta]$ with $\sup \theta \in \Theta_0^\delta |\partial^4[M_n(\theta) - M_0(\theta)]/\partial \theta_{j1}\partial \theta_{j2}\partial \theta_{j3}\partial \theta_{j4}| = o(1)$ for all $j_1,j_2,j_3,j_4 \in \{1,2,\ldots,d\}$. If $C_0(s,-s) = 0$ and if $C_0(s,t) = C_0(-s,-t)$ for all $s, t \in \mathbb{R}^d$, then $\tilde{V}_n^{\text{ND}}$ admits an approximation $\hat{V}_n^{\text{ND}}$ satisfying

$$\hat{V}_n^{\text{ND}} - \tilde{V}_n^{\text{ND}} = O_P(r_n^{-3/2}\epsilon_n^{-3/2}) + O(r_n^{-1})$$

and

$$\mathbb{E}[||\hat{V}_n^{\text{ND}} - V_n||^2] = \epsilon_n^2 \left( \sum_{k=1}^d \sum_{l=1}^d B_{kl}^2 \right) + r_n^{-3}\epsilon_n^{-3} \left( \sum_{k=1}^d \sum_{l=1}^d V_{kl} \right) + o(\epsilon_n^4 + r_n^{-3}\epsilon_n^{-3}).$$
where \( \mathbf{V}_n = -\partial^2 M_n(\theta_0)/\partial \theta \partial \theta' \) and where

\[
\mathbf{B}_{kl} = -\frac{1}{6} \left( \frac{\partial^4}{\partial \theta_k \partial \theta_l} M_0(\theta_0) + \frac{\partial^4}{\partial \theta_k^2 \partial \theta_l^2} M_0(\theta_0) \right),
\]

\[
\mathbf{V}_{kl} = \frac{1}{8} \left[ \mathcal{C}_0(\mathbf{e}_k + \mathbf{e}_l, \mathbf{e}_k + \mathbf{e}_l) + \mathcal{C}_0(\mathbf{e}_k - \mathbf{e}_l, \mathbf{e}_k - \mathbf{e}_l) - 2\mathcal{C}_0(\mathbf{e}_k + \mathbf{e}_l, \mathbf{e}_k - \mathbf{e}_l) - 2\mathcal{C}_0(\mathbf{e}_k + \mathbf{e}_l, -\mathbf{e}_k + \mathbf{e}_l) \right].
\]

Lemma 2 goes beyond consistency (Lemma 1) and develops an approximate MSE expansion for \( \tilde{\mathbf{V}}_{n,\text{ND}} \), which can be used to construct optimal choices for the numerical derivative step-size \( \epsilon_n \). Specifically, 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 \( \mathcal{C}_0 \) and the fourth order derivatives of \( M_0 \) evaluated at \( \theta_0 \):

\[
\epsilon_n^{\text{AMSE}} = \left( \frac{3}{4} \sum_{k=1}^d \sum_{l=1}^d \mathbf{V}_{kl} \right)^{1/7} r_n^{-3/7}.
\]

As the notation suggests, the constants \( \mathbf{B}_{kl} \) and \( \mathbf{V}_{kl} \) correspond to the bias and variance of the element \((k, l)\) of \( \tilde{\mathbf{V}}_{n,\text{ND}} \), so the element-wise asymptotic MSE-optimal tuning parameter choice for \( \tilde{\mathbf{V}}_{n,\text{ND}} \) is

\[
\epsilon_{n,kl}^{\text{AMSE}} = \left( \frac{3\mathbf{V}_{kl}}{4\mathbf{B}_{kl}^2} \right)^{1/7} r_n^{-3/7}.
\]

In the supplemental appendix, Section A.7 discusses plug-in implementations of \( \epsilon_n^{\text{AMSE}} \) and \( \epsilon_{n,kl}^{\text{AMSE}} \), while Section A.3 discusses alternative generic estimators of \( \mathbf{V}_0 \) based on kernel smoothing. Finally, Section 6 below explores the numerical performance of the feasible implementation of the MSE-optimal choice of \( \epsilon_n \) in the context of two distinct examples.

### 4 Examples

This section illustrates our proposed bootstrap-based distributional approximation with three examples. The first two are M-estimation problems where \( m_n \) does not change with \( n \), while the last example corresponds to a situation where \( m_n \) depends on \( n \).
4.1 Maximum Score

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

\[
y = \mathbb{I}(\beta'_0 x + u \geq 0), \quad \text{Median}(u|x) = 0,
\]

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 \( u \) is an unobserved error term. 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 \( \beta_0 \) as \( x = (x_1, x_2')' \), a maximum score estimator of \( \theta_0 \in \Theta \subseteq \mathbb{R}^d \) is any \( \hat{\theta}_{n}^{MS} \) satisfying

\[
\hat{M}_{n}^{MS}(\hat{\theta}_{n}^{MS}) \geq \sup_{\theta \in \Theta} \hat{M}_{n}^{MS}(\theta) - o_p(n^{-2/3}), \quad \hat{M}_{n}^{MS}(\theta) = \frac{1}{n} \sum_{i=1}^{n} m_{i}^{MS}(z_i, \theta),
\]

where \( m_{i}^{MS}(z, \theta) = (2y - 1) \mathbb{I}(x_1 + \theta'x_2 \geq 0) \).

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. Second, in addition to the generic estimator \( \hat{V}_{n}^{MD} \) discussed above, the maximum score estimator admits example-specific consistent estimators of the \( \mathbb{V}_0 \) associated with it. In what follows, we briefly illustrate both points; for omitted details, see Section A.4.1 of the supplemental appendix.

To state primitive conditions for Condition CRA, let \( F_{u|x_1, x_2} \) and \( F_{x_1|x_2} \) denote the conditional distribution function of \( u \) given \( (x_1, x_2) \) and \( x_1 \) given \( x_2 \), respectively.

**Condition MS** The following are satisfied:

(i) \( 0 < P(y = 1|x) < 1 \) almost surely and \( F_{u|x_1, x_2}(u|x_1, x_2) \) is differentiable in \( u \) and \( x_1 \) with bounded and continuous derivatives.

(ii) The support of \( x \) is not contained in any proper linear subspace of \( \mathbb{R}^{d+1} \), \( E[\|x_2\|^2] < \infty \), and conditional on \( x_2 \), \( x_1 \) has everywhere positive Lebesgue density. Also, \( F_{x_1|x_2}(x_1|x_2) \) is twice differentiable in \( x_1 \) with bounded and continuous derivatives.
(iii) The set $\Theta$ is compact and $\theta_0$ is an interior point of $\Theta$.

(iv) $M^{\text{MS}}$ is twice continuously differentiable near $\theta_0$ and $V^{\text{MS}} = -\partial^2 M^{\text{MS}}(\theta_0)/\partial \theta \partial \theta'$ is positive definite, where $M^{\text{MS}}(\theta) = \mathbb{E}[m^{\text{MS}}(z, \theta)]$.

Under Condition MS, $\hat{\theta}_n^{\text{MS}}$ satisfies (2) with $V_0 = V^{\text{MS}}$ and

$$C_0(s, t) = C^{\text{MS}}(s, t) = \mathbb{E}[f_{x_1|x_2}(-x_2'\theta_0|x_2) \min\{|x_2's|, |x_2't|\} \mathbb{I}(\text{sgn}(x_2's) = \text{sgn}(x_2't))],$$

where $f_{x_1|x_2}$ denotes the conditional density of $x_1$ given $x_2$. Indeed, Condition CRA is satisfied (with $q_n = 1$), so Theorem 1 is applicable. To state a maximum score version of that result, let $z_1^*, \ldots, z_n^*$, be a random sample from the empirical distribution of $z_1, \ldots, z_n$ and suppose $\tilde{\theta}_n^{\text{MS,*}}$ satisfies

$$\tilde{M}_n^{\text{MS,*}}(\tilde{\theta}_n^{\text{MS,*}}) \geq \sup_{\theta \in \Theta} \hat{M}_n^{\text{MS,*}}(\theta) - o_P(n^{-2/3}), \quad \hat{M}_n^{\text{MS,*}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \tilde{m}_n^{\text{MS}}(z_i^*, \theta),$$

where, for some estimator $\tilde{V}_n$ of $V^{\text{MS}}$,

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

**Corollary MS** If Condition MS holds and if $\tilde{V}_n \to_P V^{\text{MS}}$, then

$$\sup_{t \in \mathbb{R}^d} \left| P_n(\sqrt{n}(\tilde{\theta}_n^{\text{MS,*}} - \hat{\theta}_n^{\text{MS}}) \leq t) - P(\sqrt{n}(\tilde{\theta}_n^{\text{MS}} - \theta_0) \leq t) \right| \to_P 0.$$

Because Condition CRA is satisfied, it follows from Lemma 1 that the consistency requirement of the corollary is satisfied by the numerical derivative estimator $\tilde{V}_n^{\text{ND}}$ discussed in Section 3.1 provided the parameter $\epsilon_n$ satisfies $\epsilon_n \to 0$ and $n\epsilon_n^3 \to \infty$. Moreover, the other assumptions of Lemma 2 are mild additional regularity conditions in this example, so an MSE-optimal tuning parameter choice is feasible. In addition to the numerical derivative estimator, alternative consistent estimators of $V^{\text{MS}}$ can be constructed exploiting the specific structure of this example. One option is to employ a “plug-in” estimator of

$$V^{\text{MS}} = 2\mathbb{E} \left[ f_{u|x_1,x_2}(0) - x_2'\theta_0, x_2) f_{x_1|x_2}(-x_2'\theta_0, x_2) x_2 x_2' \right],$$

14
where \( f_{u|x_1,x_2} \) denotes the conditional density of \( u \) given \((x_1,x_2)\). To implement this estimator, nonparametric estimators of the conditional densities \( f_{u|x_1,x_2} \) and \( f_{x_1|x_2} \) are required. A simpler example-specific estimator is

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

where, for a kernel function \( K \) and a bandwidth \( h_n \), \( \hat{K}_n(u) = d \hat{K}_n(u)/du \) and \( K_n(u) = K(u/h_n)/h_n \). In words, \( \tilde{V}_n^{\text{MS}} \) is constructed by “smoothing out” the indicator function entering \( m^{\text{MS}}(z,\theta) \) and then differentiating the corresponding (regularized) objective function twice. Like the numerical derivative estimator, the estimator \( \tilde{V}_n^{\text{MS}} \) is consistent under mild conditions on its tuning parameters \((h_n \text{ and } K)\); for details, see the supplemental appendix (Section A.4.1, Lemma MS), which also reports a valid approximate MSE expansion useful to select tuning the bandwidth in an optimal way.

Section 6 below reports results of a Monte Carlo experiment evaluating the performance of our proposed bootstrap-based inference procedure for this example.

### 4.2 Empirical Risk Minimization

Mohammadi and van de Geer (2005) consider two-category classification problems in machine learning. Specifically, given a binary outcome \( y \in \{-1,1\} \) and a vector of features \( \mathbf{x} \in \mathcal{X} \), the goal is to estimate the \( \theta_0 \) that minimizes the misclassification error (or risk) \( \mathbb{P}[h_\theta(\mathbf{x}) \neq y] \) with respect to \( \theta \in \Theta \subseteq \mathbb{R}^d \), where \( \{h_\theta : \theta \in \Theta\} \) is a collection of classifiers. For simplicity, here we follow Mohammadi and van de Geer (2005, Section 2.1) and consider the case where the feature is univariate with support \( \mathcal{X} = [0,1] \) and the classifiers are of the form

\[
h_\theta(x) = \sum_{\ell=1}^{d} (-1)^\ell \mathbb{I}(\theta_{\ell-1} \leq x < \theta_\ell), \quad \theta = (\theta_1, \theta_2, \cdots, \theta_d)',
\]

where

\[
\Theta = \{(\theta_1, \theta_2, \cdots, \theta_d)' \in [0,1]^d : 0 = \theta_0 \leq \theta_1 \leq \cdots \leq \theta_d \leq \theta_{d+1} = 1\}.
\]

In Section A.4.2 of the supplemental appendix we also analyze a case where \( \mathbf{x} \) is vector-valued and
and \( \mathcal{H} \) includes non-linear classifiers, and present additional results for this example.

As an estimator of \( \theta_0 \), we consider the empirical risk minimizer \( \hat{\theta}_n^{\text{ERM}} \) satisfying

\[
\hat{M}_n^{\text{ERM}}(\hat{\theta}_n^{\text{ERM}}) \geq \sup_{\theta \in \Theta} M_n^{\text{ERM}}(\theta) - \alpha_2(n^{-2/3}), \quad M_n^{\text{ERM}}(\theta) = \frac{1}{n} \sum_{i=1}^n m_\theta(z_i, \theta),
\]

where \( m_\theta(z_i, \theta) = -1(h_\theta(x_i) \neq y_i) \) and \( z_1, \ldots, z_n \) is a random sample of \( z \). When analyzing \( \hat{\theta}_n^{\text{ERM}} \), we follow Mohammadi and van de Geer (2005, Theorem 1) and impose the following primitive conditions.

**Condition ERM** The following are satisfied:

(i) \( P(0) < 1/2 \) and \( P \) admits a continuous derivative \( p \) in a neighborhood of each element of \( \theta_0 \), where \( P(x) = \mathbb{P}[y = 1|x] \). Also, \( \mathbb{P}[h_\theta(x) \neq y] \) is continuous with respect to \( \theta \).

(ii) The distribution function of \( x \), denoted by \( F \), is absolutely continuous and its Lebesgue density \( f \) is continuously differentiable in a neighborhood of each element of \( \theta_0 \).

(iii) \( \theta_0 \) is an interior point of \( \Theta \).

(iv) \( \theta_0 = (\theta_{0,1}, \theta_{0,2}, \ldots, \theta_{0,d})' \) is the unique minimizer of \( \mathbb{P}[h_\theta(x) \neq y] \) and \( p(\theta_{0,\ell})f(\theta_{0,\ell}) \neq 0 \) for \( \ell \in \{1, \ldots, d\} \).

Under Condition ERM, \( \hat{\theta}_n^{\text{ERM}} \) satisfies (2) with

\[
V_0 = V^{\text{ERM}} = 2 \begin{pmatrix}
  p(\theta_{0,1})f(\theta_{0,1}) & 0 & \cdots & 0 \\
  0 & -p(\theta_{0,2})f(\theta_{0,2}) & \cdots & 0 \\
  \cdots & \cdots & \cdots & \cdots \\
  0 & 0 & \cdots & (-1)^{d+1}p(\theta_{0,d})f(\theta_{0,d})
\end{pmatrix}
\]

and

\[
C_0(s, t) = C^{\text{ERM}}(s, t) = \sum_{\ell=1}^d f(\theta_{0,\ell}) \min\{|s_{\ell}|, |t_{\ell}|\} \mathbb{I}\{\text{sgn}(s_{\ell}) = \text{sgn}(t_{\ell})\}
\]

for \( s = (s_1, \ldots, s_d)' \) and \( t = (t_1, \ldots, t_d)' \). Indeed, Condition CRA is satisfied (with \( q_n = 1 \)), so Theorem 1 is applicable. To state an empirical risk minimization version of that result, let \( z_{1,n}^*, \ldots, z_{n,n}^* \) be a random sample from the empirical distribution of \( z_1, \ldots, z_n \) and suppose \( \hat{\theta}_n^{\text{ERM}} \).
satisfies

\[ \tilde{M}_n^{\text{ERM}}(\hat{\theta}^{\text{ERM}}_n) \geq \sup_{\theta \in \Theta} M_n^{\text{ERM}}(\theta) - o_P(n^{-2/3}), \quad M_n^{\text{ERM}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \tilde{m}_n^{\text{ERM}}(z_{i,n}^*, \theta), \]

where, for some estimator \( \tilde{V}_n \) of \( V^{\text{ERM}} \),

\[ \tilde{m}_n^{\text{ERM}}(z, \theta) = m^{\text{ERM}}(z, \theta) - M_n^{\text{ERM}}(\theta) - \frac{1}{2} (\theta - \hat{\theta}_n^{\text{ERM}})' \tilde{V}_n (\theta - \hat{\theta}_n^{\text{ERM}}). \]

**Corollary ERM**  If Condition ERM holds and if \( \tilde{V}_n \to \tilde{V}^{\text{ERM}} \), then

\[ \sup_{t \in \mathbb{R}^d} \left| \mathbb{P}_n \left[ \tilde{V}_n^{\text{ERM},*} \left( \hat{\theta}_n^{\text{ERM}} \right) - \tilde{V}_n^{\text{ERM}} \right] \leq t \right| - \mathbb{P} \left[ \tilde{V}_n^{\text{ERM},*} \left( \hat{\theta}_n^{\text{ERM}} \right) - \theta_0 \right] \to_P 0. \]

As in the maximum score example, the consistency requirement on \( \tilde{V}_n \) can be met in various ways. First, the generic numerical derivative estimator of Section 3.1 can be used directly. Second, a “plug-in” estimator of \( V^{\text{ERM}} \) can be constructed with the help of nonparametric estimators of \( p \) and \( f \). Finally, as explained in Section A.4.2 of the supplemental appendix, kernel smoothing ideas similar to those used to construct \( \tilde{V}_n^{\text{NS}} \) can be used to obtain an estimator of \( V^{\text{ERM}} \).

### 4.3 Conditional Maximum Score

Consider the dynamic binary response model

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

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 a common scale factor. 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

\[
\hat{M}_n^{\text{CMS}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} m_n^{\text{CMS}}(z_i, \theta), \quad m_n^{\text{CMS}}(z, \theta) = y \mathbb{I}(x_1 + \theta' x_2 \geq 0) \kappa_n(w),
\]

where, for a kernel function \( \kappa \) and a bandwidth \( b_n, \kappa_n(w) = \kappa(w/b_n)/b_n^d. \)

Through its dependence on \( b_n, \) the function \( m_n^{\text{CMS}} \) 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 \( \hat{\theta}_n^{\text{CMS}} \) turn out to be qualitatively similar to those of \( \hat{\theta}_n^{\text{MS}}. \) To be specific, under regularity conditions, the conditional maximum score estimator is covered by the results of Section 3 and an example-specific alternative to the generic numerical derivative estimator is available. We outline all details in Section A.4.3 of the supplemental appendix to avoid a tedious long list of regularity conditions and other untidy notation.

5 Nonparametric Estimation under Shape Constraints

The monotone density estimator of Grenander (1956) is another well known estimator that exhibits cube root asymptotics and has the feature that the standard bootstrap-based approximation to its distribution is known to be inconsistent (e.g., Kosorok, 2008; Sen, Banerjee, and Woodroofe, 2010). In turn, the Grenander estimator can be viewed as a prototypical example of a nonparametric estimator respecting a shape constraint (e.g., Groeneboom and Jongbloed, 2014, 2018), two other well known examples being the isotonic regression estimator of Brunk (1958) and the Ayer, Brunk, Ewing, Reid, and Silverman (1955) estimator of the distribution function in the current status model. These estimators are not \( M \)-estimators and are therefore not covered by the results of the earlier sections. Nevertheless, we show in this section that in all three examples the idea of reshaping can be used to achieve consistency of bootstrap-based distributional approximations.

Before turning to the specific examples, we outline the general approach that we take. Suppose
the (scalar) estimator $\hat{\theta}_n$ of the estimand $\theta_0$ satisfies a switch relation of the form

$$\sqrt{n} (\hat{\theta}_n - \theta_0) \leq t \quad \iff \quad \text{argmax}_{s \in \mathcal{S}_n} \{ \hat{G}_n(s) + t\hat{L}_n(s) + Q_n(s) \} \leq 0,$$  

(11)

where the set $\mathcal{S}_n$, the processes $\hat{G}_n, \hat{L}_n$, and the function $Q_n$ satisfy

$$1(s \in \mathcal{S}_n) \to \mathbb{P} 1(s \in \mathbb{R}) = 1, \quad \hat{G}_n(s) \to \gamma_0 \mathcal{W}(s), \quad \hat{L}_n(s) \to \mathbb{P} -\lambda_0 s,$$  

(12)

and

$$Q_n(s) \to -\frac{1}{2} \chi_0 s^2,$$  

(13)

respectively, with $\mathcal{W}$ denoting a two-sided Wiener process with $\mathcal{W}(0) = 0$, while $\gamma_0, \lambda_0$, and $\chi_0$ being positive constants. It then stands to reason that

$$\mathbb{P} [ \sqrt{n} (\hat{\theta}_n - \theta_0) \leq t ] \to \mathbb{P} \left[ \text{argmax}_{s \in \mathbb{R}} \{ \gamma_0 \mathcal{W}(s) - t\lambda_0 s - \frac{1}{2} \chi_0 s^2 \} \leq 0 \right],$$

a result that can be stated more compactly as

$$\sqrt{n} (\hat{\theta}_n - \theta_0) \to \omega_0 \text{argmax}_{s \in \mathbb{R}} \{ \mathcal{W}(s) - s^2 \}, \quad \omega_0 = \sqrt{\frac{4\chi_0 \gamma_0^2}{\lambda_0^3}},$$  

(14)

where the equivalence of the two formulations follows from the fact that

$$\mathbb{P} \left[ \text{argmax}_{s \in \mathbb{R}} \{ \gamma_0 \mathcal{W}(s) - t\lambda_0 s - \frac{1}{2} \chi_0 s^2 \} \leq 0 \right] = \mathbb{P} \left[ \sqrt{\frac{4\chi_0 \gamma_0^2}{\lambda_0^3}} \text{argmax}_{s \in \mathbb{R}} \{ \mathcal{W}(s) - s^2 \} \leq t \right].$$

See, for instance, van der Vaart and Wellner (1996, Problem 3.2.5).

Similarly to (11), suppose the bootstrap analogue of $\hat{\theta}_n$ admits a set $\hat{\mathcal{S}}_n^*$ and processes $\hat{G}_n^*, \hat{L}_n^*$, and $\hat{Q}_n$ such that

$$\sqrt{n} (\hat{\theta}_n^* - \hat{\theta}_n) \leq t \quad \iff \quad \text{argmax}_{s \in \hat{\mathcal{S}}_n^*} \{ \hat{G}_n^*(s) + t\hat{L}_n^*(s) + \hat{Q}_n(s) \} \leq 0,$$  

(15)

where

$$1(s \in \hat{\mathcal{S}}_n^*) \to \mathbb{P} 1(s \in \mathbb{R}) = 1, \quad \hat{G}_n^*(s) \to \mathbb{P} \gamma_0 \mathcal{W}(s), \quad \hat{L}_n^*(s) \to \mathbb{P} -\lambda_0 s,$$  

(16)
but where \( \hat{Q}_n(s) \to_{\mathbb{P}} -\frac{1}{2}\chi_0 s^2 \). In that case, 
\[
\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n) \not\to_{\mathbb{P}} \omega_0 \arg\max_{s \in \mathbb{R}} \{W(s) - s^2\},
\]
the source of the bootstrap inconsistency being once again the inability of the process \( \hat{Q}_n \) to reproduce the shape of \( Q_n \).

Adapting the ideas of Section 2 to the present setting, it seems natural to attempt to construct a \( \tilde{\theta}_n^* \) satisfying
\[
\sqrt{n}(\tilde{\theta}_n^* - \hat{\theta}_n) \leq t \Leftrightarrow \arg\max_{s \in \tilde{S}_n^*} \{\tilde{G}_n^*(s) + t\tilde{L}_n^*(s) + \hat{Q}_n(s)\} \leq 0,
\]
where \((\tilde{S}_n^*, \tilde{G}_n^*, \tilde{L}_n^*) \approx (\hat{S}_n^*, \hat{G}_n^*, \hat{L}_n^*) \to_{\mathbb{P}} -\frac{1}{2}\chi_0 s^2 \), in an appropriate sense (see below), as it seems plausible that such a construction would restore bootstrap consistency by virtue of \( \tilde{\theta}_n^* \) satisfying
\[
\sqrt{n}(\tilde{\theta}_n^* - \hat{\theta}_n) \to_{\mathbb{P}} \omega_0 \arg\max_{s \in \mathbb{R}} \{W(s) - s^2\}.
\] 
As we shall see, such a construction is feasible and achieves bootstrap consistency in all three shape-contained estimation examples mentioned above. Indeed, in all three examples, the only additional ingredient needed in order to construct \( \tilde{\theta}_n^* \) is a consistent estimator \( \tilde{e}_n \) of the scalar \( \chi_0 \).

5.1 Monotone Density

Our first example is the celebrated \( \sqrt{n} \)-consistent monotone density estimator of Grenander (1956). The asymptotic properties of this 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.

Suppose the object of interest is \( \theta_0 = f_0(x) \), where \( f_0 \) is the Lebesgue density of a continuously distributed nonnegative random variable \( x \) and where \( x \in (0, \infty) \) is a given evaluation point. Assuming \( x_1, \ldots, x_n \) is a random sample from the distribution of \( x \) and assuming \( f_0 \) is non-increasing on \( [0, \infty) \), a natural estimator is \( \hat{\theta}_n = \hat{f}_{n}^\text{RD} (x) \), where \( \hat{f}_{n}^\text{RD} \) is the Grenander estimator of \( f_0 \); that is, \( \hat{f}_{n}^\text{RD} \) maximizes the log likelihood \( \sum_{i=1}^{n} \log f(x_i) \) over all non-increasing densities \( f \) on \( [0, \infty) \).

As is well known, \( \hat{f}_{n}^\text{RD} \) is the left derivative of the least concave majorant (LCM) of the empirical distribution function \( \hat{F}_n(\cdot) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(x_i \leq \cdot) \). Using this fact, it can be shown that \( \hat{\theta}_n \) satisfies
(11) with
\[ \hat{S}_n = [-\sqrt[3]{n}x, \infty), \]
\[ \hat{G}_n(s) = n^{2/3}\{\hat{F}_n(x + sn^{-1/3}) - \hat{F}_n(x) - F_0(x + sn^{-1/3}) + F_0(x)\}, \]
\[ \hat{L}_n(s) = -s, \]
and
\[ Q_n(s) = n^{2/3}\{F_0(x + sn^{-1/3}) - F_0(x) - f_0(x)sn^{-1/3}\}, \]
where \( F_0 \) is the distribution function of \( x \). As a consequence, (12) holds with \( \gamma_0 = \sqrt{f_0(x)} \) and \( \lambda_0 = 1 \). If also \( f_0 \) admits a negative derivative \( f_0'(x) \) at \( x \), then (13) holds with \( \chi_0 = -f_0'(x) \) and (14) holds with \( \omega_0 = \sqrt{-4f_0'^2(x)f_0(x)} \).

Letting \( x^*_{1,n}, \ldots, x^*_{n,n} \) denote a random sample from the empirical distribution of \( x_1, \ldots, x_n \), a natural bootstrap analogue of \( \hat{\theta}_n \) is given by \( \hat{\theta}^*_n = \hat{f}^{\text{MD}*}_n(x) \), where \( \hat{f}^{\text{MD}*}_n \) is the left derivative of the LCM of \( \hat{F}^*_n(\cdot) = n^{-1}\sum_{i=1}^n \mathbb{I}(x^*_i \leq \cdot) \). This estimator satisfies (15) with
\[ \hat{S}^*_n = [-\sqrt[3]{n}x, \infty), \]
\[ \hat{G}^*_n(s) = n^{2/3}\{\hat{F}^*_n(x + sn^{-1/3}) - \hat{F}^*_n(x) - \hat{F}_n(x + sn^{-1/3}) + \hat{F}_n(x)\} \]
\[ \hat{L}^*_n(s) = -s, \]
and
\[ \hat{Q}_n(s) = n^{2/3}\{\hat{F}_n(x + sn^{-1/3}) - \hat{F}_n(x) - \hat{f}^{\text{MD}*}_n(x)sn^{-1/3}\}. \]

Moreover, it can be shown that (16) holds with \( \gamma_0 = \sqrt{f_0(x)} \) and \( \lambda_0 = 1 \). On the other hand, being non-smooth, the empirical distribution function \( \hat{F}_n \) does not admit a quadratic approximation around \( x \), so the natural bootstrap analog of (13) fails; that is, \( \hat{Q}_n(s) \not\rightarrow \frac{1}{2}f_0'(x)s^2 \). Sen, Banerjee, and Woodroofe (2010, p. 1968) attribute the inconsistency of the bootstrap to this fact and show that bootstrap consistency can be restored by generating the bootstrap sample from a smoothed/regularized approximation to \( \hat{F}_n \). A similar result was obtained by Kosorok (2008).

To define our alternative bootstrap procedures, let \( \hat{f}^l_n(x) \) be an estimator of \( f_0'(x) \) and let \( \hat{f}^{\text{MD}*}_n \)
be the left derivative of the LCM of $\hat{F}^*_n$, where $\hat{F}^*_n$ is the following reshaped version of $\hat{F}^*_n$:

$$\hat{F}^*_n(\cdot) = \hat{F}^*_n(\cdot) - \hat{F}_n(\cdot) + \hat{F}_n(x) + \hat{f}_{\text{MD}}^n(\cdot - x) + \frac{1}{2}\tilde{f}^n(x)(\cdot - x)^2.$$ 

Then $\tilde{\theta}^*_n = \hat{f}_{\text{MD}}^n(x)$ satisfies (17) with $(\hat{S}^*_n, \hat{\hat{G}}_n, \hat{L}^*_n) = (\hat{S}^*_n, \hat{\hat{G}}_n, \hat{L}^*_n)$ and $\hat{Q}_n(s) = \frac{1}{2}\tilde{f}^n(x)s^2$. As one would hope, (18) is satisfied provided $\tilde{\theta}^*_n(x)$ is consistent.

**Theorem MD** If $f_0$ is differentiable at $x$ with derivative $f_0'(x) < 0$ and if $\hat{f}_n^*(x) \rightarrow_P f_0'(x)$, then

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}^*\left[ \sqrt{n}(\hat{f}_{\text{MD}}^n(x) - f_{\text{MD}}^n(x)) \leq t \right] - \mathbb{P}\left[ \sqrt{n}(f_{\text{MD}}^n(x) - f_0(x)) \leq t \right] \right| \rightarrow_P 0.$$ 

The (pointwise) consistency requirement $\tilde{\theta}^*_n(x) \rightarrow_P f_0'(x)$ can be met in various ways. One possibility is to apply a version of our generic numerical derivative estimator, but an arguably more natural option is to employ a standard kernel (derivative) density estimator with its corresponding plug-in MSE-optimal bandwidth selector; see Section A.5.1 of the supplemental appendix for details.

Section 6 below evaluates the numerical performance of our new bootstrap-based inference procedure for the monotone density estimator in a simulation study.

### 5.2 Monotone Regression

Our second example is the monotone regression estimator of Brunk (1958). In this example, the object of interest is $\theta_0 = \mu_0(x)$, where $\mu_0$ is the regression function $\mu_0(x) = \mathbb{E}(y|x)$ for a dependent variable $y$ and (scalar) continuously distributed regressor $x$, and $x \in \mathbb{R}$ is a given evaluation point. Assuming $z_1, \ldots, z_n$ is a random sample from the distribution of $z = (y, x)'$ and assuming $\mu_0$ is non-decreasing on $\mathcal{X}$, a natural estimator is $\hat{\theta}_n = \hat{\mu}_n^{\text{MR}}(x)$, where $\hat{\mu}_n^{\text{MR}}$ minimizes $\sum_{i=1}^n (y_i - \mu(x_i))^2$ over all non-decreasing functions $\mu$.

We do not incorporate interpolation in the construction of the estimator $\hat{\theta}_n$ only for simplicity, but all our results continue to hold if interpolation is used. Therefore, we can take $\hat{\theta}_n$ to be the left derivative of the greatest convex minorant (GCM) of $\hat{U}_n \circ \hat{F}_n^{-1}$ at $\hat{F}_n(x)$, where

$$\hat{U}_n(\cdot) = \frac{1}{n} \sum_{i=1}^n y_i \mathbb{I}(x_i \leq \cdot), \quad \hat{F}_n(\cdot) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(x_i \leq \cdot).$$
Using this fact and defining

$$\hat{M}_n(\cdot; \theta) = -[\hat{U}_n(\cdot) - \theta \hat{F}_n(\cdot)], \quad M_0(\cdot; \theta) = \mathbb{E}[\hat{M}_n(\cdot; \theta)],$$

it can be shown that $\hat{\theta}_n$ satisfies (11) with

$$\hat{S}_n = [\sqrt[n]{\min_{1 \leq i \leq n} x_i - x}, \sqrt[n]{\max_{1 \leq i \leq n} x_i - x}],$$

$$\hat{G}_n(s) = n^{2/3} \{ \hat{M}_n(x + sn^{-1/3}; \theta_0) - M_0(x) + M_0(x + sn^{-1/3}; \theta_0) \},$$

$$\hat{L}_n(s) = -n^{1/3} \{ \hat{F}_n(x + sn^{-1/3}) - \hat{F}_n(x) \},$$

and

$$Q_n(s) = n^{2/3} \{ M_0(x + sn^{-1/3}; \theta_0) - M_0(x; \theta_0) \}. $$

Under the following condition, (12) holds with $\gamma_0 = \sqrt{\sigma_0^2(x)f_0(x)}$ and $\lambda_0 = f_0(x)$, (13) holds with $\chi_0 = \chi_{\text{MR}} = \mu_0'(x)f_0(x)$, and (14) holds with $\omega_0 = \sqrt{4\sigma_0^2(x)\mu_0'(x)/f_0(x)}$.

**Condition MR** For some neighborhood $\mathcal{X}$ of $x$, the following are satisfied:

(i) $\mathbb{E}[\{y - \mu_0(x)\}^4] < \infty$, $\mu_0$ admits a continuous derivative $\mu_0'$ on $\mathcal{X}$, and $\sigma_0^2$ is continuous on $\mathcal{X}$, where $\sigma_0^2(x) = \mathbb{V}[y|x]$.

(ii) The density $f_0$ of $x$ has $f_0(x) > 0$ and admits a bounded derivative $f_0'$ on $\mathcal{X}$.

Letting $\mathbf{z}_{1,n}^*, \ldots, \mathbf{z}_{n,n}^*$ denote a random sample from the empirical distribution of $\mathbf{z}_1, \ldots, \mathbf{z}_n$, a natural bootstrap analogue of $\hat{\mu}_{n_{\text{MR}}}^*(x)$ is given by $\hat{\mu}_{n_{\text{MR}},*}^*(x)$, the left derivative of the GCM of $\hat{U}_n^* \circ \hat{F}_n^{*,-1}$ at $\hat{F}_n^*(x)$, where

$$\hat{U}_n^*(\cdot) = \frac{1}{n} \sum_{i=1}^n y_{i,n}^* \mathbb{I}(x_{i,n}^* \leq \cdot), \quad \hat{F}_n^*(\cdot) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(x_{i,n}^* \leq \cdot).$$

Defining

$$\hat{M}_n^*(\cdot; \theta) = -[\hat{U}_n^*(\cdot) - \theta \hat{F}_n^*(\cdot)].$$
it can be shown that \( \hat{\theta}_n^* = \hat{\mu}_n^{MR,*}(x) \) satisfies (15) with

\[
\hat{S}_n^* = \left[ \frac{1}{\sqrt{n}} ( \min_{1 \leq i \leq n} x_{i,n}^* - x), \frac{1}{\sqrt{n}} ( \max_{1 \leq i \leq n} x_{i,n}^* - x) \right],
\]

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

\[
\hat{L}_n^*(s) = -n^{1/3} \{ \hat{F}_n^*(x + sn^{-1/3}) - \hat{F}_n^*(x) \},
\]

and

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

Moreover, it can be shown that (16) holds with \( \gamma_0 = \sqrt{\sigma_0^2(x)f_0(x)} \) and \( \lambda_0 = f_0(x) \). On the other hand, the process \( \hat{M}_n(\cdot; \hat{\theta}_n) \) does not admit a quadratic approximation around \( x \), so the natural bootstrap analog of (13) fails; that is, \( \hat{Q}_n(s) \toP -\frac{1}{2} \mu'_0(x)f_0(x)s^2 \).

Bootstrap consistency can be restored by reshaping \( \hat{U}_n^* \). Let \( \bar{\chi}_n \) is an estimator of \( \chi^{MR} \) and let \( \hat{\mu}_n^{MR,*}(x) \), the left derivative of the GCM of \( \hat{U}_n^* \circ \hat{F}_n^{-1} \) at \( \hat{F}_n(x) \), where

\[
\bar{U}_n^*(\cdot) = \bar{U}_n^*(\cdot) - \bar{U}_n(\cdot) + \bar{U}_n(x) + \hat{\mu}_n^{MR}(x) \{ \hat{F}_n(\cdot) - \hat{F}_n(x) \} + \frac{1}{2} \bar{\chi}_n(\cdot - x)^2.
\]

By construction, \( \hat{\theta}_n^* = \hat{\mu}_n^{MR,*}(x) \) satisfies (17) with \( (\hat{S}_n^*, \hat{G}_n^*, \hat{L}_n^*) = (\hat{S}_n^*, \hat{G}_n^*, \hat{L}_n^*) \) and \( \hat{Q}_n(s) = -\frac{1}{2} \hat{\chi}_n s^2 \).

As one would hope, (18) is satisfied under mild conditions.

**Theorem MR**  If Condition MR holds and if \( \bar{\chi}_n \toP \chi^{MR} \), then

\[
\sup_{t \in \mathbb{R}} | \mathbb{P}^* \left[ \frac{1}{\sqrt{n}} (\hat{\mu}_n^{MR}(x) - \mu_0(x)) \leq t \right] - \mathbb{P} \left[ \frac{1}{\sqrt{n}} (\hat{\mu}_n^{MR}(x) - \mu_0(x)) \leq t \right] | \toP 0.
\]

The consistency requirement \( \bar{\chi}_n \toP \chi^{MR} = \mu'_0(x)f_0(x) \) can be met in various ways. One possibility is to apply a version of our generic numerical derivative estimator and another option is to use a plug-in estimator employing separate nonparametric estimators of \( \mu'_0(x) \) and \( f_0(x) \); see Section A.5.2 of the supplemental appendix for details.
5.3 Current Status

Our final example corresponds to the basic current status model considered by Ayer, Brunk, Ewing, Reid, and Silverman (1955). In this example, the object of interest is $F_0(x)$, where $x \in \mathbb{R}$ is a given evaluation point and where $F_0$ is the distribution function of an unobserved (scalar) continuously distributed $x$. If $y$ is independent of $x$ and continuously distributed, then $F_0$ is identified (on the support of $y$) from the distribution of $z = (y, d)'$, where $d = \mathbb{I}(x \leq y)$. Assuming $z_1, \ldots, z_n$ is a random sample from the distribution of $z$, a natural estimator of $\theta_0 = F_0(x)$ is $\hat{\theta}_n = \hat{F}_n^{\text{CS}}(x)$, where $\hat{F}_n^{\text{CS}}$ maximizes the (partial) log likelihood $\sum_{i=1}^n \{d_i \log F(y_i) + (1 - d_i) \log [1 - F(y_i)]\}$ over all distribution functions $F$.

It can be shown that $\hat{F}_n^{\text{CS}}$ minimizes $\sum_{i=1}^n (d_i - F(y_i))^2$ over all distribution functions $F$. In other words, $\hat{F}_n^{\text{CS}}(x)$ is of the same form as the monotone regression estimator $\hat{\mu}_n^{\text{MR}}(x)$ analyzed in the previous subsection. In particular, $\hat{\theta}_n$ can be taken to be the left derivative of the GCM of $\hat{V}_n \hat{H}_n^{-1}$ at $\hat{H}_n(x)$, where

$$\hat{V}_n(\cdot) = \frac{1}{n} \sum_{i=1}^n d_i \mathbb{I}(y_i \leq \cdot), \quad \hat{H}_n(\cdot) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(y_i \leq \cdot),$$

and it can be shown that (12) holds with $\gamma_0 = \sqrt{F_0(x) [1 - F_0(x)] g_0(x)}$ and $\lambda_0 = g_0(x)$, (13) holds with $\chi_0 = \chi^{\text{CS}} = f_0(x) g_0(x)$, and (14) holds with $\omega_0 = \sqrt{4 F_0(x) [1 - F_0(x)] f_0(x) / g_0(x)}$ under the following condition.

**Condition CS** For some neighborhood $\mathcal{X}$ of $x$, the densities $f_0$ and $g_0$ of $x$ and $y$ are continuous on $\mathcal{X}$ with $f_0(x) > 0$ and $g_0(x) > 0$.

Letting $z_{1,n}', \ldots, z_{n,n}'$ denote a random sample from the empirical distribution of $z_1, \ldots, z_n$, a natural bootstrap analogue of $\hat{F}_n^{\text{CS}}(x)$ is given by $\hat{F}_n^{\text{CS}*}(x)$, the left derivative of the GCM of $\hat{V}_n^* \hat{H}_n^{-1}$ at $\hat{H}_n^*(x)$, where

$$\hat{V}_n^*(\cdot) = \frac{1}{n} \sum_{i=1}^n d_{i,n}' \mathbb{I}(y_{i,n}' \leq \cdot), \quad \hat{H}_n^*(\cdot) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(y_{i,n}' \leq \cdot).$$

Once again, the distributional approximation based on $\hat{\theta}_n^*$ is inconsistent, but bootstrap consistency can be restored by reshaping $\hat{V}_n^*$. Let $\hat{\chi}_n$ be an estimator of $\chi^{\text{CS}}$ and let $\hat{F}_n^{\text{CS}*}(x)$ be the left derivative.
of the GCM of $\tilde{V}_n^* \circ \tilde{H}_n^{*-1}$ at $\hat{H}_n^*(x)$, where

$$\tilde{V}_n^*(\cdot) = \hat{V}_n^*(\cdot) - \hat{V}_n(\cdot) + \hat{F}_{CS}(x)(\hat{H}_n(\cdot) - \hat{H}_n(x)) + \frac{1}{2} \hat{\chi}_n(\cdot - x)^2.$$ 

**Theorem CS** If Condition CS holds and if $\hat{\chi}_n \to_{P} \chi^{CS}$, then

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}[\sqrt{n}(\hat{F}_{CS}^*(x) - \hat{F}_{CS}(x)) \leq t] - \mathbb{P}[\sqrt{n}(\hat{F}_{CS}(x) - F_0(x)) \leq t] \right| \to_{P} 0.$$ 

The consistency requirement $\hat{\chi}_n \to_{P} \chi^{CS} = f_0(x)g_0(x)$ can be met by using either our generic numerical derivative approach or an example-specific plug-in estimator; see Section A.5.3 of the supplemental appendix for details.

### 6 Numerical Results

We illustrate the numerical performance of our proposed bootstrap-based inference methods for the maximum score estimator and the monotone density estimator. In both cases we find that the resulting confidence intervals exhibit very good coverage and length properties in the simulation designs considered, outperforming other available resampling methods for cube root consistent estimators.

#### 6.1 The Maximum Score Estimator

We consider the setup of Section 4.1, and generate data from a model with $d = 1$, $\theta_0 = 1$,

$$x = (x_1, x_2)' \sim N\left(\begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right),$$

and $u$ generated by three distinct distributions. Specifically, DGP 1 sets $u \sim \text{Logistic}(0, 1)/\sqrt{2\pi^2/3}$, DGP 2 sets $u \sim T_3/\sqrt{3}$, where $T_k$ denotes a Student’s $t$ distribution with $k$ degrees of freedom, and DGP 3 sets $u \sim (1 + 2(x_1 + x_2)^2 + (x_1 + x_2)^4)\text{Logistic}(0, 1)/\sqrt{\pi^2/48}$.

The Monte Carlo experiment employs a sample size $n = 1,000$ with $B = 2,000$ bootstrap replications and $S = 2,000$ simulations. For each of the three DGPs, we implement the stan-
standard non-parametric bootstrap, \( m \)-out-of-\( n \) bootstrap, and our proposed method using the two estimators \( \nabla_{\text{MS}}^n \) and \( \nabla_{\text{ND}}^n \) of \( V_0 \). We report empirical coverage for nominal 95\% confidence intervals and their average interval length. For the case of our proposed procedures, we investigate their performance using both (i) a grid of fixed tuning parameter values (bandwidth/derivative step) around the MSE-optimal choice and (ii) infeasible and feasible AMSE-optimal choices of the tuning parameter.

Table 1 presents the main results, which are consistent across all three simulation designs. First, as expected, the standard nonparametric bootstrap (labeled “Standard”) does not perform well, leading to confidence intervals with an average 64\% empirical coverage rate. Second, the \( m \)-out-of-\( n \) bootstrap (labeled “\( m \)-out-of-\( n \)”) performs somewhat better for small subsamples, but leads to very large average interval length of the resulting confidence intervals. Our proposed methods, on the other hand, exhibit excellent finite sample performance in this Monte Carlo experiment. Results employing the example-specific plug-in estimator \( \nabla_{\text{MS}}^n \) are presented under the label “Plug-in” while results employing the generic numerical derivative estimator \( \nabla_{\text{ND}}^n \) are reported under the label “Num Deriv”. Empirical coverage appears stable across different values of the tuning parameters for each method, with better performance in the case of \( \nabla_{\text{MS}}^n \). We conjecture that \( n = 1,000 \) is too small for the numerical derivative estimator \( \nabla_{\text{ND}}^n \) to lead to as good inference performance as \( \nabla_{\text{MS}}^n \) (e.g., note that the MSE-optimal choice \( \epsilon_{\text{MSE}} \) is greater than 1). Nevertheless, empirical coverage of confidence intervals constructed using our proposed bootstrap-based method is close to 95\% in all cases except when \( \nabla_{\text{ND}}^n \) is used with either the infeasible asymptotic choice \( \epsilon_{\text{AMSE}} \) or its estimated counterpart \( \hat{\epsilon}_{\text{AMSE}} \), and with an average interval length of at most half that of any of the \( m \)-out-of-\( n \) competing confidence intervals. In particular, confidence intervals based on \( \nabla_{\text{MS}}^n \) implemented with the feasible bandwidth \( \hat{h}_{\text{AMSE}} \) perform quite well across the three DGPs considered.

6.2 The Monotone Density Estimator

We consider the setup of Section 5.1 and employ the DGPs and simulation setting previously considered in Sen, Banerjee, and Woodroofe (2010). We estimate \( f_0(x) \) at the evaluation point \( x = 1 \) using a random sample of observations, where three distinct distributions are considered: DGP 1 sets \( x \sim \text{Exponential}(1) \), DGP 2 sets \( x \sim |\text{Normal}(0,1)| \), and DGP 3 sets \( x \sim |T_3| \). As in the case of the maximum score example, the Monte Carlo experiment employs a sample size
$n = 1,000$ with $B = 2,000$ bootstrap replications and $S = 2,000$ simulations, and compares three types of bootstrap-based inference procedures: the standard non-parametric bootstrap, $m$-out-of-$n$ bootstrap, and our proposed method using two distinct estimators of $f_0'(x)$, an “off-the-shelf” plug-in estimator and our generic numerical derivative estimator $\tilde{V}_n^{\text{ND}}$.

Table 2 presents the numerical results for this example. We continue to report empirical coverage for nominal 95% confidence intervals and their average interval length. For the case of our proposed procedures, we again investigate their performance using both (i) a grid of fixed tuning parameter value (derivative step/bandwidth) and (ii) infeasible and feasible AMSE-optimal choice of tuning parameter. Also in this case, the numerical evidence is very encouraging. Our proposed bootstrap-based inference method leads to confidence intervals with excellent empirical coverage and average interval length, outperforming both the standard nonparametric bootstrap (which is inconsistent) and the $m$-out-of-$n$ bootstrap (which is consistent). In particular, in this example, the plug-in method employs an off-the-shelf kernel derivative estimator, which in this case leads to confidence intervals that are very robust (i.e., insensitive) to the choice of bandwidth. Furthermore, when the corresponding feasible off-the-shelf MSE-optimal bandwidth is used, the resulting confidence intervals continue to perform excellently. Finally, the generic numerical derivative estimator also leads to very good performance of bootstrap-based infeasible and feasible confidence intervals.

7 Conclusion

We developed a valid resampling procedure for cube root asymptotics based on the nonparametric bootstrap. While it is well known that for cube root consistent (and related) estimators exhibiting a Chernoff (1964) type distribution the standard nonparametric bootstrap is invalid, our approach restores bootstrap validity by applying a carefully tailored reshapement of the objective function defining the estimator. Such reshapement is easy to implement both in general and in specific cases, as illustrated by the six distinct examples we considered.

From a more general perspective, our approach can be heuristically explained as follows. Restating the result in (9) as

$$r_n(\hat{\theta}_n - \theta_0) \sim S_0(G_0), \quad S_0(G) = \arg\max_{s \in \mathbb{R}^d} \{G(s) + Q_0(s)\},$$
our procedure approximates the distribution of $S_0(G_0)$ by that of $\tilde{S}_n(\widehat{G}_n^*)$, where

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

is a bootstrap process whose distribution approximates that of $G_0(s)$ and where

$$\tilde{S}_n(G) = \operatorname{argmax}_{s \in \mathbb{R}^d} \{G(s) + \bar{Q}_n(s)\}, \quad \bar{Q}_n(s) = -\frac{1}{2} s' \tilde{V}_n s,$$

is an estimator of $S_0(G)$. In other words, our procedure replaces the functional $S_0$ with a consistent estimator (namely, $\tilde{S}_n$) and its random argument $G_0$ with a bootstrap approximation (namely, $\widehat{G}_n^*$).

Furthermore, 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 $\bar{M}_n$ is asymptotically normal and that the nonparametric bootstrap is consistent. In fact, $\bar{\theta}_n^*$ is asymptotically equivalent to $\widehat{\theta}_n^*$ 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.

Finally, Seo and Otsu (2018) give conditions under which results of the form (9) can be obtained also when the data is weakly dependent. See also Bagchi, Banerjee, and Stoev (2016), and references therein. 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 paper to substantiate that conjecture.

References


Table 1: Simulations, Maximum Score Estimator, 95% Confidence Intervals.

(a) \( n = 1,000, S = 2,000, B = 2,000 \)

<table>
<thead>
<tr>
<th></th>
<th>DGP 1</th>
<th>DGP 2</th>
<th>DGP 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Standard</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( h, \epsilon )</td>
<td>0.639</td>
<td>0.645</td>
<td>0.640</td>
</tr>
<tr>
<td>Coverage</td>
<td>0.475</td>
<td>0.480</td>
<td>0.247</td>
</tr>
<tr>
<td>Length</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| **m-out-of-n**       |       |       |       |
| \( m = [n^{1/2}] \)  | 0.998 | 0.997 | 0.999 |
| \( m = [n^{2/3}] \)  | 0.979 | 0.979 | 0.985 |
| \( m = [n^{4/5}] \)  | 0.902 | 0.894 | 0.904 |

| **Plug-in: \( V_n^{\text{MS}} \)** |       |       |       |
| 0.7 \( h_{\text{MSE}} \)   | 0.434 | 0.406 | 0.105 |
| 0.8 \( h_{\text{MSE}} \)   | 0.496 | 0.464 | 0.120 |
| 0.9 \( h_{\text{MSE}} \)   | 0.558 | 0.522 | 0.135 |
| 1.0 \( h_{\text{MSE}} \)   | 0.620 | 0.580 | 0.150 |
| 1.1 \( h_{\text{MSE}} \)   | 0.682 | 0.638 | 0.165 |
| 1.2 \( h_{\text{MSE}} \)   | 0.744 | 0.696 | 0.180 |
| 1.3 \( h_{\text{MSE}} \)   | 0.806 | 0.754 | 0.195 |

| \( h_{\text{AMSE}} \)  | 0.385 | 0.367 | 0.119 |
| \( \hat{h}_{\text{AMSE}} \) | 0.446 | 0.415 | 0.155 |

| **Num Deriv: \( V_n^{\text{ND}} \)** |       |       |       |
| 0.7 \( \epsilon_{\text{MSE}} \) | 0.980 | 0.904 | 0.203 |
| 0.8 \( \epsilon_{\text{MSE}} \) | 1.120 | 1.033 | 0.232 |
| 0.9 \( \epsilon_{\text{MSE}} \) | 1.260 | 1.163 | 0.261 |
| 1.0 \( \epsilon_{\text{MSE}} \) | 1.400 | 1.292 | 0.290 |
| 1.1 \( \epsilon_{\text{MSE}} \) | 1.540 | 1.421 | 0.319 |
| 1.2 \( \epsilon_{\text{MSE}} \) | 1.680 | 1.550 | 0.348 |
| 1.3 \( \epsilon_{\text{MSE}} \) | 1.820 | 1.679 | 0.377 |

| \( \epsilon_{\text{AMSE}} \)  | 0.483 | 0.476 | 0.216 |
| \( \hat{\epsilon}_{\text{AMSE}} \)  | 0.518 | 0.513 | 0.368 |

Notes:
(i) Panel **Standard** refers to standard nonparametric bootstrap, Panel **m-out-of-n** refers to \( m \)-out-of-\( n \) nonparametric bootstrap with subsample \( m \), Panel **Plug-in: \( V_n^{\text{MS}} \)** refers to our proposed bootstrap-based implemented using the example-specific plug-in drift estimator, and Panel **Num Deriv: \( V_n^{\text{ND}} \)** refers to our proposed bootstrap-based implemented using the generic numerical derivative drift estimator.

(ii) Column “\( h, \epsilon \)” reports tuning parameter value used or average across simulations when estimated, and Columns “Coverage” and “Length” report empirical coverage and average length of bootstrap-based 95% percentile confidence intervals, respectively.

(iii) \( h_{\text{MSE}} \) and \( \epsilon_{\text{MSE}} \) correspond to the simulation MSE-optimal choices, \( h_{\text{AMSE}} \) and \( \epsilon_{\text{AMSE}} \) correspond to the AMSE-optimal choices, and \( \hat{h}_{\text{MSE}} \) and \( \hat{\epsilon}_{\text{MSE}} \) correspond to the ROT feasible implementation of \( h_{\text{MSE}} \) and \( \epsilon_{\text{MSE}} \) described in the supplemental appendix.
Table 2: Simulations, Isotonic Density Estimator, 95% Confidence Intervals.

(a) \( n = 1,000, S = 2,000, B = 2,000 \)

<table>
<thead>
<tr>
<th></th>
<th>DGP 1</th>
<th></th>
<th>DGP 2</th>
<th></th>
<th>DGP 3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( h, \epsilon )</td>
<td>Coverage</td>
<td>Length</td>
<td>( h, \epsilon )</td>
<td>Coverage</td>
<td>Length</td>
</tr>
<tr>
<td><strong>Standard</strong></td>
<td></td>
<td>0.828</td>
<td>0.146</td>
<td></td>
<td>0.808</td>
<td>0.172</td>
</tr>
<tr>
<td><strong>m-out-of-n</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( m = [n^{1/2}] )</td>
<td>1.000</td>
<td>0.438</td>
<td></td>
<td></td>
<td>0.995</td>
<td>0.495</td>
</tr>
<tr>
<td>( m = [n^{2/3}] )</td>
<td>0.989</td>
<td>0.314</td>
<td></td>
<td></td>
<td>0.979</td>
<td>0.360</td>
</tr>
<tr>
<td>( m = [n^{1/5}] )</td>
<td>0.953</td>
<td>0.235</td>
<td></td>
<td></td>
<td>0.937</td>
<td>0.274</td>
</tr>
<tr>
<td><strong>Plug-in:</strong> ( V_{n}^{\text{ID}} )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( 0.7 \cdot h_{\text{MSE}} )</td>
<td>0.264</td>
<td>0.955</td>
<td>0.157</td>
<td></td>
<td>0.202</td>
<td>0.947</td>
</tr>
<tr>
<td>( 0.8 \cdot h_{\text{MSE}} )</td>
<td>0.302</td>
<td>0.954</td>
<td>0.157</td>
<td></td>
<td>0.231</td>
<td>0.946</td>
</tr>
<tr>
<td>( 0.9 \cdot h_{\text{MSE}} )</td>
<td>0.339</td>
<td>0.951</td>
<td>0.156</td>
<td></td>
<td>0.260</td>
<td>0.944</td>
</tr>
<tr>
<td>( 1.0 \cdot h_{\text{MSE}} )</td>
<td>0.377</td>
<td>0.949</td>
<td>0.154</td>
<td></td>
<td>0.289</td>
<td>0.941</td>
</tr>
<tr>
<td>( 1.1 \cdot h_{\text{MSE}} )</td>
<td>0.415</td>
<td>0.940</td>
<td>0.151</td>
<td></td>
<td>0.318</td>
<td>0.938</td>
</tr>
<tr>
<td>( 1.2 \cdot h_{\text{MSE}} )</td>
<td>0.452</td>
<td>0.934</td>
<td>0.147</td>
<td></td>
<td>0.347</td>
<td>0.934</td>
</tr>
<tr>
<td>( 1.3 \cdot h_{\text{MSE}} )</td>
<td>0.490</td>
<td>0.922</td>
<td>0.142</td>
<td></td>
<td>0.376</td>
<td>0.928</td>
</tr>
<tr>
<td>( h_{\text{AMSE}} )</td>
<td>0.380</td>
<td>0.949</td>
<td>0.154</td>
<td></td>
<td>0.300</td>
<td>0.940</td>
</tr>
<tr>
<td>( \hat{h}_{\text{AMSE}} )</td>
<td>0.364</td>
<td>0.950</td>
<td>0.155</td>
<td></td>
<td>0.290</td>
<td>0.941</td>
</tr>
<tr>
<td><strong>Num Deriv:</strong> ( V_{n}^{\text{ND}} )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( 0.7 \cdot \epsilon_{\text{MSE}} )</td>
<td>0.726</td>
<td>0.954</td>
<td>0.158</td>
<td></td>
<td>0.527</td>
<td>0.947</td>
</tr>
<tr>
<td>( 0.8 \cdot \epsilon_{\text{MSE}} )</td>
<td>0.830</td>
<td>0.956</td>
<td>0.159</td>
<td></td>
<td>0.602</td>
<td>0.947</td>
</tr>
<tr>
<td>( 0.9 \cdot \epsilon_{\text{MSE}} )</td>
<td>0.933</td>
<td>0.956</td>
<td>0.160</td>
<td></td>
<td>0.678</td>
<td>0.944</td>
</tr>
<tr>
<td>( 1.0 \cdot \epsilon_{\text{MSE}} )</td>
<td>1.037</td>
<td>0.956</td>
<td>0.159</td>
<td></td>
<td>0.753</td>
<td>0.942</td>
</tr>
<tr>
<td>( 1.1 \cdot \epsilon_{\text{MSE}} )</td>
<td>1.141</td>
<td>0.955</td>
<td>0.159</td>
<td></td>
<td>0.828</td>
<td>0.940</td>
</tr>
<tr>
<td>( 1.2 \cdot \epsilon_{\text{MSE}} )</td>
<td>1.244</td>
<td>0.956</td>
<td>0.160</td>
<td></td>
<td>0.904</td>
<td>0.936</td>
</tr>
<tr>
<td>( 1.3 \cdot \epsilon_{\text{MSE}} )</td>
<td>1.348</td>
<td>0.960</td>
<td>0.163</td>
<td></td>
<td>0.979</td>
<td>0.935</td>
</tr>
<tr>
<td>( \epsilon_{\text{AMSE}} )</td>
<td>0.927</td>
<td>0.956</td>
<td>0.160</td>
<td></td>
<td>0.731</td>
<td>0.943</td>
</tr>
<tr>
<td>( \hat{\epsilon}_{\text{AMSE}} )</td>
<td>0.888</td>
<td>0.956</td>
<td>0.159</td>
<td></td>
<td>0.708</td>
<td>0.943</td>
</tr>
</tbody>
</table>

Notes:
(i) Panel **Standard** refers to standard nonparametric bootstrap, Panel **m-out-of-n** refers to \( m \)-out-of-\( n \) nonparametric bootstrap with subsample \( m \), Panel **Plug-in:** \( V_{n}^{\text{ID}} \) refers to our proposed bootstrap-based implemented using the example-specific plug-in drift estimator, and Panel **Num Deriv:** \( V_{n}^{\text{ND}} \) refers to our proposed bootstrap-based implemented using the generic numerical derivative drift estimator.
(ii) Column “\( h, \epsilon \)” reports tuning parameter value used or average across simulations when estimated, and Columns “Coverage” and “Length” report empirical coverage and average length of bootstrap-based 95% percentile confidence intervals, respectively.
(iii) \( h_{\text{MSE}} \) and \( \epsilon_{\text{MSE}} \) correspond to the simulation MSE-optimal choices, \( h_{\text{AMSE}} \) and \( \epsilon_{\text{AMSE}} \) correspond to the AMSE-optimal choices, and \( \hat{h}_{\text{AMSE}} \) and \( \hat{\epsilon}_{\text{AMSE}} \) correspond to the ROT feasible implementation of \( h_{\text{AMSE}} \) and \( \epsilon_{\text{AMSE}} \) described in the supplemental appendix.