Local M-estimation with discontinuous criterion for dependent and limited observations

LOCAL M-ESTIMATION WITH DISCONTINUOUS CRITERION FOR DEPENDENT AND LIMITED OBSERVATIONS

MYUNG HWAN SEO AND TAISUKE OTSU

Abstract. This paper examines asymptotic properties of local M-estimators under three sets of high-level conditions. These conditions are sufficiently general to cover the minimum volume predictive region, conditional maximum score estimator for a panel data discrete choice model, and many other widely used estimators in statistics and econometrics. Specifically, they allow for discontinuous criterion functions of weakly dependent observations, which may be localized by kernel smoothing and contain nuisance parameters whose dimension may grow to infinity. Furthermore, the localization can occur around parameter values rather than around a fixed point and the observation may take limited values, which leads to set estimators. Our theory produces three different nonparametric cube root rates and enables valid inference for the local M-estimators, building on novel maximal inequalities for weakly dependent data. Our results include the standard cube root asymptotics as a special case. To illustrate the usefulness of our results, we verify our conditions for various examples such as the Hough transform estimator with diminishing bandwidth, maximum score-type set estimator, and many others.

1. Introduction

There is a class of estimation problems in statistics where a point (or set-valued) estimator is obtained by maximizing a discontinuous and possibly localized criterion function. As a prototype, consider the estimation of a simplified version of the minimum volume predictive region for y given $x = c$ (Polonik and Yao, 2000). Let $\mathbb{I}\{\cdot\}$ be the indicator function, $K(\cdot)$ be a kernel function, and h_n be a bandwidth. For a significance level α, the estimator $[\hat{\theta} \pm \hat{\nu}]$ is obtained by the M-estimation:

$$\max_{\theta \in \Theta} \sum_{t=1}^{n} \mathbb{I}\{|y_t - \theta| \leq \hat{\nu}\} K\left(\frac{x_t - c}{h_n}\right), \tag{1}$$

where Θ is some parameter space and

$$\hat{\nu} = \inf\left\{ \nu \in \mathbb{R} : \max_{\theta \in \Theta} \frac{\sum_{t=1}^{n} \mathbb{I}\{|y_t - \theta| \leq \nu\} K\left(\frac{x_t - c}{h_n}\right)}{\sum_{t=1}^{n} K\left(\frac{x_t - c}{h_n}\right)} \geq \alpha \right\}.$$

Key words and phrases. Cube root asymptotics, Maximal inequality, Mixing process, Partial identification, Parameter-dependent localization.

The authors would like to thank Aureo de Paula, Marine Carrasco, Javier Hidalgo, Dennis Kristensen, Benedikt Pötscher, Peter Robinson, Kyungchul Song, Yoon-Jae Whang, and seminar and conference participants at Cambridge, CIREQ in Montreal, CORE in Louvain, CREATES in Aarhus, IHS in Vienna, ISNPS in Cádiz, LSE, Surrey, UCL, Vienna, and York for helpful comments. The authors also acknowledge helpful comments from an associate editor and anonymous referees. This research was partly supported by Promising-Pioneering Researcher Program through Seoul National University (Seo) and the ERC Consolidator Grant (SNP 615882) (Otsu).

This estimation exhibits several distinguishing features such as discontinuity of the criterion function, localization by kernel smoothing, and serial dependence in time series data, which have prevented a full-blown asymptotic analysis of the M-estimator $\hat{\theta}$. Only the consistency is reported in the literature.

This type of M-estimation has numerous applications. Since Chernoff's (1964) study on estimation of the mode, many papers raised such estimation problems such as the shorth (Andrews *et al.,* 1972), least median of squares (Rousseeuw, 1984), nonparametric monotone density estimation (Prakasa Rao, 1969), and maximum score estimation (Manski, 1975). These classical examples are studied in a seminal work by Kim and Pollard (1990), which explained elegantly how this type of estimation problems induces the so-called cube root asymptotics in a unified framework by means of empirical process theory. See also van der Vaart and Wellner (1996) and Kosorok (2008) for a general theory of M-estimation via empirical processes. However, these works do not cover the estimation problem in (1) due to their focus on cross-sectional data among others. It should be emphasized that this is not a pathological example. We provide several other relevant examples in Section 3 and the supplementary material, which include well-known Honoré and Kyriazidou's (2000) estimator for a dynamic panel discrete choice model. Furthermore, we propose a new binary choice model with random coefficients and analyze a localized maximum score estimator in Section 3.2.

This paper covers a broader class of M-estimators than the above examples suggest. The baseline scenario above (called *local M-estimation* due to the localization at $x = c$) is generalized into two directions. First, we accommodate not only variables taking limited values (e.g., interval-valued data) which typically lead to estimation of a set rather than a point, but also nuisance parameters with growing dimension. Set estimation problems due to limited data are also known as partial identification problems in econometrics literature (e.g., Manski and Tamer, 2002). It is also novel to accommodate high-dimensional nuisance parameters in the M-estimation with discontinuous criterion functions. Second, we allow for the localization to be dependent on the parameter θ instead of a prespecified value c. For instance, the criterion function may take the form of $\sum_{t=1}^{n} \mathbb{I}\{|y_t - \theta| \leq h_n\}$ with $h_n \to 0$. Relevant examples include mode estimation (Chernoff, 1964, and Lee, 1989) and the Hough transform estimator in image analysis (Goldenshluger and Zeevi, 2004). Henceforth, we call this case *parameter-dependent local M-estimation*. Parameter-dependence brings some new feature in our asymptotic analysis but in a different way from a classical example of parameter-dependency on support such as the maximum likelihood estimator for Uniform$[0, \theta]$.

The main contribution of this paper is to develop a general asymptotic theory for such M-estimation problems. Our theoretical results cover all the examples above and can be used to establish limit laws for point estimators and convergence rates for set estimators. To this end, we develop certain maximal inequalities, which enable us to obtain nonparametric cube root rates of $(nh_n)^{1/3}$, $\{nh_n/\log(nh_n)\}^{1/3}$, and $(nh_n^2)^{1/3}$, for the local M-estimation, limited variable case, and parameter-dependent case, respectively. These inequalities are extended to establish stochastic

asymptotic equicontinuity of normalized processes of the criterion functions so that an argmax theorem delivers limit laws of the M-estimators. It is worth noting that all the conditions are characterized through moment conditions that can be easily verified, as illustrated in the examples. Thus, our results can be applied without prior knowledge on empirical process theory. It is often not trivial to verify the entropy conditions such as uniform manageability in Kim and Pollard (1990). Particularly for dependent data, the covering or bracketing numbers often need to be calculated using a norm that hinges on the mixing coefficients and distribution of the data (e.g., the $L_{2,\beta}$-norm in Doukhan, Massart and Rio, 1995).

Another contribution is that we allow for weakly dependent data, which are associated with exponential mixing decay rates of the absolutely regular processes. In some applications, the cube root asymptotics has been extended to time series data such as Anevski and Hössjer (2006) for monotone density estimation, Zinde-Walsh (2002) for least median of squares, de Jong and Woutersen (2011) for maximum score, and Koo and Seo (2015) for the break estimation under misspecification. However, it is not clear whether they are able to handle a general class of estimation problems in this paper.

The paper is organized as follows. Section 2 develops an asymptotic theory for the local M-estimation, and Section 3 provides some examples. In Section 4, we generalize the asymptotic theory to the cases of limited variables (Section 4.1) and parameter-dependent localization (Section 4.2). Section 5 concludes. All proofs, detailed illustrations of our asymptotic theory for the examples in Section 3, and some additional examples are contained in the supplementary material.

2. Local M-estimation with discontinuous criterion function

Let us consider the M-estimator $\hat{\theta}$ that maximizes

$$\mathbb{P}_n f_{n,\theta} = \frac{1}{n}\sum_{t=1}^{n} f_{n,\theta}(z_t),$$

where $\{f_{n,\theta} : \theta \in \Theta\}$ is a sequence of criterion functions indexed by the parameters $\theta \in \Theta \subseteq \mathbb{R}^d$ and $\{z_t\}$ is a strictly stationary sequence of random variables with marginal P. We introduce a set of conditions for $f_{n,\theta}$, which induces a possibly localized counterpart of Kim and Pollard's (1990) cube root asymptotics. Their cube root asymptotics can be viewed as a special case of ours, where $f_{n,\theta} = f_\theta$ for all n. Let $Pf = \int f dP$ for a function f, $|\cdot|$ be the Euclidean norm of a vector, and $\|\cdot\|_2$ be the $L_2(P)$-norm of a random variable. The class of criterion functions of our interest is characterized as follows.

Assumption M. *For a sequence $\{h_n\}$ of positive numbers with $nh_n \to \infty$, $\{f_{n,\theta} : \theta \in \Theta\}$ satisfies the following conditions.*

(i): *$\{h_n f_{n,\theta} : \theta \in \Theta\}$ is a class of uniformly bounded functions. $\lim_{n\to\infty} Pf_{n,\theta}$ is uniquely maximized at θ_0. $Pf_{n,\theta}$ is twice continuously differentiable at θ_0 for all n large enough and*

admits the expansion

$$P(f_{n,\theta} - f_{n,\theta_0}) = \frac{1}{2}(\theta - \theta_0)' V (\theta - \theta_0) + o(|\theta - \theta_0|^2) + o((nh_n)^{-2/3}), \tag{2}$$

for a negative definite matrix V.

(ii): *There exist positive constants C and C' such that*

$$|\theta_1 - \theta_2| \leq C h_n^{1/2} \left\| f_{n,\theta_1} - f_{n,\theta_2} \right\|_2,$$

for all n large enough and all $\theta_1, \theta_2 \in \{\theta \in \Theta : |\theta - \theta_0| \leq C'\}$.

(iii): *There exists a positive constant C'' such that*

$$P \sup_{\theta \in \Theta : |\theta - \theta'| < \varepsilon} h_n |f_{n,\theta} - f_{n,\theta'}|^2 \leq C'' \varepsilon, \tag{3}$$

for all n large enough, $\varepsilon > 0$ small enough, and θ' in a neighborhood of θ_0.

Although we are primarily interested in the case of $h_n \to 0$, we do not exclude the case of $h_n = 1$. When $h_n \to 0$, $\{h_n\}$ is usually a sequence of bandwidths for localization. Although we cover Kim and Pollard's setup as a special case, our conditions appear somewhat different from theirs. In fact, our conditions consist of directly verifiable moment conditions without resorting to the notion of empirical process theory such as uniform manageability.

Assumption M (i) contains boundedness, point identification of θ_0, and a quadratic approximation of $Pf_{n,\theta}$ at $\theta = \theta_0$. Boundedness of $\{h_n f_{n,\theta}\}$ is a major requirement, but is satisfied for all examples in this paper and Kim and Pollard (1990). In Section 4, we relax the assumption of point identification of θ_0. When the criterion function involves kernel smoothing for localization, it typically takes the form of $f_{n,\theta}(x,y) = \frac{1}{h_n} K\left(\frac{x-c}{h_n}\right) m(x,y,\theta)$ (see (1) and examples in Section 3).

Despite of discontinuity of $f_{n,\theta}$, the population criterion function $Pf_{n,\theta}$ is smooth and approximated by a quadratic function as in (2). This distinguishes our estimation problem from that of a change-point in the regression, which also involves a discontinuous criterion function but the change-point estimator is super-consistent (e.g., Chan, 1993) unless the estimating equation is misspecified as in the split point estimator for decision trees (Bühlmann and Yu, 2002, and Banerjee and McKeague, 2007).

Assumption M (ii) is used to relate the $L_2(P)$-norm $\|f_{n,\theta} - f_{n,\theta_0}\|_2$ for the criterion functions to the Euclidean norm $|\theta - \theta_0|$ for the parameters. This condition is implicit in Kim and Pollard (1990, Condition (v)) under independent observations and is often verified in the course of checking the expansion in (2).

Assumption M (iii), an envelope condition for the class $\mathcal{F}_n = \{f_{n,\theta} - f_{n,\theta'} : |\theta - \theta'| \leq \varepsilon\}$, plays a key role for the cube root asymptotics. It should be noted that for the familiar squared root asymptotics, the upper bound in (3) is of order ε^2 instead of ε. It is often the case that verifying the envelope condition for arbitrary θ' in a neighborhood of θ_0 is not more demanding than that for θ_0.

In particular, Assumption M (iii) is used to guarantee an integrability condition on the metric entropy with bracketing for $\mathcal{F}_n$ in the $L_{2,\beta}$-norm so that the maximal inequality in Doukhan, Massart and Rio (1995, Theorem 3) can be applied to establish Lemma M below. On the other hand, Kim and Pollard (1990) used the concept of uniform manageability (Pollard, 1989) to control the size of $\mathcal{F}_n$, which is defined by the ϵ-capacity or metric entropy by covering numbers. Generally the bracketing and covering numbers approaches are not directly comparable (see Section 2.5 of van der Vaart and Wellner, 1996, for example). It would be interesting to explore how the symmetrization argument combined with certain manageability concept can be applied in our setup.

We now study the limiting behavior of the M-estimator, which is precisely defined as a random variable $\hat{\theta}$ satisfying

$$\mathbb{P}_n f_{n,\hat{\theta}} \geq \sup_{\theta\in\Theta} \mathbb{P}_n f_{n,\theta} - o_p((nh_n)^{-2/3}). \tag{4}$$

The first step is to establish weak consistency $\hat{\theta} \xrightarrow{p} \theta_0$, which is rather standard and usually shown by establishing the uniform convergence $\sup_{\theta\in\Theta} |\mathbb{P}_n f_{n,\theta} - P f_{n,\theta}| \xrightarrow{p} 0$. Thus, this section simply assumes the consistency of $\hat{\theta}$. See Section 3 and the supplementary material for some illustrations.

The next step is to derive the convergence rate of $\hat{\theta}$. A key ingredient for this step is to obtain the modulus of continuity of the centered empirical process $\{\mathbb{G}_n h_n^{1/2}(f_{n,\theta} - f_{n,\theta_0}) : \theta \in \Theta\}$ by certain maximum inequality, where $\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n f - Pf)$ for a function f. If $f_{n,\theta}$ does not vary with n and $\{z_t\}$ is independent, several maximal inequalities are available in the literature (e.g., page 199 of Kim and Pollard, 1990). If $f_{n,\theta}$ varies with n and $\{z_t\}$ is dependent, to the best of our knowledge, there is no maximal inequality which can be applied to the class of functions in Assumption M. Our first task is to establish such a maximal inequality.

To proceed, we now characterize the dependence structure of the data. Among several notions of dependence, this paper focuses on an absolutely regular process. See Doukhan, Massart and Rio (1995) for a discussion on empirical process theory of absolutely regular processes. Let $\mathcal{F}_{-\infty}^0$ and $\mathcal{F}_m^\infty$ be σ-fields of $\{\ldots, z_{t-1}, z_0\}$ and $\{z_m, z_{m+1}, \ldots\}$, respectively. Define the β-mixing coefficient as $\beta_m = \frac{1}{2}\sup \sum_{(i,j)\in I\times J} |P\{A_i \cap B_j\} - P\{A_i\}P\{B_j\}|$, where the supremum is taken over all the finite partitions $\{A_i\}_{i\in I}$ and $\{B_j\}_{j\in J}$ respectively $\mathcal{F}_{-\infty}^0$ and $\mathcal{F}_m^\infty$ measurable. Throughout the paper, we maintain the following assumption on the data $\{z_t\}$.

Assumption D. *$\{z_t\}$ is a strictly stationary and absolutely regular process with β-mixing coefficients $\{\beta_m\}$ such that $\beta_m = O(\rho^m)$ for some $0 < \rho < 1$.*

This assumption obviously covers the case of independent observations and says the mixing coefficient β_m should decay at an exponential rate.[1] For example, various Markov, GARCH, and stochastic volatility models satisfy this assumption (Carrasco and Chen, 2002).

The maximal inequality for the empirical process $\mathbb{G}_n h_n^{1/2}(f_{n,\theta} - f_{n,\theta_0})$ is presented as follows.

[1]Indeed, polynomial decays of β_m are often associated with strong dependence and long memory type behaviors in sample statistics. See Chen, Hansen and Carrasco (2010) and references therein. In this case, asymptotic analysis for the M-estimator will become very different.

Lemma M. *Under Assumptions M and D, there exist positive constants C and C' such that*

$$P \sup_{\theta\in\Theta:|\theta-\theta_0|<\delta} |\mathbb{G}_n h_n^{1/2}(f_{n,\theta} - f_{n,\theta_0})| \leq C\delta^{1/2},$$

for all n large enough and $\delta \in [(nh_n)^{-1/2}, C']$.

This lemma implies the following result.

Lemma 1. *Under Assumptions M and D, for each $\varepsilon > 0$, there exist random variables $\{R_n\}$ of order $O_p(1)$ and a positive constant C such that*

$$|\mathbb{P}_n(f_{n,\theta} - f_{n,\theta_0}) - P(f_{n,\theta} - f_{n,\theta_0})| \leq \varepsilon|\theta - \theta_0|^2 + (nh_n)^{-2/3}R_n^2,$$

for all $\theta \in \Theta$ satisfying $(nh_n)^{-1/3} \leq |\theta - \theta_0| \leq C$.

We now derive the convergence rate of $\hat{\theta}$. Suppose $|\hat{\theta} - \theta_0| \geq (nh_n)^{-1/3}$. Then by (4), Lemma 1, and Assumption M (i), we can take a positive constant c such that

$$\begin{aligned} o_p((nh_n)^{-2/3}) &\leq \mathbb{P}_n(f_{n,\hat{\theta}} - f_{n,\theta_0}) \leq P(f_{n,\hat{\theta}} - f_{n,\theta_0}) + \varepsilon|\hat{\theta} - \theta_0|^2 + (nh_n)^{-2/3}R_n^2 \\ &\leq (-c + \varepsilon)|\hat{\theta} - \theta_0|^2 + o(|\hat{\theta} - \theta_0|^2) + O_p((nh_n)^{-2/3}), \end{aligned}$$

for each $\varepsilon > 0$. Taking ε small enough to satisfy $c - \varepsilon > 0$ yields the convergence rate $\hat{\theta} - \theta_0 = O_p((nh_n)^{-1/3})$.

Given the convergence rate, the final step is to establish the limiting distribution of $\hat{\theta}$. To this end, we apply a continuous mapping theorem of an argmax element (e.g., Theorem 2.7 of Kim and Pollard, 1990). A key ingredient for this argument is to show weak convergence of the standardized empirical process

$$Z_n(s) = n^{1/6}h_n^{2/3}\mathbb{G}_n(f_{n,\theta_0+s(nh_n)^{-1/3}} - f_{n,\theta_0}),$$

for $|s| \leq K$ with any $K > 0$. Weak convergence of Z_n may be characterized by its finite dimensional convergence and tightness (or stochastic asymptotic equicontinuity). If $f_{n,\theta}$ does not vary with n and $\{z_t\}$ is independent as in Kim and Pollard (1990), a classical central limit theorem combined with the Cramér-Wold device implies finite dimensional convergence, and a maximal inequality on a suitably regularized class of functions guarantees tightness of the process of criterion functions. We adapt this approach to our local M-estimation problem with dependent observations satisfying Assumption D.

Consider a function $\beta(\cdot)$ such that $\beta(t) = \beta_{[t]}$ if $t \geq 1$ and $\beta(t) = 1$ otherwise and denote its càdlàg inverse by $\beta^{-1}(\cdot)$. Let $Q_g(u)$ be the inverse function of the tail probability function $x \mapsto P\{|g(z_t)| > x\}$. For finite dimensional convergence, we employ Rio's (1997, Corollary 1) central limit theorem for α-mixing arrays to our setup.

Lemma C. *Suppose Assumption D holds true, $Pg_n = 0$, and*

$$\sup_{n\in\mathbb{N}} \int_0^1 \beta^{-1}(u)Q_{g_n}(u)^2 du < \infty. \tag{5}$$

Then $\Sigma = \lim_{n\to\infty} \mathrm{Var}(\mathbb{G}_n g_n)$ *exists and* $\mathbb{G}_n g_n \xrightarrow{d} N(0, \Sigma)$.

The finite dimensional convergence of Z_n follows from Lemma C by setting g_n as any finite dimensional projection of the process $\{g_{n,s} - Pg_{n,s} : s\}$ with $g_{n,s} = n^{1/6}h_n^{2/3}(f_{n,\theta_0+s(nh_n)^{-1/3}} - f_{n,\theta_0})$. The requirement in (5) is the Lindeberg-type condition in Rio (1997, Corollary 1). This condition excludes polynomial decay of β_m. Noted that for criterion functions satisfying Assumption M, the $(2+\delta)$-th moments $P|g_{n,s}|^{2+\delta}$ for $\delta > 0$ typically diverge because $g_{n,s}$ usually involves indicator functions. To verify (5), the following lemma is often useful.

Lemma 2. *Suppose Assumptions M and D hold true, and there is a positive constant* c *such that*

$$P\{|g_{n,s}| \geq c\} \leq c(nh_n^{-2})^{-1/3}, \tag{6}$$

for all n *large enough and* s. *Then (5) holds true.*

In our examples, $g_{n,s}$ is zero or close to zero with high probability so that (6) is easily satisfied. See Section 3.1 for an illustration.

We provide another maximal inequality that is useful to establish tightness of the process Z_n.

Lemma M'. *Suppose Assumption D holds true. Consider a sequence of classes of functions* $\mathcal{G}_n = \{g_{n,s} : |s| \leq K\}$ *for some* $K > 0$ *with envelope functions* G_n. *Suppose there is a positive constant* C *such that*

$$P \sup_{s:|s-s'|<\varepsilon} |g_{n,s} - g_{n,s'}|^2 \leq C\varepsilon, \tag{7}$$

for all n *large enough,* $|s'| \leq K$, *and* $\varepsilon > 0$ *small enough. Also, assume that there exist* $0 \leq \kappa < 1/2$ *and* $C' > 0$ *such that* $G_n \leq C'n^{\kappa}$ *and* $\|G_n\|_2 \leq C'$ *for all* n *large enough. Then for any* $\sigma > 0$, *there exist* $\delta > 0$ *and a positive integer* N_δ *such that*

$$P \sup_{(s,s'):|s-s'|<\delta} |\mathbb{G}_n(g_{n,s} - g_{n,s'})| \leq \sigma,$$

for all $n \geq N_\delta$.

Tightness of Z_n is implied from Lemma M' by setting

$$g_{n,s} = n^{1/6}h_n^{2/3}(f_{n,\theta_0+s(nh_n)^{-1/3}} - f_{n,\theta_0}).$$

Note that (7) is satisfied by Assumption M (iii).[2] Compared to Lemma M used to derive the convergence rate of $\hat{\theta}$, Lemma M' is applied only to establish tightness of the process Z_n. Therefore, we do not need an exact decay rate on the right hand side of the maximal inequality.[3]

Finite dimensional convergence and tightness of Z_n imply its weak convergence. Then the continuous mapping theorem for an argmax element (Theorem 2.7 of Kim and Pollard, 1990) yields the limiting distribution of $\hat{\theta}$. The main theorem of this section is presented as follows.

[2]The upper bound in (7) can be relaxed to $\varepsilon^{1/p}$ for $1 \leq p < \infty$. However, it is typically satisfied with $p = 1$ for the examples we consider.

[3]In particular, Z_n itself does not satisfy Assumption M (ii).

Theorem 1. *Suppose that Assumptions M and D hold, $\hat{\theta}$ defined in (4) converges in probability to $\theta_0 \in \text{int}\Theta$, and (5) holds with $g_{n,s} - Pg_{n,s}$ for each s. Then*

$$(nh_n)^{1/3}(\hat{\theta} - \theta_0) \xrightarrow{d} \arg\max_{s\in\mathbb{R}^d} Z(s), \tag{8}$$

where $Z(s)$ is a Gaussian process with continuous sample paths, expected value $s'Vs/2$, and covariance kernel $H(s_1, s_2) = \lim_{n\to\infty} \sum_{t=-n}^{n} \text{Cov}(g_{n,s_1}(z_0), g_{n,s_2}(z_t)) < \infty$.

This theorem can be considered as an extension of the main theorem of Kim and Pollard (1990) to the case where the criterion function $f_{n,\theta}$ can vary with the sample size and the observations $\{z_t\}$ can obey a dependent process. To the best of our knowledge, the (nonparametric) cube root convergence rate $(nh_n)^{1/3}$ is new in the literature. It is interesting to note that similar to standard nonparametric estimation, nh_n still plays a role as the "effective sample size."

2.1. **Nuisance parameters.** It is often the case that the criterion function contains some nuisance parameters, which can be estimated with rates faster than $(nh_n)^{-1/3}$ such as $\hat{\nu}$ in (1). For the rest of this section, let $\hat{\theta}$ and $\tilde{\theta}$ satisfy

$$\begin{aligned}
\mathbb{P}_n f_{n,\hat{\theta},\hat{\nu}} &\geq \sup_{\theta\in\Theta} \mathbb{P}_n f_{n,\theta,\hat{\nu}} + o_p((nh_n)^{-2/3}), \\
\mathbb{P}_n f_{n,\tilde{\theta},\nu_0} &\geq \sup_{\theta\in\Theta} \mathbb{P}_n f_{n,\theta,\nu_0} + o_p((nh_n)^{-2/3}),
\end{aligned}$$

respectively, where ν_0 is a vector of nuisance parameters and $\hat{\nu}$ is its estimator satisfying $\hat{\nu} - \nu_0 = o_p((nh_n)^{-1/3})$. Theorem 1 is extended as follows.

Theorem 2. *Suppose Assumption D holds true. Let $\{f_{n,\theta,\nu_0} : \theta \in \Theta\}$ satisfy Assumption M and $\{f_{n,\theta,\nu} : \theta \in \Theta, \nu \in \Lambda\}$ satisfy Assumption M (iii). Suppose there exists some negative definite matrix V_1 such that*

$$P(f_{n,\theta,\nu} - f_{n,\theta_0,\nu_0}) = \frac{1}{2}(\theta - \theta_0)'V_1(\theta - \theta_0) + o(|\theta - \theta_0|^2) + O(|\nu - \nu_0|^2) + o((nh_n)^{-2/3}), \tag{9}$$

for all θ and ν in neighborhoods of θ_0 and ν_0, respectively. Then $\hat{\theta} = \tilde{\theta} + o_p((nh_n)^{-1/3})$. Additionally, if (5) holds with $(g_{n,s} - Pg_{n,s})$ for each s with $g_{n,s}$ being $n^{1/6}h_n^{2/3}(f_{n,\theta_0+s(nh_n)^{-1/3},\nu_0} - f_{n,\theta_0,\nu_0})$, then

$$(nh_n)^{1/3}(\hat{\theta} - \theta_0) \xrightarrow{d} \arg\max_{s\in\mathbb{R}^d} Z(s),$$

where $Z(s)$ is a Gaussian process with continuous sample paths, expected value $s'V_1s/2$ and covariance kernel $H(s_1, s_2) = \lim_{n\to\infty} \sum_{t=-n}^{n} \text{Cov}(g_{n,s_1}(z_0), g_{n,s_2}(z_t)) < \infty$.

A key step for the proof of this theorem is to confirm that the empirical process $\mathbb{G}_n f_{n,\theta,\nu_0+c(nh_n)^{-1/3}}$ is well approximated by $\mathbb{G}_n f_{n,\theta,\nu_0}$ over $|\theta - \theta_0| \leq \epsilon$ and $|c| \leq \epsilon$ (see, (A.10) in the supplementary material). This is shown by applying Lemma M' with $g_{n,s} = n^{1/6}h_n^{2/3}(f_{n,\theta,c(nh_n)^{-1/3}} - f_{n,\theta,\nu_0})$. The condition (7) in Lemma M' demands more precise control on the size of the envelope for the class

of $g_{n,s}$ than the comparable condition in the Z-estimation with nuisance parameters (e.g., eq. (3) of van der Vaart and Wellner, 2007).

2.2. **Discussions.**

2.2.1. *Inference.* Once we show that the M-estimator has a proper limiting distribution in Theorem 1 or 2, Politis, Romano and Wolf (1999, Theorem 3.3.1) justify the use of subsampling to construct confidence intervals. Since Assumption D satisfies the requirement of their theorem, subsampling inference based on s consecutive observations with $s/n \to \infty$ is asymptotically valid (in a pointwise sense explained below). See Politis, Romano and Wolf (1999, Section 3.6) for a discussion on data-dependent choices of s.

We note that the asymptotic validity of subsampling inference mentioned above is in a pointwise sense rather than uniform. To be specific, suppose that $\{z_t\}$ is an independent and identically distributed (iid) sample from the probability measure P that belongs to a class of probability measures $\mathcal{P}$. Also denote the true parameters by $\theta_0(P)$ to make explicit the dependence on P. Based on Romano and Shaikh (2008), the confidence set $\mathcal{C}_n$ for $\theta_0(P)$ is called *pointwise* valid in $(1-\alpha)$ level if

$$\liminf_{n\to\infty} P\{\theta_0(P) \in \mathcal{C}_n\} \geq 1-\alpha,$$

for each $P \in \mathcal{P}$, and is called *uniformly* valid in $(1-\alpha)$ level if

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}} P\{\theta_0(P) \in \mathcal{C}_n\} \geq 1-\alpha.$$

Our Theorems 1 and 2 combined with Politis, Romano and Wolf (1999, Theorem 3.3.1) guarantee the pointwise validity of the subsampling confidence set based on the quantiles of the subsample statistic $(sh_s)^{1/3}(\hat{\theta}_s - \hat{\theta})$, where s is the size of subsamples, and $\hat{\theta}_s$ and $\hat{\theta}$ are the M-estimators based on the subsample and full sample, respectively. Also the pointwise valid confidence interval for each element of $\theta_0(P)$ can be obtained in a similar manner.

To investigate whether we can construct a uniformly valid confidence set in our setup, we assume that $\{z_t\}$ is iid and the distribution $J_n(\cdot,\theta,P)$ of $Q_n(\theta) = (nh_n)^{2/3}\{\max_{\vartheta\in\Theta} \mathbb{P}_n f_{n,\vartheta} - \mathbb{P}_n f_{n,\theta}\}$ satisfies

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta} \sup_{P\in\mathcal{P}:\theta=\theta_0(P)} \sup_{x\in\mathbb{R}} \{J_s(x,\theta,P) - J_n(x,\theta,P)\} \leq 0, \tag{10}$$

Then, Romano and Shaikh (2008, Theorems 3.1 and 3.3) imply the uniform validity of the confidence set

$$\mathcal{C}_n = \{\theta \in \Theta : Q_n(\theta) \leq q_s(\theta, 1-\alpha)\},$$

over $\mathcal{P}$, where $q_s(\theta, 1-\alpha)$ is the $(1-\alpha)$-th quantile of the distribution of the subsample statistic $Q_s(\theta)$. By inspection of Romano and Shaikh (2008), we can see that (10) is satisfied if $Q_n(\theta_0(P_n))$ converges in law to a unique continuous distribution for any sequence of $P_n \in \mathcal{P}$ yielding a row-wise iid triangular array. Our lemmas to obtain Theorem 1 can be readily extended to the array setting by restating Assumptions M and D and the additional conditions for Theorem 1 in the array setup.

We note that the computation of $\mathcal{C}_n$ may require an extensive numerical search over $\theta \in \Theta$, where the quantile $q_s(\theta, 1-\alpha)$ needs to be computed for each θ.

The above uniformity result relies upon the general results in Romano and Shaikh (2008, Theorems 3.1 and 3.3), and there are at least three issues to be further considered. First, the iid assumption for the sample does not allow serial dependence as in Assumption D. To accommodate time series data, Romano and Shaikh (2008, Theorems 3.1) that provide high-level assumptions for uniform validity should be modified. Second, it is not a trivial task to extend the results in Romano and Shaikh (2008) to inference on subvectors (or functions) of θ except for a conservative projection of $\mathcal{C}_n$ to lower dimension. Third, a key result in Romano and Shaikh (2008, Theorems 3.1) holds for objects centered at the true parameter $\theta_0(P)$ instead of the estimator $\hat{\theta}$. Therefore, their result does not apply to the subsample statistic $(sh_s)^{1/3}(\hat{\theta}_s - \hat{\theta})$. All of these issues require full length papers and are beyond the scope of this paper.

Another candidate to conduct inference based on the M-estimator is the bootstrap. However, even for independent observations, it is known that the naive nonparametric bootstrap is typically invalid under the cube root asymptotics (Abrevaya and Huang, 2005, and Sen, Banerjee and Woodroofe, 2010).

2.2.2. *Generalization of Assumption D.* All the results in this section build upon Assumption D that requires $\{z_t\}$ to be strictly stationary and absolutely regular (or β-mixing) with exponential decaying mixing coefficients. Assumption D is used for both the maximal inequality (Lemma M) and central limit theorem (Lemma C), which are building blocks to derive the asymptotic distribution of $\hat{\theta}$. It is of interest whether we can establish analogous results under more general setups, such as α-mixing, by utilizing some recent developments in the empirical process theory for dependent data. For instance, Merlevède, Peligrad and Rio (2009, 2011) obtained Bernstein type inequalities for α-mixing processes. Baraud (2010) and Nickl and Söhl (2015, Section 3) explored the generic chaining argument by Talagrand (2005) for Markov chains.

Since the central limit theorem in Rio (1997, Corollary 1) holds for α-mixing arrays, we can modify Lemma C to accommodate α-mixing processes. Thus, we focus on extending Lemma M, the maximal inequality. A crucial step for this extension is whether we can replace the key lemma in Doukhan, Massart and Rio (1995, Lemma 3), which leads to the maximal inequality for β-mixing (in eq. (6) of the supplementary material) through a chaining argument. Specifically, consider a finite subclass $\mathcal{F}$ of bounded functions with cardinality $p \geq \exp(1)$. By a decoupling technique for β-mixing, Doukhan, Massart and Rio (1995, Lemma 3) showed that for positive constants c and c_1, there exists a universal positive constant C such that

$$P \max_{f \in \mathcal{F}} |\mathbb{G}_n f| \leq C \left(c\sqrt{\log p} + c_1 q \frac{\log p}{\sqrt{n}} + c_1 \beta_q \sqrt{n} \right),$$

for all $q = 1, \ldots, n$. Note that the above upper bound reduces to the first term $Cc\sqrt{\log p}$ for the iid case. By properly choosing q, the first term still dominates in the β-mixing case even if $\log p$

is close to n so that Lemma M can be established. In contrast, the maximal inequality implied by Merlevède, Peligrad and Rio (2009, (2.1) in Theorem 1) for α-mixing would be written in the form of $C\left(c\sqrt{\log p}+c_1\log n\log\log n\frac{\log p}{\sqrt{n}}\right)$. Therefore, as $\log p$ becomes close to n, the second term will dominate. Since this order of cardinality p (i.e., $\log p$ close to n) is required in the proof of Doukhan, Massart and Rio (1995, Theorem 2), the upper bound in Lemma M for α-mixing would become larger.[4]

Another direction to extend our result is to accommodate Markov chains. To this end, the chaining argument (see, Baraud, 2010, and Nickl and Söhl, 2015) based on the Bernstein type inequalities for Markov chains (e.g., Adamczak, 2008, and Paulin, 2015) may yield an analog of Lemma M. Although this is an intriguing question, existing time series examples on the cube root asymptotics mostly focus on mixing data (e.g., Polonik and Yao, 2000, and de Jong and Woutersen, 2011) and also typically involve additional conditioning or exogenous variables. Thus, we leave this extension for future work.

3. Examples

We provide several examples to demonstrate the usefulness of the asymptotic theory in the last section. For the sake of space, we only sketch the arguments to verify the conditions to apply the theorems in Section 2. Detailed verifications under primitive conditions are delegated to the supplementary material. Also, the supplementary material contains two additional examples (dynamic least median of squares and monotone density estimation) which are omitted for brevity.

3.1. **Dynamic panel discrete choice.** For a binary response y_{it} and k-dimensional covariates x_{it}, consider a dynamic panel data model

$$\begin{aligned} P\{y_{i0}=1|x_i,\alpha_i\} &= F_0(x_i,\alpha_i), \\ P\{y_{it}=1|x_i,\alpha_i,y_{i0},\ldots,y_{it-1}\} &= F(x_{it}'\beta_0+\gamma_0 y_{it-1}+\alpha_i), \end{aligned}$$

for $i=1,\ldots,n$ and $t=1,2,3$, where α_i is unobservable and both F_0 and F are unknown. Honoré and Kyriazidou (2000) proposed the conditional maximum score estimator for (β_0,γ_0)

$$(\hat\beta,\hat\gamma)=\arg\max_{(\beta,\gamma)\in\Theta}\sum_{i=1}^{n}K\left(\frac{x_{i2}-x_{i3}}{b_n}\right)(y_{i2}-y_{i1})\mathrm{sgn}\{(x_{i2}-x_{i1})'\beta+(y_{i3}-y_{i0})\gamma\},$$

where K is a kernel function and b_n is a bandwidth. Kernel smoothing is introduced to deal with the unknown link function F. Honoré and Kyriazidou (2000) obtained consistency of this estimator but the convergence rate and limiting distribution are unknown. Since the criterion function varies with the sample size through the bandwidth b_n, the cube root asymptotic theory of Kim and Pollard (1990) is not applicable here.

[4]Although full investigation is beyond the scope of this paper, we conjecture that it is also the case for the generic chaining argument by Talagrand (2005). Indeed, Talagrand (2005, eq. (1.9) in p. 10) explains that the generic chaining needs partitions of cardinality up to 2^{2^n}.

This open question can be addressed by Theorem 1. Let $z = (z_1', z_2, z_3')'$ with $z_1 = x_2 - x_3$, $z_2 = y_2 - y_1$, and $z_3 = ((x_2 - x_1)', y_3 - y_0)'$. The above estimator for $\theta_0 = (\beta_0', \gamma_0)'$ can be written as the M-estimator using the criterion function

$$f_{n,\theta}(z) = e_n(z)(\mathbb{I}\{z_3'\theta \geq 0\} - \mathbb{I}\{z_3'\theta_0 \geq 0\}), \tag{11}$$

where $e_n(z) = b_n^{-k} K(b_n^{-1} z_1) z_2$. To apply Theorem 1, it is enough to show that $f_{n,\theta}$ in (11) satisfies Assumption M with $h_n = b_n^k$ and (6). Then the limiting distribution of Honoré and Kyriazidou's (2000) estimator is obtained as in (8).

Here we sketch the verification. See the supplementary material (Section B.1) for a detail and primitive conditions. For Assumption M (i), $\{h_n f_{n,\theta}\}$ is bounded for the kernel K bounded, and (2) is obtained by a Taylor expansion combined with the argument in Kim and Pollard (1990, pp. 214-215). For Assumption M (ii), take any θ_1 and θ_2 and note that

$$\begin{aligned} h_n^{1/2} \left\| f_{n,\theta_1} - f_{n,\theta_2} \right\|_2 &= \left(P \left\{ h_n E[e_n(z)^2 | z_3] | \mathbb{I}\{z_3'\theta_1 \geq 0\} - \mathbb{I}\{z_3'\theta_2 \geq 0\} | \right\} \right)^{1/2} \\ &\geq \left(cP | \mathbb{I}\{z_3'\theta_1 \geq 0\} - \mathbb{I}\{z_3'\theta_2 \geq 0\} | \right)^{1/2} \\ &\geq c^{1/2} P\{z_3'\theta_1 \geq 0 > z_3'\theta_2 \text{ or } z_3'\theta_2 \geq 0 > z_3'\theta_1\}, \end{aligned}$$

for some $c > 0$, where the first inequality follows from $h_n E[e_n(z)^2 | z_3] > c$ (by a change of variables and condition on the density $z_1 | z_2 \neq 0, z_3$ bounded away from zero) and the second inequality follows by Jensen's inequality. The right hand side is probability for a pair of wedge shaped regions with an angle of order $|\theta_1 - \theta_2|$. Thus, Assumption M (ii) is satisfied if the density of z_3 is bounded away from zero in a neighborhood of 0. Assumption M (iii) can be verified in a similar way (by considering the upper bound instead). The Markov inequality and boundedness of the density imply (6).

3.2. **Random coefficient binary choice.** As a new statistical model which can be covered by our asymptotic theory, let us consider the regression model with a random coefficient $y_t = x_t'\theta(w_t) + u_t$. Suppose we observe x_t, w_t, and only the sign of y_t and wish to estimate $\theta_0 = \theta(c)$ at some given c.[5] We propose a localized version of the maximum score estimator

$$\hat{\theta} = \arg\max_{\theta \in S} \sum_{t=1}^{n} K\left(\frac{w_t - c}{b_n}\right) [\mathbb{I}\{y_t \geq 0, x_t'\theta \geq 0\} + \mathbb{I}\{y_t < 0, x_t'\theta < 0\}], \tag{12}$$

where S is the surface of the unit sphere. Again, the cube root asymptotic theory of Kim and Pollard (1990) is not applicable due to the bandwidth b_n.

[5]Gautier and Kitamura (2013) studied identification and estimation of the random coefficient binary choice model, where $\theta_t = \theta(w_t)$ is unobservable. Here we study the model where heterogeneity in the slope is caused by the observables w_t.

Theorem 1 can be applied to obtain the limiting distribution of this estimator. Note that $\hat{\theta}$ in (12) can be written as the M-estimator using the criterion function

$$f_{n,\theta}(x,w,u) = \frac{1}{h_n} K\left(\frac{w-c}{h_n^{1/k}}\right) h(x,u)[\mathbb{I}\{x'\theta \geq 0\} - \mathbb{I}\{x'\theta_0 \geq 0\}], \tag{13}$$

for $h_n = b_n^k$ and $h(x,u) = \mathbb{I}\{x'\theta_0 + u \geq 0\} - \mathbb{I}\{x'\theta_0 + u < 0\}$. Once we check Assumption M and (6), Theorem 1 implies the limiting distribution.

The verification is sketched as follows. See the supplementary material (Section B.2) for a detail and primitive conditions. Assumption M (i)-(ii) and (6) can be checked similarly as in Section 3.1. Here we verify Assumption M (iii). By the change of variables and $h(x,u)^2 = 1$, there exists a positive constant C' such that

$$\begin{aligned} & P \sup_{\theta\in\Theta:|\theta-\vartheta|<\varepsilon} h_n |f_{n,\theta} - f_{n,\vartheta}|^2 \\ = & \int\int K(s)^2 \sup_{\theta\in\Theta:|\theta-\vartheta|<\varepsilon} |[\mathbb{I}\{x'\theta \geq 0\} - \mathbb{I}\{x'\vartheta \geq 0\}]|^2 p(x, c+sb_n) dx ds \\ \leq & C' E\left[\sup_{\theta\in\Theta:|\theta-\vartheta|<\varepsilon} |[\mathbb{I}\{x'\theta \geq 0\} - \mathbb{I}\{x'\vartheta \geq 0\}]|^2 \middle| w = c \right], \end{aligned}$$

for all $\varepsilon > 0$, ϑ in a neighborhood of θ_0, and n large enough, where p is the joint density of (x_t, w_t). Since the right hand side is the conditional probability for a pair of wedge shaped regions with an angle of order ε, Assumption M (iii) is guaranteed by some boundedness condition on the conditional density of x_t given $w_t = c$.

3.3. **Minimum volume predictive region.** As an illustration of Theorem 2, we now consider the example in (1). Polonik and Yao's (2000) minimum volume predictor for y at $x = c$ in the class $\mathcal{I}$ of intervals at level α is defined as

$$\hat{I} = \arg\min_{S\in\mathcal{I}} \mu(S) \quad \text{s.t. } \hat{P}(S) \geq \alpha,$$

where μ is the Lebesgue measure and $\hat{P}(S) = \sum_{t=1}^n \mathbb{I}\{y_t \in S\} K\left(\frac{x_t - c}{h_n}\right) / \sum_{t=1}^n K\left(\frac{x_t - c}{h_n}\right)$ is the kernel estimator of $P\{y_t \in S | x_t = c\}$. Since $\hat{I}$ is an interval, it can be written as $\hat{I} = [\hat{\theta} - \hat{\nu}, \hat{\theta} + \hat{\nu}]$, where

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \hat{P}([\theta - \hat{\nu}, \theta + \hat{\nu}]), \qquad \hat{\nu} = \inf\left\{ \nu \in \mathbb{R} : \max_{\theta\in\Theta} \hat{P}([\theta - \nu, \theta + \nu]) \geq \alpha \right\}. \tag{14}$$

For notational convenience, assume $\theta_0 = 0$ and $\nu_0 = 1$. By applying Lemma M', the convergence rate of the nuisance parameter estimator is obtained as $\hat{\nu} - 1 = O_p((nh_n)^{-1/2} + h_n^2)$ (see Section B.3 in the supplementary material).

Note that $\hat{\theta}$ in (14) can be written as the M-estimator using the criterion function

$$f_{n,\theta,\hat{\nu}}(y,x) = \frac{1}{h_n} K\left(\frac{x-c}{h_n}\right) [\mathbb{I}\{y \in [\theta - \hat{\nu}, \theta + \hat{\nu}]\} - \mathbb{I}\{y \in [-\hat{\nu}, \hat{\nu}]\}].$$

We apply Theorem 2 to obtain the convergence rate of $\hat{\theta}$. Details are provided in the supplementary material (Section B.3). Assumptions M for $f_{n,\theta,1}$ and M (iii) for $f_{n,\theta,\nu}$ are verified similarly as in Sections 3.1 and 3.2. To check (9), a Taylor expansion yields

$$\begin{aligned} & P(f_{n,\theta,\nu} - f_{n,0,1}) = P(f_{n,\theta,\nu} - f_{n,0,\nu}) + P(f_{n,0,\nu} - f_{n,0,1}) \\ = \; & -\frac{1}{2}\{-\dot{\gamma}_{y|x}(1|c) + \dot{\gamma}_{y|x}(-1|c)\}\gamma_x(c)\theta^2 + \{\dot{\gamma}_{y|x}(1|c) + \dot{\gamma}_{y|x}(-1|c)\}\gamma_x(c)\theta\nu + o(\theta^2 + |\nu-1|^2) + O(h_n^2), \end{aligned}$$

where γ and $\dot{\gamma}$ mean the density and its derivative, respectively. Also, (9) holds with $V_1 = \{\dot{\gamma}_{y|x}(1|c) - \dot{\gamma}_{y|x}(-1|c)\}\gamma_x(c)$.

Therefore, Theorem 2 implies $\hat{\theta} - \theta_0 = O_p((nh_n)^{-1/3} + h_n)$, which confirms positively the conjecture of Polonik and Yao (2000, Remark 3b) on the exact convergence rate of $\hat{I}$.

3.4. **Dynamic maximum score.** To illustrate the derivation of the covariance kernel H in Theorem 1 for dependent data, we consider the maximum score estimator (Manski, 1975) for the regression model $y_t = x_t'\theta_0 + u_t$, that is

$$\hat{\theta} = \arg\max_{\theta \in S} \sum_{t=1}^{n} [\mathbb{I}\{y_t \geq 0, x_t'\theta \geq 0\} + \mathbb{I}\{y_t < 0, x_t'\theta < 0\}],$$

where S is the surface of the unit sphere. This estimator can be written as the M-estimator using the criterion function

$$f_\theta(x,u) = h(x,u)[\mathbb{I}\{x'\theta \geq 0\} - \mathbb{I}\{x'\theta_0 \geq 0\}],$$

where $h(x,u) = \mathbb{I}\{x'\theta_0 + u \geq 0\} - \mathbb{I}\{x'\theta_0 + u < 0\}$. The conditions to apply Theorem 1 can be verified similarly as in the above examples (see Section B.4 of the supplementary material). Here we focus on the derivation of the covariance kernel for the limiting distribution under Assumption D.

Let $q_{n,t} = f_{\theta_0 + n^{-1/3}s_1}(x_t, u_t) - f_{\theta_0 + n^{-1/3}s_2}(x_t, u_t)$. The covariance kernel is written as $H(s_1, s_2) = \frac{1}{2}\{L(s_1,0) + L(0,s_2) - L(s_1,s_2)\}$, where

$$L(s_1,s_2) = \lim_{n\to\infty} n^{4/3}\mathrm{Var}(\mathbb{P}_n q_{n,t}) = \lim_{n\to\infty} n^{1/3}\left\{\mathrm{Var}(q_{n,t}) + \sum_{m=1}^{\infty}\mathrm{Cov}(q_{n,t}, q_{n,t+m})\right\}.$$

The limit of $n^{1/3}\mathrm{Var}(q_{n,t})$ is given in Kim and Pollard (1990, p. 215). For $\mathrm{Cov}(q_{n,t}, q_{n,t+m})$, we note that $q_{n,t}$ takes only three values, -1, 0, or 1. The definition of β_m and Assumption D imply

$$|P\{q_{n,t} = j, q_{n,t+m} = k\} - P\{q_{n,t} = j\}P\{q_{n,t+m} = k\}| \leq n^{-2/3}\beta_m,$$

for all $n, m \geq 1$ and $j, k = -1, 0, 1$. Thus, $\{q_{n,t}\}$ is a β-mixing array with mixing coefficients bounded by $n^{-2/3}\beta_m$. This in turn implies that $\{q_{n,t}\}$ is an α-mixing array with mixing coefficients bounded by $2n^{-2/3}\beta_m$. By applying the α-mixing inequality, the covariance is bounded as

$$\mathrm{Cov}(q_{n,t}, q_{n,t+m}) \leq Cn^{-2/3}\beta_m \|q_{n,t}\|_p^2,$$

for some $C > 0$ and $p > 2$. Note that

$$\|q_{n,t}\|_p^2 \leq [P|\mathbb{I}\{x'(\theta_0 + s_1 n^{-1/3}) > 0\} - \mathbb{I}\{x'(\theta_0 + s_2 n^{-1/3}) > 0\}|]^{2/p} = O(n^{-2/(3p)}).$$

Combining these results, we get $n^{1/3} \sum_{m=1}^{\infty} \mathrm{Cov}(q_{n,t}, q_{n,t+m}) \to 0$ as $n \to \infty$. Therefore, the covariance kernel H is same as the independent case in Kim and Pollard (1990, p. 215).

3.5. **Other examples.** In the supplementary material, we present additional examples on the dynamic least median of squares estimator (Section B.5) and monotone density estimator (Section B.6).

4. Generalizations

In this section, we consider two generalizations of the asymptotic theory in Section 2. The first concerns data taking limited values such as interval-valued regressors and the second is to allow for the localization to depend on the parameter values.

4.1. **Limited variables.** We consider the case where some of the variables take limited values. In particular, we relax the assumption of point identification of θ_0 and study the case where the limiting criterion function is maximized at any element of a set $\Theta_I \subset \Theta$. The set Θ_I is called the identified set. In order to estimate Θ_I, we consider a collection of approximate maximizers of the sample criterion function

$$\hat{\Theta} = \{\theta \in \Theta : \max_{\theta \in \Theta} \mathbb{P}_n f_{n,\theta} - \mathbb{P}_n f_{n,\theta} \leq \hat{c}(nh_n)^{-1/2}\},$$

i.e., the level set based on $\mathbb{P}_n f_{n,\theta}$ from the maximum with a cutoff value $\hat{c}(nh_n)^{-1/2}$. This section studies the convergence rate of $\hat{\Theta}$ to Θ_I under the Hausdorff distance defined below. We assume that Θ_I is convex. Then the projection $\pi_\theta = \arg\min_{\theta' \in \Theta_I} |\theta' - \theta|$ of $\theta \in \Theta$ on Θ_I is uniquely defined. To deal with the partially identified case, we modify Assumption M as follows.

Assumption S. *For a sequence $\{h_n\}$ of positive numbers satisfying $nh_n \to \infty$, $\{f_{n,\theta} : \theta \in \Theta\}$ satisfies the following conditions.*

(i): *$\{h_n f_{n,\theta} : \theta \in \Theta\}$ is a class of uniformly bounded functions. Also, $\lim_{n\to\infty} P f_{n,\theta}$ is maximized at any θ in a bounded convex set Θ_I. There exist positive constants c and c' such that*

$$P(f_{n,\pi_\theta} - f_{n,\theta}) \geq c|\theta - \pi_\theta|^2 + o(|\theta - \pi_\theta|^2) + o((nh_n)^{-2/3}), \tag{15}$$

for all n large enough and all $\theta \in \{\theta \in \Theta : 0 < |\theta - \pi_\theta| \leq c'\}$.

(ii): *There exist positive constants C and C' such that*

$$|\theta - \pi_\theta| \leq C h_n^{1/2} \|f_{n,\theta} - f_{n,\pi_\theta}\|_2,$$

for all n large enough and all $\theta \in \{\theta \in \Theta : 0 < |\theta - \pi_\theta| \leq C'\}$.

(iii): *There exists a positive constant C'' such that*

$$P \sup_{\theta\in\Theta:0<|\theta-\pi_\theta|<\varepsilon} h_n|f_{n,\theta} - f_{n,\pi_\theta}|^2 \leq C''\varepsilon,$$

for all n large enough and all $\varepsilon > 0$ small enough.

We allow $h_n = 1$ for the case without a bandwidth in the criterion function. Similar comments to Assumption M apply. The main difference is that the conditions are imposed on the contrast $f_{n,\theta} - f_{n,\pi_\theta}$ using the projection π_θ. Assumption S (i) contains boundedness and expansion conditions. The inequality in (15) can be checked by a one-sided Taylor expansion using the directional derivative. Assumption S (ii) and (iii) play similar roles as Assumption M (ii) and (iii) and can be verified by similar arguments to the point identified case.

We establish the following maximal inequality for the criterion functions satisfying Assumption S. Let $r_n = nh_n/\log(nh_n)$.

Lemma MS. *Under Assumptions D and S, there exist positive constants C and $C' < 1$ such that*

$$P \sup_{\theta\in\Theta:0<|\theta-\pi_\theta|<\delta} |\mathbb{G}_n h_n^{1/2}(f_{n,\theta} - f_{n,\pi_\theta})| \leq C(\delta\log(1/\delta))^{1/2},$$

for all n large enough and $\delta \in [r_n^{-1/2}, C']$.

Compared to Lemma M, the additional log term on the right hand side is due to the fact that the supremum is taken over the δ-tube (or manifold) instead of the δ-ball, which increases the entropy. This maximal inequality is applied to obtain the following analog of Lemma 1.

Lemma 3. *Under Assumptions D and S, for each $\varepsilon > 0$, there exist random variables $\{R_n\}$ of order $O_p(1)$ and a positive constant C such that*

$$|\mathbb{P}_n(f_\theta - f_{\pi_\theta}) - P(f_\theta - f_{\pi_\theta})| \leq \varepsilon|\theta - \pi_\theta|^2 + r_n^{-2/3}R_n^2,$$

for all $\theta \in \{\theta \in \Theta : r_n^{-1/3} \leq |\theta - \pi_\theta| \leq C\}$.

Let $\rho(A, B) = \sup_{a\in A}\inf_{b\in B}|a - b|$ and $H(A, B) = \max\{\rho(A, B), \rho(B, A)\}$ be the Hausdorff distance of sets $A, B \subset \mathbb{R}^d$. Based on Lemmas MS and 3, the convergence rate of the set estimator $\hat{\Theta}$ is obtained as follows.

Theorem 3. *Suppose Assumption D holds true. Let $\{f_{n,\theta} : \theta \in \Theta\}$ satisfy Assumption S and $\{h_n^{1/2}f_{n,\theta} : \theta \in \Theta_I\}$ be a P-Donsker class. Assume $H(\hat{\Theta}, \Theta_I) \xrightarrow{p} 0$ and $\hat{c} = o_p((nh_n)^{1/2})$. Then*

$$\rho(\hat{\Theta}, \Theta_I) = O_p(\hat{c}^{1/2}(nh_n)^{-1/4} + r_n^{-1/3}).$$

Furthermore, if $\hat{c} \to \infty$, then $P\{\Theta_I \subset \hat{\Theta}\} \to 1$ and

$$H(\hat{\Theta}, \Theta_I) = O_p(\hat{c}^{1/2}(nh_n)^{-1/4}).$$

Note that ρ is asymmetric in its arguments. In contrast to the convergence rate of $\rho(\hat{\Theta}, \Theta_I)$ obtained in the first part of this theorem, the second part says $P\{\Theta_I \subset \hat{\Theta}\} \to 1$ (i.e., $\rho(\Theta_I, \hat{\Theta})$ can

converge to zero at an arbitrary rate) as far as $\hat{c} \to \infty$. For example, we may set $\hat{c} = \log(nh_n)$. These results are combined to imply the convergence rate $H(\hat{\Theta}, \Theta_I) = O_p(\hat{c}^{1/2}(nh_n)^{-1/4})$ under the Hausdorff distance. The cube root term of order $r_n^{-1/3}$ in the rate of $\rho(\hat{\Theta}, \Theta_I)$ is dominated by the term of order $\hat{c}^{1/2}(nh_n)^{-1/4}$.

We next consider the case where the criterion function contains nuisance parameters. In particular, we allow that the dimension k_n of the nuisance parameters ν can grow as the sample size increases. For instance the nuisance parameters might be coefficients in sieve estimation. It is important to allow the growing dimension of ν to cover Manski and Tamer's (2002) set estimator, where the criterion function contains some nonparametric estimate and its transform by the indicator. The rest of this subsection considers the set estimator

$$\hat{\Theta} = \{\theta \in \Theta : \max_{\theta \in \Theta} \mathbb{P}_n f_{n,\theta,\hat{\nu}} - \mathbb{P}_n f_{n,\theta,\hat{\nu}} \leq \hat{c}(nh_n)^{-1/2}\},$$

with some preliminary estimator $\hat{\nu}$ and cutoff value $\hat{c}$.

To derive the convergence rate of $\hat{\Theta}$, we establish a maximal inequality over a sequence of sets of functions that are indexed by parameters with increasing dimension. Let $g_{n,s} = h_n^{1/2}(f_{n,\theta,\nu} - f_{n,\theta,\nu_0})$ with $s = (\theta', \nu')'$ and consider $\mathcal{G}_n = \{g_{n,s} : |\theta - \pi_\theta| \leq K_1, |\nu - \nu_0| \leq a_n K_2\}$ for some $K_1, K_2 > 0$ with the envelope function $G_n = \sup_{\mathcal{G}_n} |g_{n,s}|$. The maximal inequality in Lemma MS is modified as follows.

Lemma MS'. *Suppose Assumption D holds true. Suppose there exists a positive constant C such that*

$$P \sup_{s:\theta\in\Theta, |\nu-\nu_0|\leq\varepsilon} |g_{n,s}|^2 \leq C\sqrt{k_n}\varepsilon, \tag{16}$$

$$\sup_{s:\theta\in\Theta, |\nu-\nu_0|\leq\varepsilon} |\{|\nu - \nu_0| - C \|g_{n,s}\|_2\} \leq 0, \tag{17}$$

for all n large enough and all ε small enough. Also assume that there exist $0 \leq \kappa < 1/4$ and $C' > 0$ such that $G_n \leq C' n^\kappa$ and $\|G_n\|_2 \leq C'$ for all n large enough. Then there exists $K_3 > 0$ such that

$$P \sup_{g_{n,s}\in\mathcal{G}_n} |\mathbb{G}_n g_{n,s}| \leq K_3 a_n^{1/2} k_n^{3/4} \sqrt{\log k_n a_n^{-1}},$$

for all n large enough.

The increasing dimension k_n of ν affects the upper bound via two routes. First, it increases the size of envelope by the factor of $\sqrt{k_n}$, which in turn increases the entropy of the space. Second, it also demands us to consider an inflated class of functions to apply the more fundamental maximal inequality by Doukhan, Massart and Rio (1995), which relies on the $\|\cdot\|_{2,\beta}$ norm. Note that the envelope condition in (16) allows for step functions containing some nonparametric estimates.

Based on this lemma, the convergence rate of the set estimator $\hat{\Theta}$ is characterized as follows.

Theorem 4. *Suppose Assumption D holds true. Let $\{f_{n,\theta,\nu_0} : \theta \in \Theta\}$ satisfy Assumption S and $\{h_n^{1/2} f_{n,\theta,\nu_0} : \theta \in \Theta_I\}$ be a P-Donsker class. Assume $\rho(\hat{\Theta}, \Theta_I) \xrightarrow{p} 0$, $\hat{c} = o_p((nh_n)^{1/2})$, $k_n \to \infty$, and*

$|\hat{\nu} - \nu_0| = o_p(a_n)$ *for some* $\{a_n\}$ *such that* $h_n/a_n \to \infty$. *Furthermore, there exist some* $\varepsilon > 0$ *and neighborhoods* $\{\theta \in \Theta : |\theta - \pi_\theta| < \varepsilon\}$ *and* $\{\nu : |\nu - \nu_0| \leq \varepsilon\}$, *where* $h_n^{1/2}(f_{n,\theta,\nu} - f_{n,\theta,\nu_0})$ *satisfies the conditions (16) and (17) in Lemma MS' and*

$$P(f_{n,\theta,\nu} - f_{n,\theta,\nu_0}) - P(f_{n,\pi_\theta,\nu} - f_{n,\pi_\theta,\nu_0}) = o(|\theta - \pi_\theta|^2) + O(|\nu - \nu_0|^2 + r_n^{-2/3}). \tag{18}$$

Then

$$\rho(\hat{\Theta}, \Theta_I) = O_p(\hat{c}^{1/2}(nh_n)^{-1/4} + r_n^{-1/3} + (nh_n a_n^{-1})^{-1/4}(\log k_n)^{1/2}) + o(a_n). \tag{19}$$

Furthermore, if $\hat{c} \to \infty$, *then* $P\{\Theta_I \subset \hat{\Theta}\} \to 1$ *and*

$$H(\hat{\Theta}, \Theta_I) = O_p(\hat{c}^{1/2}(nh_n)^{-1/4} + (nh_n)^{-1/4} a_n^{1/4} k_n^{3/8} \log^{1/4} n) + o(a_n). \tag{20}$$

Compared to Theorem 3, we have two extra terms in the convergence rate of $H(\hat{\Theta}, \Theta_I)$ due to (nonparametric) estimation of ν_0. However, they can be shown to be dominated by the first term under standard conditions. Suppose that $k_n^4 \log k_n / n \to 0$ and the preliminary estimator $\hat{\nu}$ satisfies $|\hat{\nu} - \nu_0| = O_p(n^{-1/2}(k_n \log k_n)^{1/2})$, which is often the case as in sieve estimation (see, e.g., Chen, 2007).[6] Then we can set $a_n = n^{-1/2}(k_n \log k_n)^{1/2}$ so that $a_n^{1/4} k_n^{3/8} \to 0$. Now by choosing $\hat{c} = \log n$, the first term in (20) dominates the other terms.

4.1.1. *Example: Binary choice with interval regressor.* As an illustration of partially identified models, we consider a binary choice model with an interval-valued regressor studied by Manski and Tamer (2002). More precisely, let $y = \mathbb{I}\{x'\theta_0 + w + u \geq 0\}$, where x is a vector of observable regressors, w is an unobservable regressor, and u is an unobservable error term satisfying $P\{u \leq 0|x, w\} = \alpha$ (we set $\alpha = .5$ to simplify the notation). Instead of w, we observe the interval $[w_l, w_u]$ such that $P\{w_l \leq w \leq w_u\} = 1$. Here we normalize that the coefficient of w to determine y equals one. In this setup, the parameter θ_0 is partially identified and its identified set is written as (Manski and Tamer 2002, Proposition 2)

$$\Theta_I = \{\theta \in \Theta : P\{x'\theta + w_u \leq 0 < x'\theta_0 + w_l \text{ or } x'\theta_0 + w_u \leq 0 < x'\theta + w_l\} = 0\}.$$

Let $\tilde{x} = (x', w_l, w_u)'$ and $q_{\hat{\nu}}(\tilde{x})$ be an estimator for $P\{y = 1|\tilde{x}\}$ with the estimated parameters $\hat{\nu}$. Suppose $P\{y = 1|\tilde{x}\} = q_{\nu_0}(\tilde{x})$. By exploring the maximum score approach, Manski and Tamer (2002) developed the set estimator for Θ_I

$$\hat{\Theta} = \{\theta \in \Theta : \max_{\theta \in \Theta} S_n(\theta) - S_n(\theta) \leq \epsilon_n\}, \tag{21}$$

where

$$S_n(\theta) = \mathbb{P}_n(y - .5)[\mathbb{I}\{q_{\hat{\nu}}(\tilde{x}) > .5\}\text{sgn}(x'\theta + w_u) + \mathbb{I}\{q_{\hat{\nu}}(\tilde{x}) \leq .5\}\text{sgn}(x'\theta + w_l)].$$

Manski and Tamer (2002) established the consistency of $\hat{\Theta}$ to Θ_I under the Hausdorff distance. To establish the consistency, they assumed the cutoff value ϵ_n is bounded from below by the (almost

[6] Alternatively ν_0 can be estimated by some high-dimensional method (e.g. Belloni, Chen, Chernozhukov and Hansen, 2012), which also typically guarantees $a_n = o(n^{-1/4})$.

sure) decay rate of $\sup_{\theta\in\Theta}|S_n(\theta)-S(\theta)|$, where $S(\theta)$ is the limiting object of $S_n(\theta)$. As Manski and Tamer (2002, Footnote 3) argued, characterization of the decay rate is a complex task because $S_n(\theta)$ is a step function and $\mathbb{I}\{q_{\hat{\nu}}(\tilde{x})>.5\}$ is a step function transform of the nonparametric estimate of $P\{y=1|\tilde{x}\}$. Therefore, it has been an open question. Obtaining the lower bound rate of ϵ_n is important because we wish to minimize the volume of the estimator $\hat{\Theta}$ without losing the asymptotic validity. By applying Theorem 4, we can explicitly characterize the decay rate for the lower bound of ϵ_n and establish the convergence rate of $\hat{\Theta}$.

A little algebra shows that the set estimator in (21) is written as

$$\hat{\Theta}=\{\theta\in\Theta:\max_{\theta\in\Theta}\mathbb{P}_n f_{\theta,\hat{\nu}}-\mathbb{P}_n f_{\theta,\hat{\nu}}\leq \hat{c}n^{-1/2}\},$$

where $z=(x',w,w_l,w_u,u)'$, $h(x,w,u)=\mathbb{I}\{x'\theta_0+w+u\geq 0\}-\mathbb{I}\{x'\theta_0+w+u<0\}$, and

$$f_{\theta,\nu}(z)=h(x,w,u)[\mathbb{I}\{x'\theta+w_u\geq 0,q_\nu(\tilde{x})>.5\}-\mathbb{I}\{x'\theta+w_l<0,q_\nu(\tilde{x})\leq .5\}]. \tag{22}$$

To apply Theorem 4, we check Assumption S with $h_n=1$ for $\{f_{\theta,\nu_0}:\theta\in\Theta\}$ in the supplementary material (Section B.7). Here we illustrate the verification of (16). Let $I_\nu(\tilde{x})=\mathbb{I}\{q_\nu(\tilde{x})>.5\geq q_{\nu_0}(\tilde{x})$ or $q_\nu(\tilde{x})\leq .5<q_{\nu_0}(\tilde{x})\}$ and note that $|f_{\theta,\nu}-f_{\theta,\nu_0}|^2\leq \mathbb{I}\{x'\theta\geq -w_u\geq x'\pi_\theta$ or $x'\theta<-w_u<x'\pi_\theta\}I_\nu(\tilde{x})\leq I_\nu(\tilde{x})$. Also we have

$$P\sup_{\nu\in\Lambda:|\nu-\nu_0|<\varepsilon}\mathbb{I}\{q_\nu(\tilde{x})>.5\geq q_{\nu_0}(\tilde{x})\}\leq CP\sup_{\nu\in\Lambda:|\nu-\nu_0|<\varepsilon}|q_\nu(\tilde{x})-q_{\nu_0}(\tilde{x})|\leq C\sqrt{k_n}\varepsilon,$$

where the first inequality holds under boundedness of the conditional density of $q_{\nu_0}(\tilde{x})$ and the second under smoothness of q_ν. This verifies (16). Also (17) is verified in the same manner as Assumption S (ii).

For (18), note that

$$\begin{aligned}
&|P(f_{\theta,\nu}-f_{\theta,\nu_0})-P(f_{\pi_\theta,\nu}-f_{\pi_\theta,\nu_0})|\\
\leq\ & P\mathbb{I}\{x'\theta\geq -w_u\geq x'\pi_\theta \text{ or } x'\theta<-w_u<x'\pi_\theta\}I_\nu(\tilde{x})\\
&+P\mathbb{I}\{x'\theta\geq -w_l\geq x'\pi_\theta \text{ or } x'\theta<-w_l<x'\pi_\theta\}I_\nu(\tilde{x}),
\end{aligned}\tag{23}$$

for each $\theta\in\{\theta\in\Theta:|\theta-\pi_\theta|<\varepsilon\}$ and ν in a neighborhood of ν_0. For the first term of (23), the law of iterated expectation and an expansion of $q_\nu(\tilde{x})$ around ν_0 imply

$$\begin{aligned}
&P\mathbb{I}\{x'\theta\geq -w_u\geq x'\pi_\theta \text{ or } x'\theta<-w_u<x'\pi_\theta\}I_\nu(\tilde{x})\\
\leq\ & P\mathbb{I}\{x'\theta\geq -w_u\geq x'\pi_\theta \text{ or } x'\theta<-w_u<x'\pi_\theta\}A(w_u,x)|v-\nu_0|,
\end{aligned}$$

for some bounded function A. The second term of (23) is bounded in the same manner. Therefore, $|P(f_{\theta,\nu}-f_{\theta,\nu_0})-P(f_{\pi_\theta,\nu}-f_{\theta,\nu_0})|=O(|\theta-\pi_\theta||v-\nu_0|)$ and (18) is verified. Since all conditions of Theorem 4 are satisfied, we conclude that the convergence rate of Manski and Tamer's (2002) set estimator $\hat{\Theta}$ in (21) is characterized by (19) and (20).

Manski and Tamer (2002) proved the consistency of $\hat{\Theta}$ to Θ_I in terms of the Hausdorff distance. We provide a sharper lower bound on the their tuning parameter ϵ_n, which is $\hat{c}n^{-1/2}$ with $\hat{c} \to \infty$. For example, if we set $\hat{c} = \log n$, the convergence rate becomes $H(\hat{\Theta}, \Theta_I) = O_p(n^{-1/4}(\log n)^{1/2})$. We basically verify the high level assumption of Chernozhukov, Hong and Tamer (2007, Condition C.2) in the cube root context. However, we mention that in the above setup, the criterion function contains nuisance parameters with increasing dimension and the result in Chernozhukov, Hong and Tamer (2007) does not apply.

Furthermore, our result enables us to construct the confidence set by subsampling as described by Chernozhukov, Hong and Tamer (2007). Specifically, the maximal inequality in Lemma MS' and the assumption that $\{h_n^{1/2} f_{n,\theta,\nu_0} : \theta \in \Theta_I\}$ is P-Donsker are sufficient to satisfy their Conditions C.4 and C.5.

4.2. **Parameter-dependent local M-estimation.** We consider a setup where localization of the criterion function depends on the parameter values. A leading example is the mode estimation. Chernoff (1964) studied the asymptotic property of the mode estimator that maximizes $(nh)^{-1} \sum_{t=1}^n \mathbb{I}\{|y_t - \beta| \leq h\}$ with respect to β for some fixed h. This estimator was extended to the regression case by Lee (1989). Lee (1989) established consistency of the mode regression estimator and conjectured the cube root convergence rate. To estimate β consistently for a broader family of distributions, however, we need to treat h as a bandwidth parameter and let $h \to 0$ as in Yao, Lindsay and Li (2012) for example.

This parameter-dependent localization alters Assumption M (iii) because it increases the size (in terms of the L_2-norm) of the envelope of the class $\{h^{-1}(\mathbb{I}\{|y_t - \beta| \leq h\} - \mathbb{I}\{|y_t - \beta_0| \leq h\}) : |\beta - \beta_0| \leq \varepsilon\}$. More precisely, we replace Assumption M (iii) with the following.

Assumption M (iii'). *There exists a positive constant C'' such that*

$$P \sup_{\theta \in \Theta : |\theta - \theta'| < \varepsilon} h_n^2 |f_{n,\theta} - f_{n,\theta'}|^2 \leq C'' \varepsilon,$$

for all n large enough, $\varepsilon > 0$ small enough, and θ' in a neighborhood of θ_0.

Under this assumption, Lemma M in Section 2 is modified as follows.

Lemma M1. *Under Assumption M (i), (ii), and (iii'), there exist positive constants C and C' such that*

$$P \sup_{\theta \in \Theta : |\theta - \theta_0| < \delta} |\mathbb{G}_n h_n^{1/2} (f_{n,\theta} - f_{n,\theta_0})| \leq C h_n^{-1/2} \delta^{1/2},$$

for all n large enough and $\delta \in [(nh_n^2)^{-1/2}, C']$.

Parameter dependency arises in different contexts and may lead to different types of non-standard distributions. For instance, the maximum likelihood estimator for Uniform$[0, \theta]$ yields super consistency (see Hirano and Porter, 2003, for a general discussion). This contrast is similar to the difference between estimation of a change point in regression analysis and mode regression.

Once we have obtained Lemma 4.2, the remaining steps are similar to those in Section 2 by replacing "h_n" with "h_n^2". Here we present the result without nuisance parameters ν for the sake of expositional simplicity.

Theorem 5. *Let $\{f_{n,\theta} : \theta \in \Theta\}$ satisfy Assumption M (i), (ii), and (iii'). Also suppose (5) holds with $(g_{n,s} - Pg_{n,s})$ for each s, where $g_{n,s} = n^{1/6}h_n^{4/3}(f_{n,\theta_0+s(nh_n^2)^{-1/3}} - f_{n,\theta_0})$. Then*

$$(nh_n^2)^{1/3}(\hat{\theta} - \theta_0) \xrightarrow{d} \arg\max_{s\in\mathbb{R}^d} Z(s), \tag{24}$$

where $Z(s)$ is a Gaussian process with continuous sample paths, expected value $s'Vs/2$ and covariance kernel $H(s_1, s_2) = \lim_{n\to\infty} \sum_{t=-n}^{n} \mathrm{Cov}(g_{n,s_1}(z_0), g_{n,s_2}(z_t)) < \infty$.

4.2.1. *Example: Hough transform estimator.* In the statistics literature on computer vision algorithm, Goldenshluger and Zeevi (2004) investigated the so-called Hough transform estimator for the regression model

$$\hat{\beta} = \arg\max_{\beta\in B} \sum_{t=1}^{n} \mathbb{I}\{|y_t - x_t'\beta| \le h|x_t|\}, \tag{25}$$

where B is some parameter space, $x_t = (1, \tilde{x}_t)'$ for a scalar $\tilde{x}_t$, and h is a fixed tuning constant. Goldenshluger and Zeevi (2004) derived the cube root asymptotics for $\hat{\beta}$ with fixed h and discussed carefully about the practical choice of h. However, for this estimator, h plays the role of the bandwidth and the analysis for the case of $h_n \to 0$ is a substantial open question (see pp. 1915-6 of Goldenshluger and Zeevi, 2004). Here we study the asymptotic property of $\hat{\beta}$ in (25) with $h = h_n \to 0$. The estimators by Chernoff (1964) and Lee (1989) with varying h can be analyzed in the same manner.

Let $z = (x, u)$. Note that $\hat{\theta} = \hat{\beta} - \beta_0$ is written as the M-estimator using the criterion function

$$f_{n,\theta}(z) = h_n^{-1}\mathbb{I}\{|u - x'\theta| \le h_n|x|\}.$$

The consistency of $\hat{\theta}$ follows from the uniform convergence $\sup_{\theta\in\Theta} |\mathbb{P}_n f_{n,\theta} - Pf_{n,\theta}| \xrightarrow{p} 0$ by applying Nobel and Dembo (1993, Theorem 1).

To apply Theorem 5, we need to verify that $\{f_{n,\theta}\}$ satisfies Assumption M (i), (ii), and (iii'). Here we show it for (iii') explicitly while other details are found in the supplementary material (Section B.8). Observe that

$$\begin{aligned} P\sup_{\theta\in\Theta:|\theta-\vartheta|<\varepsilon} h_n^2|f_{n,\theta} - f_{n,\vartheta}|^2 \;\le\; & P\sup_{\theta\in\Theta:|\theta-\vartheta|<\varepsilon} \mathbb{I}\{|u - x'\vartheta| \le h_n|x|,\ |u - x'\theta| > h_n|x|\} \\ & + P\sup_{\theta\in\Theta:|\theta-\vartheta|<\varepsilon} \mathbb{I}\{|u - x'\theta| \le h_n|x|,\ |u - x'\vartheta| > h_n|x|\}, \end{aligned}$$

for all ϑ in a neighborhood of 0. Since the same argument applies to the second term, we focus on the first term (say, T). If $\varepsilon \le 2h_n$, then an expansion around $\varepsilon = 0$ implies

$$T \le P\{(h_n - \varepsilon)|x| \le u \le h_n|x|\} = P\gamma(h_n|x|)|x|\varepsilon + o(\varepsilon),$$

assuming independence between u and x. Also, if $\varepsilon > 2h_n$, then an expansion around $h_n = 0$ implies

$$T \leq P\{-h_n|x| \leq u \leq h_n|x|\} \leq P\gamma(0)|x|\varepsilon + o(h_n).$$

Therefore, Assumption M (iii') is satisfied.

Finally, the covariance kernel is obtained by a similar way as Section 3.1. Let $r_n = (nh_n^2)^{1/3}$ be the convergence rate in this example. The covariance kernel is written by $H(s_1, s_2) = \frac{1}{2}\{L(s_1, 0) + L(0, s_2) - L(s_1, s_2)\}$, where $L(s_1, s_2) = \lim_{n\to\infty} \mathrm{Var}(r_n^2 \mathbb{P}_n g_{n,t})$ with $g_{n,t} = f_{n,s_1/r_n} - f_{n,s_2/r_n}$. An expansion implies $n^{-1}\mathrm{Var}(r_n^2 g_{n,t}) \to 2\gamma(0)P|x'(s_1 - s_2)|$. We can also see that the covariance term $n^{-1}\sum_{m=1}^{\infty} \mathrm{Cov}(r_n^2 g_{n,t}, r_n^2 g_{n,t+m})$ is negligible. Therefore, Theorem 5 implies the limiting distribution of the Hough transform estimator with the bandwidth h_n is obtained as in (24) with $V = \ddot{\gamma}(0)P(|x|xx')$ and $H(s_1, s_2) = 2\gamma(0)P|x'(s_1 - s_2)|$.

5. Conclusion

This paper develops general asymptotic theory, which encompasses a wide class of non-regular M-estimation problems. Many of these problems have been left without a proper inference method for a long time. It is worthwhile to emphasize that our theory validates inference based on subsampling for this important class of estimators, including the confidence set construction for set-valued parameters in Manski and Tamer's (2002) binary choice model with an interval regressor. An interesting future research is to develop valid bootstrap methods for these estimators. Naive applications of the standard bootstrap resampling lead to inconsistent inference as shown by Abrevaya and Huang (2005) and Sen, Banerjee and Woodroofe (2010) among others.

Supplementary material. This paper has an on-line supplementary material, which contains all the proofs of the theorems and lemmas and additional examples.

References

[1] Abrevaya, J. and J. Huang (2005) On the bootstrap of the maximum score estimator, *Econometrica*, 73, 1175-1204.

[2] Adamczak, R. (2008) A tail inequality for suprema of unbounded empirical processes with applications to Markov chains, *Electronic Journal of Probability*, 13, 1000-1034.

[3] Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H. and J. W. Tukey (1972) *Robust Estimates of Location*, Princeton University Press, Princeton.

[4] Andrews, D. W. K. (1993) An introduction to econometric applications of empirical process theory for dependent random variables, *Econometric Reviews*, 12, 183-216.

[5] Anevski, D. and O. Hössjer (2006) A general asymptotic scheme for inference under order restrictions, *Annals of Statistics*, 34, 1874-1930.

[6] Arcones, M. A. and B. Yu (1994) Central limit theorems for empirical and U-processes of stationary mixing sequences, *Journal of Theoretical Probability*, 7, 47-71.

[7] Banerjee, M. and I. W. McKeague (2007) Confidence sets for split points in decision trees, *Annals of Statistics*, 35, 543-574.

[8] Baraud, Y. (2010) A Bernstein-type inequality for suprema of random processes with applications to model selection in non-Gaussian regression, *Bernoulli*, 16, 1064-1085.

[9] Belloni, A., Chen, D., Chernozhukov, V. and C. Hansen (2012) Sparse models and methods for optimal instruments with application to eminent domain, *Econometrica*, 80, 2369-2429.

[10] Bühlmann, P. and B. Yu (2002) Analyzing bagging, *Annals of Statistics*, 30, 927-961.

[11] Carrasco, M. and X. Chen (2002) Mixing and moment properties of various GARCH and stochastic volatility models, *Econometric Theory*, 18, 17-39.

[12] Chan, K. S. (1993) Consistency and Limiting Distribution of the Least Squares Estimator of a Threshold Autoregressive Model. *Ann. Statist.* 21, 520-533.

[13] Chen, X. (2007) Large sample sieve estimation of semi-nonparametric models, *Handbook of Econometrics*, vol. 6B, ch. 76, Elsevier.

[14] Chen, X., Hansen, L. P. and M. Carrasco (2010) Nonlinearity and temporal dependence, *Journal of Econometrics*, 155, 155-169.

[15] Chernoff, H. (1964) Estimation of the mode, *Annals of the Institute of Statistical Mathematics*, 16, 31-41.

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

[17] de Jong, R. M. and T. Woutersen (2011) Dynamic time series binary choice, *Econometric Theory*, 27, 673-702.

[18] Doukhan, P., Massart, P. and E. Rio (1994) The functional central limit theorems for strongly mixing processes, *Annales de l'Institut Henri Poincaré, Probability and Statistics*, 30, 63-82.

[19] Doukhan, P., Massart, P. and E. Rio (1995) Invariance principles for absolutely regular empirical processes, *Annales de l'Institut Henri Poincaré, Probability and Statistics*, 31, 393-427.

[20] Gautier, E. and Y. Kitamura (2013) Nonparametric estimation in random coefficients binary choice models, *Econometrica*, 81, 581-607.

[21] Goldenshluger, A. and A. Zeevi (2004) The Hough transform estimator, *Annals of Statistics*, 32, 1908-1932.

[22] Hirano, K. and J. R. Porter, (2003) Asymptotic efficiency in parametric structural models with parameter-dependent support, *Econometrica*, 71, 1307-1338.

[23] Honoré, B. E. and E. Kyriazidou (2000) Panel data discrete choice models with lagged dependent variables, *Econometrica*, 68, 839-874.

[24] Kim, J. and D. Pollard (1990) Cube root asymptotics, *Annals of Statistics*, 18, 191-219.

[25] Koo, B. and M. H. Seo (2015) Structural Break Models under Misspecification: Implication for Forecasting, *Journal of Econometrics*, 188, 166-181.

[26] Kosorok, M. R. (2008) *Introduction to Empirical Processes and Semiparametric Inference*, Springer.

[27] Lee, M.-J. (1989) Mode regression, *Journal of Econometrics*, 42, 337-349.

[28] Manski, C. F. (1975) Maximum score estimation of the stochastic utility model of choice, *Journal of Econometrics*, 3, 205-228.

[29] Manski, C. F. (1985) Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator, *Journal of Econometrics*, 27, 313-333.

[30] Manski, C. F. and E. Tamer (2002) Inference on regressions with interval data on a regressor or outcome, *Econometrica*, 70, 519-546.

[31] Merlevède, F., Peligrad, M. and E. Rio (2009) Bernstein inequality and moderate deviations under strong mixing conditions, *IMS Collections: High Dimensional Probability V*, 5, 273-292.

[32] Merlevède, F., Peligrad, M. and E. Rio (2011) A Bernstein type inequality and moderate deviations for weakly dependent sequences, *Probability Theory and Related Fields*, 151, 435-474.

[33] Moon, H. R. (2004) Maximum score estimation of a nonstationary binary choice model, *Journal of Econometrics*, 122, 385-403.

[34] Nickl, R. and J. Söhl (2016) Nonparametric Bayesian posterior contraction rates for discretely observed scalar diffusions, Working paper, arXiv:1510.05526v2.

[35] Nobel, A. and A. Dembo (1993) A note on uniform laws of averages for dependent processes, *Statistics and Probability Letters*, 17, 169-172.

[36] Paulin, D. (2015) Concentration inequalities for Markov chains by Marton couplings and spectral methods, *Electronic Journal of Probability*, 20, 1-32.

[37] Prakasa Rao, B. L. S. (1969) Estimation of a unimodal density, *Sankhyā*, A, 31, 23-36.

[38] Politis, D. N., Romano, J. P. and M. Wolf (1999) *Subsampling*, New York: Springer-Verlag.

[39] Pollard, D. (1989) Asymptotics via empirical processes, Statistica Sinica, 4, 341-354.

[40] Pollard, D. (1993) The asymptotics of a binary choice model, Working paper.

[41] Polonik, W. and Q. Yao (2000) Conditional minimum volume predictive regions for stochastic processes, *Journal of the American Statistical Association*, 95, 509-519.

[42] Rio, E. (1997) About the Lindeberg method for strongly mixing sequences, *ESAIM: Probability and Statistics*, 1, 35-61.

[43] Romano, J. P. and A. M. Shaikh (2008) Inference for identifiable parameters in partially identified econometric models, *Journal of Statistical Planning and Inference*, 138, 2786-2807.

[44] Rousseeuw, P. J. (1984) Least median of squares regression, *Journal of the American Statistical Association*, 79, 871-880.

[45] Sen, B., Banerjee, M. and M. Woodroofe (2010) Inconsistency of bootstrap: the Grenander estimator, *Annals of Statistics*, 38, 1953-1977.

[46] Talagrand, M. (2005) *The Generic Chaining*, Springer.

[47] van der Vaart, A. W. and J. A. Wellner (1996) *Weak Convergence and Empirical Processes*, Springer, New York.

[48] van der Vaart, A. W. and J. A. Wellner (2007) Empirical processes indexed by estimated functions, *IMS Lecture Notes: Asymptotics: Particles, Processes and Inverse Problems,* 55, 234-252.

[49] Yao, W., Lindsay, B. G. and R. Li (2012) Local modal regression, *Journal of Nonparametric Statistics*, 24, 647-663.

[50] Zinde-Walsh, V. (2002) Asymptotic theory for some high breakdown point estimators, *Econometric Theory*, 18, 1172-1196.

Department of Economics, Seoul National University, Seoul, Korea.

E-mail address: myunghseo@snu.ac.kr

Department of Economics, London School of Economics, Houghton Street, London, WC2A 2AE, UK.

E-mail address: t.otsu@lse.ac.uk