Semiparametric M-Estimation with Non-Smooth Criterion Functions

Laurent Delsol, Ingrid van Keilegom

Semiparametric M-Estimation with Non-Smooth Criterion Functions

Laurent Delsol [*†] Ingrid Van Keilegom [‡]

February 23, 2015

Abstract

We are interested in the estimation of a parameter θ that maximizes a certain criterion function depending on an unknown, possibly infinite dimensional nuisance parameter h. A common estimation procedure consists in maximizing the corresponding empirical criterion, in which the nuisance parameter is replaced by a non-parametric estimator. In the literature, this research topic, commonly referred to as semiparametric M-estimation, has received a lot of attention in the case where the criterion function satisfies certain smoothness properties. In certain applications, these smoothness conditions are however not satisfied. The aim of this paper is therefore to extend the existing theory on semiparametric M-estimation problems, in order to cover non-smooth M-estimation problems as well. In particular, we develop 'high-level' conditions under which the proposed M-estimator is consistent and has an asymptotic limit. We also check these conditions in detail for a specific example of a semiparametric M-estimation problem, which comes from the area of classification with missing data, and which cannot be dealt with using the existing results in the literature. Finally, we perform a small simulation study to verify the small sample performance of the proposed estimator, and we briefly describe a number of other situations in which the general theory can be applied, and which are not covered by the existing theory for semiparametric M-estimators.

Key Words: Asymptotic distribution; Classification; Empirical processes; M-estimation; Missing data; Nonstandard asymptotics; Nuisance parameter; Semiparametric regression.

[*]MAPMO, Université d'Orléans, Bâtiment de mathématiques, Rue de Chartres B.P. 6759, 45067 Orléans cedex 2, France, E-mail address: laurent.delsol@univ-orleans.fr

[†]Most of the research leading to this paper was carried out while L. Delsol was a postdoctoral fellow at the Université catholique de Louvain, financed by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement No. 203650.

[‡]Institute of Statistics, Université catholique de Louvain, Voie du Roman Pays 20, B 1348 Louvain-la-Neuve, Belgium. E-mail address: ingrid.vankeilegom@uclouvain.be. Research supported by IAP research network grant nr. P7/06 of the Belgian government (Belgian Science Policy), by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement No. 203650, and by the contract 'Projet d'Actions de Recherche Concertées' (ARC) 11/16-039 of the 'Communauté française de Belgique', granted by the 'Académie universitaire Louvain'.

1 Introduction

Consider the estimation of a parameter of interest θ_0 that maximizes a criterion function $M(\theta, h_0)$, where h_0 is the true value of an unknown, possibly infinite dimensional nuisance parameter h. A common estimation procedure consists in maximizing an empirical criterion function $M_n(\theta, \widehat{h})$, where M_n is an estimator of the unknown function M and $\widehat{h}$ is a nonparametric estimator of the unknown nuisance parameter h_0. In the literature, this research topic, commonly referred to as semiparametric M-estimation, has received a lot of attention in the case where the criterion function M satisfies certain smoothness properties. In certain applications, these smoothness conditions are however not satisfied. The aim of this paper is therefore to extend the existing theory on semiparametric M-estimation problems, in order to cover non-smooth M-estimation problems as well. In particular, we develop 'high-level' conditions under which the proposed M-estimator is consistent and has an asymptotic limit. We check these conditions in detail for a specific example of a semiparametric M-estimation problem, which comes from the area of classification with missing data, and which cannot be dealt with using the existing results in the literature. We also mention briefly a number of other examples that are not covered by the current literature on semiparametric estimation. These examples come from various areas in econometrics and statistics, and illustrate the usefulness of our results.

Non-smooth semiparametric M-estimation problems form a basically unsolved open problem in the literature. We aim at filling this gap in the literature by combining results for non-smooth parametric M-estimation problems with smooth semiparametric M-estimation problems. However, as will be seen later, the problem requires much more than 'simply' combining ideas from these two domains. In fact, delicate mathematical derivations will be required to cope with estimators of the nuisance parameters inside non-smooth criterion functions. This feature is not present for parametric M-estimators nor for smooth semiparametric M-estimators, and is the source of the complicated nature of this problem.

In the literature it is often assumed that $M(\theta, h)$ can be written as

$$M(\theta, h) = \mathbb{E}\big[m(Z, \theta, h(Z, \theta))\big], \tag{1}$$

where m is a known function and h is allowed to depend on θ and on a random vector Z taking values in some space $\mathcal{F}$. A common estimation procedure consists then in

maximizing the corresponding empirical criterion:

$$M_n(\theta, \widehat{h}) = \frac{1}{n} \sum_{i=1}^{n} m(Z_i, \theta, \widehat{h}(Z_i, \theta)), \tag{2}$$

with respect to θ, where the random vectors $Z_1, \ldots, Z_n$ have the same distribution as Z. We assume in the remainder of this paper that for all θ the functions $m(\cdot, \theta, \widehat{h}(\cdot, \theta))$ and $m(\cdot, \theta, h_0(\cdot, \theta))$ are measurable.

When the function $m(z, \theta, h(z, \theta))$ is differentiable with respect to θ and when $M(\theta, h_0)$ is concave in θ, then the M-estimation problem can be reduced to a Z-estimation problem, by solving the equation $\partial M_n(\theta, \widehat{h})/\partial\theta = 0$ (or by minimizing the norm of $\partial M_n(\theta, \widehat{h})/\partial\theta$ if a solution would not exist). A general result on (two-step) semiparametric Z-estimators can be found in Chen, Linton and Van Keilegom (2003). In that paper high-level conditions are given under which the estimator of θ_0 is weakly consistent and asymptotically normal. The criterion function $\partial m/\partial\theta$ is not required to be smooth in θ nor in h. See also Van der Vaart and Wellner (2007) for high-level conditions for the stochastic equicontinuity in semiparametric Z-estimation problems. For specific examples of semiparametric Z-estimation problems we refer (among others) to Chen and Fan (2006), Linton, Sperlich and Van Keilegom (2008), Mammen, Rothe and Schienle (2011), Escanciano, Jacho-Chavez and Lewbel (2012, 2014), and the references therein. Finally, for (one-step) sieve estimation in semiparametric Z-estimation problems, see Chen (2007), Chen and Pouzo (2009), Ding and Nan (2011), Chen and Liao (2012) and Cheng and Shang (2015), among others. See also the book by Horowitz (2009) for a number of other examples.

On the other hand, when either $m(z, \theta, h(z, \theta))$ is not differentiable with respect to θ, or when $M(\theta, h_0)$ has more than one (local) maximum, then the M-estimation problem can not be reduced to a Z-estimation problem, and we need to use other procedures. In the parametric case, where no infinite dimensional nuisance parameter is present, we refer to Kim and Pollard (1990) for a general result on parametric M-estimators that have $n^{1/3}$-rate of convergence, and to Van der Vaart and Wellner (1996) for a result on both estimators that converge at $n^{1/2}$-rate in the smooth case, and at a rate slower than $n^{1/2}$ for non-smooth functions. See also Groeneboom and Wellner (1993), Groeneboom, Jongbloed and Wellner (2001), Goldenshluger and Zeevi (2004), Mohammadi and Van de Geer (2005) and Radchenko (2008), among others for important contributions on results for specific parametric M-estimation problems with slower than $n^{1/2}$-rate of convergence.

The problem becomes more difficult when the model is semiparametric. Basically two main approaches can be considered in that case. In the first approach $M_n(\theta, h)$ is

maximized jointly with respect to θ and h, and then the criterion function is modified in order to obtain an estimator of θ_0 converging at $n^{1/2}$-rate. The second approach, which we will follow, consists in maximizing $M_n(\theta, \widehat{h})$ with respect to θ, where $\widehat{h}$ is a preliminary estimator of h_0. When the m-function is in some sense 'smooth' (e.g. Lipschitz continuous in L_p-norm) several contributions on both approaches can be found in the literature. See e.g. Van der Vaart and Wellner (1996), Van de Geer (2000), Ma and Kosorok (2005), Kosorok (2008), Ichimura and Lee (2010), and Kristensen and Salanié (2013). In these cases, the estimator of θ_0 is $n^{1/2}$-consistent, even when the nuisance parameter is estimated at slower rate. This rate is obtained thanks to the regularity of the criterion function m.

However, in numerous situations we are faced to semiparametric M-estimation problems, where the function m does not satisfy the smoothness property that makes the estimator of θ_0 $n^{1/2}$-consistent. Examples can be found (among many others) in classification problems with variables missing at random, and in partially linear binary choice models (see Section 6 for more details and examples). This general context has, to the best of our knowledge, not been considered so far in the literature. It is substantially more difficult than the 'smooth' case. This can be understood e.g. from the fact that it leads to non-standard asymptotics and to estimators of θ_0 that are not $n^{1/2}$-consistent and that converge to non-normal limits. To achieve this we need to apply second order Taylor expansions (as opposed to first order Taylor expansions in the smooth case), the application of delicate empirical process results, and the analysis of messy and complicated remainder terms, which do not show up in the smooth case or the case without nonparametric nuisance functions. The results that we will obtain allow to show that in the case where m is not smooth (e.g. when m includes an indicator function), we can obtain the same rate of convergence (and sometimes even the same asymptotic distribution) as in the case where the nuisance parameter would be known, even when the nuisance parameter is estimated at slower rate.

The paper is organized as follows. In the next section we introduce some notations and give the formal definition of the M-estimator. In Section 3 we show under which conditions the estimator of θ_0 is weakly consistent. Section 4 deals with the development of the rate of convergence of the estimator, whereas in Section 5 we state the asymptotic distribution of the estimator. In Section 6 a particular example of a non-smooth semiparametric M-estimation problem is considered, for which we check the conditions of the asymptotic results in detail, and a number of other examples are briefly outlined. Finally, the Appendix contains the proofs of the asymptotic results.

2 Notations and definitions

Throughout the paper we assume that the data $Z_1, \ldots, Z_n$ are identically distributed random vectors. In many applications the vector Z_i $(i = 1, \ldots, n)$ will consist of a random vector Y_i (representing a response) and a random vector X_i (representing a vector of explanatory variables). The set Θ denotes a compact parameter set (usually but not necessarily of finite dimension) with non empty interior and $\mathcal{H}$ denotes an infinite dimensional parameter set. Suppose there exists a non-random measurable real-valued function $M : \Theta \times \mathcal{H} \to I\!R$, such that

$$\theta_0 = \text{argmax}_{\theta \in \Theta} M(\theta, h_0(\cdot, \theta)),$$

and suppose θ_0 is unique and belongs to the interior of Θ. Let θ_0 and $h_0 \in \mathcal{H}$ be the true unknown finite and infinite dimensional parameters. We allow that the functions $h \in \mathcal{H}$ depend on the parameters θ and the vector Z, but for notational convenience we will often suppress this dependence when no confusion is possible. For instance, we often use the following abbreviated notations : $(\theta, h) \equiv (\theta, h(\cdot, \theta))$, $(\theta, h_0) \equiv (\theta, h_0(\cdot, \theta))$, and $(\theta_0, h_0) \equiv (\theta_0, h_0(\cdot, \theta_0))$. The sets Θ and $\mathcal{H}$ are supposed to be metric spaces. Their metrics are denoted by d and $d_{\mathcal{H}}$ respectively. Since the nuisance parameter is allowed to depend on θ we implicitly define $d_{\mathcal{H}}(h, h_0)$ uniformly over θ, i.e. $d_{\mathcal{H}}(h, h_0) := \sup_{\theta \in \Theta} d^1_{\mathcal{H}}(h(., \theta), h_0(., \theta))$ for some metric $d^1_{\mathcal{H}}$.

Suppose there exists a random real-valued function $M_n : \Theta \times \mathcal{H} \to I\!R$ depending on the data $Z_1, \ldots, Z_n$, such that $M_n(\theta, h_0)$ is an approximation of $M(\theta, h_0)$ (the precise conditions on M_n will be given in the next sections). In many applications we have that $M(\theta, h) = E[m(Z, \theta, h)]$ and $M_n(\theta, h) = n^{-1} \sum_{i=1}^n m(Z_i, \theta, h)$, where m is a measurable real-valued function such that $\theta_0 = \text{argmax}_{\theta \in \Theta} E[m(Z, \theta, h_0)]$. However, the conditions on M_n do not impose this particular structure and allow for more general situations as well. Suppose that for each θ there is an initial nonparametric estimator $\widehat{h}(\cdot, \theta)$ for $h_0(\cdot, \theta)$. This nonparametric estimator depends on the particular model, and can be based on e.g. kernels, splines or neural networks. Again for notational ease we let $(\theta, \widehat{h}) \equiv (\theta, \widehat{h}(\cdot, \theta))$. We estimate θ_0 by any $\widehat{\theta} \in \Theta$ that 'approximately solves' the sample maximization problem:

$$\max_{\theta \in \Theta} M_n(\theta, \widehat{h}). \tag{3}$$

In the set of conditions given in the next sections we will formalize what we mean with 'approximate solution'.

3 Consistency

We focus in this section on the development of sufficient conditions under which the estimator $\widehat{\theta}$ is weakly consistent. This consistency will be used as a preliminary step for the subsequent sections, where we will deal with the rate of convergence and the asymptotic distribution of the estimator. In the remainder of the paper the notations P^* and E^* will be used to denote outer probabilities and outer expectations, to take into account potential measurability issues.

Consider the following assumptions:

(A1) $\widehat{\theta} \in \Theta$ and $M_n(\widehat{\theta}, \widehat{h}) \geq M_n(\theta_0, \widehat{h}) + o_{P^*}(1)$.

(A2) For all $\epsilon > 0$ there exists a $\delta(\epsilon) > 0$ such that $d(\theta, \theta_0) > \epsilon$ implies $M(\theta_0, h_0) - M(\theta, h_0) > \delta(\epsilon)$.

(A3) $\mathbb{P}(\widehat{h} \in \mathcal{H}) \to 1$ as $n \to \infty$ and $d_{\mathcal{H}}(\widehat{h}, h_0) \xrightarrow{P^*} 0$.

(A4)
$$\sup_{\theta \in \Theta,\, h \in \mathcal{H}} \frac{|M_n(\theta, h) - M_n(\theta_0, h) - M(\theta, h) + M(\theta_0, h)|}{1 + |M_n(\theta, h) - M_n(\theta_0, h)| + |M(\theta, h) - M(\theta_0, h)|} = o_{P^*}(1).$$

(A5) $\lim_{d_{\mathcal{H}}(h, h_0) \to 0} \sup_{\theta \in \Theta} |M(\theta, h) - M(\theta, h_0)| = 0$.

Below we illustrate and interpret these conditions, and we give some remarks, extensions, sufficient conditions, etc., that are useful for verifying these conditions in practice.

Remark 1

(i) In assumption (A4), we take the supremum with respect to h over the whole family $\mathcal{H}$. However, it is enough to assume the same type of convergence only for $h = \widehat{h}$ or for $\{h : d_{\mathcal{H}}(h, h_0)\} \leq \delta_n$ for some $\delta_n \to 0$ such that $d_{\mathcal{H}}(\widehat{h}, h_0)\delta_n^{-1} = o_{P^*}(1)$. Assumption (A5) could be changed in the same way.

(ii) Assumption (A4) is closely related to the compactness of the sets Θ and $\mathcal{H}$ and is automatically fulfilled when the following standard assumption holds:

$$\sup_{\theta \in \Theta,\, h \in \mathcal{H}} |M_n(\theta, h) - M(\theta, h)| = o_{P^*}(1).$$

This last condition holds when the family $\mathcal{F} = \{m(., \theta, h),\, \theta \in \Theta, h \in \mathcal{H}\}$ is Glivenko-Cantelli, and $M(\theta, h) = \mathbb{E}[m(Z, \theta, h)]$.

(iii) We do not require here that M is the mean of the random function $m(Z, ., .)$. We only require that assumption (A4) holds. Moreover, we do not impose any smoothness assumptions on M_n. We only require that the function M satisfies assumptions (A2) and (A5).

(iv) In this section we do not assume that θ belongs to an Euclidean space. It is possible that θ and h belong to general metric spaces. Theoretically, it is also possible to consider semimetric spaces, however, this only allows to get the consistency (without rate of convergence) with respect to the corresponding semimetrics.

(v) The assumption that the variables Z_i are independent is not necessary here. Our result could be used even for dependent data as soon as it is possible to fulfill assumptions (A1), (A3) and (A4).

(vi) Assumption (A1) is trivially fulfilled when $M_n(\widehat{\theta}, \widehat{h}) \geq \sup_{\theta \in \Theta} M_n(\theta, \widehat{h}) + o_{P^*}(1)$, which allows to deal with approximations of the value that actually maximizes $\theta \mapsto M_n(\theta, \widehat{h})$.

Theorem 1 *Under assumptions (A1)-(A5) we have that*

$$d(\widehat{\theta}, \theta_0) \xrightarrow{P^*} 0.$$

The proof is given in the Appendix.

4 Rate of convergence

In the previous section we have shown the consistency of general M-estimators. We are now interested in going one step further and give their convergence rates. In this section, the consistency of our estimators is used as a preliminary assumption. Of course, Theorem 1 can be used to obtain this consistency. We introduce the following assumptions:

(B1) $d(\widehat{\theta}, \theta_0) \xrightarrow{P^*} 0$ and $v_n d_{\mathcal{H}}(\widehat{h}, h_0) = O_{P^*}(1)$ for some sequence $v_n \to \infty$.

(B2) For all $\delta_1 > 0$, there exist $\alpha < 2$, $K > 0$, $\delta_0 > 0$ and $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$ there exists a function Φ_n for which $\delta \mapsto \Phi_n(\delta)/\delta^\alpha$ is decreasing on $(0, \delta_0]$ and for all $\delta \leq \delta_0$,

$$\mathbb{E}^* \left[\sup_{d(\theta,\theta_0) \leq \delta, d_{\mathcal{H}}(h,h_0) \leq \frac{\delta_1}{v_n}} |M_n(\theta, h) - M_n(\theta_0, h) - M(\theta, h) + M(\theta_0, h)| \right] \leq K \frac{\Phi_n(\delta)}{\sqrt{n}}.$$

(B3) There exist a constant $C > 0$, a sequence $r_n \to \infty$, and variables $W_n = O_{P^*}(r_n^{-1})$ and $\beta_n = o_{P^*}(1)$, such that for all $\theta \in \Theta$ satisfying $d(\theta, \theta_0) \leq \delta_0$:

$$M(\theta, \widehat{h}) - M(\theta_0, \widehat{h}) \leq W_n d(\theta, \theta_0) - C d(\theta, \theta_0)^2 + \beta_n d(\theta, \theta_0)^2.$$

(B4) $M_n(\widehat{\theta}, \widehat{h}) \geq M_n(\theta_0, \widehat{h}) + O_{P^*}(r_n^{-2})$ and $r_n^2 \Phi_n(r_n^{-1}) \leq \sqrt{n}$.

Under the above assumptions, we will prove that the estimator $\widehat{\theta}$ is r_n^{-1}-consistent. Hence, the sequence r_n plays an important role in the above assumptions and should be chosen in the sharpest possible way. Before stating and proving this result, we first discuss the above assumptions in more detail.

Remark 2

(i) Assumption (B1) is a 'high-level' assumption. Many asymptotic results allow to get such conditions on both the M-estimator $\widehat{\theta}$ and the nuisance estimator $\widehat{h}$. In general the rate of convergence of the nuisance estimator is slower than the best convergence rate of the M-estimator. We are interested in studying cases where the convergence rate of the M-estimator is not affected by the fact that we need to estimate the nuisance parameter.

(ii) Assumption (B2) is also a 'high-level' assumption. Assume that for any z the function $(\theta, h) \to m(z, \theta, h(z, \theta)) - m(z, \theta_0, h(z, \theta_0))$ is uniformly bounded on an open neighborhood of (θ_0, h_0), i.e. on $\{(\theta, h) : d(\theta, \theta_0) \leq \delta_0,\, d_{\mathcal{H}}(h, h_0) \leq \delta_1'\}$ for some $\delta_0, \delta_1' > 0$. Let us consider the class $\mathcal{F}_{\delta,\delta_1'} = \{m(., \theta, h(\cdot, \theta)) - m(., \theta_0, h(\cdot, \theta_0)) : d(\theta, \theta_0) \leq \delta,\, d_{\mathcal{H}}(h, h_0) \leq \delta_1'\}$ for any $\delta \leq \delta_0$ and denote its envelope by $M_{\delta,\delta_1'}$. For any δ_1, we have $\delta_1 v_n^{-1} \leq \delta_1'$ for n large enough. Then, under entropy conditions on $\mathcal{F}_{\delta,\delta_1'}$, as for instance

$$\sup_{\delta \leq \delta_0} \int_0^1 \sqrt{1 + \log N_{[]}(\epsilon \|M_{\delta,\delta_1'}\|_{\mathbb{L}_2(\mathbb{P}^*)},\, \mathcal{F}_{\delta,\delta_1'},\, \mathbb{L}_2(\mathbb{P}))} d\epsilon < +\infty \tag{4}$$

(where $N_{[]}$ denotes the bracketing number, i.e. the smallest number of brackets that are needed to cover the space), there exists a positive constant K_1 (not depending on δ) such that for all $\delta \leq \delta_0$,

$$\mathbb{E}^* \left[\sup_{d(\theta,\theta_0) \leq \delta, d_{\mathcal{H}}(h,h_0) \leq \delta_1'} |M_n(\theta, h) - M_n(\theta_0, h) - M(\theta, h) + M(\theta_0, h)| \right] \leq K_1 \frac{\sqrt{\mathbb{E}^*[M_{\delta,\delta_1'}^2]}}{\sqrt{n}}$$

(see Theorems 2.14.1 and 2.14.2 in Van der Vaart and Wellner (1996)). Hence, in this case, the last part of (B2) holds if $\Phi_n(\delta)$ can be chosen such that

$$\exists K_0, \forall \delta \leq \delta_0 : \sqrt{\mathbb{E}^*[M_{\delta,\delta_1'}^2]} \leq K_0 \Phi_n(\delta). \tag{5}$$

The function $\Phi_n(\delta)$ is closely related to the 'smoothness' of the functions $\theta \to m(z, \theta, h(z, \theta))$. When these functions are Lipschitz (respectively Hölder of order γ) uniformly over z and h, it is possible to take $\Phi_n(\delta) = \delta$ (respectively $\Phi_n(\delta) = \delta^\gamma$). In other situations, for instance when m contains an indicator function involving θ, such regularity assumptions may fail but it is possible to state (B2) with $\Phi_n(\delta) = \sqrt{\delta}$ (see Section 6).

(iii) The way $\Phi_n(\delta)$ decreases when δ tends to zero has a crucial impact on the convergence rate r_n through the condition $r_n^2 \Phi_n(r_n^{-1}) \leq \sqrt{n}$. When $\Phi_n(\delta) = \delta$, this last condition is equivalent to $r_n \leq \sqrt{n}$ and we may obtain $\sqrt{n}$ convergence rates. However, when $\Phi_n(\delta) = \delta^\gamma$ with $\gamma < 1$, this condition is equivalent to $r_n^{2-\gamma} \leq \sqrt{n}$ and hence only $n^{\frac{1}{2(2-\gamma)}}$ rates may be considered. In the case of non continuous criterion functions (involving e.g. indicator functions) a $n^{\frac{1}{3}}$ convergence rate may be obtained, analogously to the case of parametric M-estimation.

(iv) Assumption (B4) is automatically fulfilled under the following classical assumption:

$$M_n(\widehat{\theta}, \widehat{h}) \geq \sup_{\theta \in \Theta} M_n(\theta, \widehat{h}) + O_{P^*}(r_n^{-2}).$$

As in the previous section, this allows to consider approximations of the value that actually maximizes the empirical criterion.

(v) Assumption (B3) is automatically fulfilled when the following conditions hold:

(a) $\Theta \subset \mathbb{R}^k$ for some k, and $d(\theta_1, \theta_2) = \|\theta_1 - \theta_2\|$, where $\|\cdot\|$ is the Euclidean norm.

(b) There exists $\delta_2 > 0$ such that for any h satisfying $d_{\mathcal{H}}(h, h_0) \leq \delta_2$, the function $\theta \to M(\theta, h)$ is twice continuously differentiable on an open neighborhood of θ_0. Hereafter, $\Gamma(\theta_0, h)$ and $\Lambda(\theta_0, h)$ denote respectively its gradient and Hessian matrix for $\theta = \theta_0$. Moreover,

$$\lim_{\|\theta-\theta_0\| \to 0} \sup_{d_{\mathcal{H}}(h,h_0) \leq \delta_2} \|\theta - \theta_0\|^{-2} \Big| M(\theta, h) - M(\theta_0, h) - \Gamma(\theta_0, h)(\theta - \theta_0) - \frac{1}{2}(\theta - \theta_0)^T \Lambda(\theta_0, h)(\theta - \theta_0) \Big| = 0.$$

(c) $\|\Gamma(\theta_0, \widehat{h})\| = O_{P^*}(r_n^{-1})$ and $\Gamma(\theta_0, h_0) = 0$.

(d) $\Lambda(\theta_0, h_0)$ is negative definite, and $h \mapsto \Lambda(\theta_0, h)$ is continuous in h_0 (i.e. $\lim_{d_{\mathcal{H}}(h,h_0)\to 0} \sup_{u\in\mathbb{R}^k,\|u\|=1} \|(\Lambda(\theta_0, h) - \Lambda(\theta_0, h_0))u\| = 0$).

Now denote the greatest eigenvalue of $\Lambda(\theta_0, h_0)$ by λ_m. When $d_{\mathcal{H}}(\widehat{h}, h_0) \leq \delta_2$, assumptions (a)-(d) above imply that

$$\begin{aligned} & M(\theta, \widehat{h}) - M(\theta_0, \widehat{h}) \\ &= \langle \Gamma(\theta_0, \widehat{h}), \gamma_\theta \rangle + \frac{1}{2}(\gamma_\theta)^T \Lambda(\theta_0, h_0)(\gamma_\theta) + \|\gamma_\theta\|^2 o_{P^*}(1) + o(\|\gamma_\theta\|^2) \\ &\leq \|\Gamma(\theta_0, \widehat{h})\|\|\gamma_\theta\| + \frac{\lambda_m}{2}\|\gamma_\theta\|^2 + \|\gamma_\theta\|^2 o_{P^*}(1) + o(\|\gamma_\theta\|^2), \end{aligned}$$

where $\gamma_\theta = \theta - \theta_0$ and where the notation $o(\|\gamma_\theta\|^2)$ means $\lim_{\|\gamma_\theta\|\to 0} \frac{o(\|\gamma_\theta\|^2)}{\|\gamma_\theta\|^2} = 0$. By taking δ_0 such that $\|\theta - \theta_0\| \leq \delta_0$, the last term above is bounded by $-\frac{\lambda_m}{4}\|\theta - \theta_0\|^2$, and hence (B3) holds with $W_n = \|\Gamma(\theta_0, \widehat{h})\|$ and $C = -\frac{\lambda_m}{4}$.

(vi) Finally, it is possible to modify slightly the proof of the following theorem by considering the following extensions of assumptions (B3) and (B4):

(B3′) There exist $\eta_0 > 0$, and two positive and non-decreasing functions Ψ_1 and Ψ_2 on $(0, \eta_0]$ such that for all θ satisfying $d(\theta, \theta_0) \leq \eta_0$:

$$M(\theta, \widehat{h}) - M(\theta_0, \widehat{h}) \leq W_n \Psi_1(d(\theta, \theta_0)) - (1 + o_{P^*}(1))\Psi_2(d(\theta, \theta_0)).$$

Moreover, there exist $\beta_2 > \alpha$, $\beta_1 < \beta_2$, $\delta_0 > 0$ such that $\delta \mapsto \Psi_1(\delta)\delta^{-\beta_1}$ is non-increasing and $\delta \mapsto \Psi_2(\delta)\delta^{-\beta_2}$ is non-decreasing on $(0, \delta_0]$, and such that $\Psi_1(r_n^{-1})W_n = O_{P^*}(\Psi_2(r_n^{-1}))$ for some sequence $r_n \to \infty$.

(B4′) $M_n(\widehat{\theta}, \widehat{h}) \geq M_n(\theta_0, \widehat{h}) + O_p(\Psi_2(r_n^{-1}))$ and $\Phi_n(r_n^{-1}) \leq \sqrt{n}\Psi_2(r_n^{-1})$.

It is possible to consider the case where $\beta_1 = \beta_2$ if we assume that $\Psi_1(r_n^{-1})W_n = o_p(\Psi_2(r_n^{-1}))$.

We are now ready to state the rate of convergence of the estimator $\widehat{\theta}$. The proof of this result can be found in the Appendix.

Theorem 2 *Under assumptions (B1)-(B4) we have that*

$$r_n d(\widehat{\theta}, \theta_0) = O_{P^*}(1).$$

5 Asymptotic distribution

In the previous section, we have shown that $r_n d(\widehat{\theta}, \theta_0) = O_{P^*}(1)$. Our aim is now to study the asymptotic distribution of $r_n(\widehat{\theta} - \theta_0)$. We will assume throughout this section that Θ is equipped with the Euclidean norm $\|\cdot\|$. We start with introducing a number of notations. For any $\theta \in \Theta$ and $h \in \mathcal{H}$, let $B_n(\theta, h) = M_n(\theta, h) - M_n(\theta_0, h)$ and $B(\theta, h) = M(\theta, h) - M(\theta_0, h)$, and define

$$M_\delta(.) \geq \sup_{\|\theta - \theta_0\| \leq \delta} |m(., \theta, h_0) - m(., \theta_0, h_0)|$$

for any $\delta > 0$. Also, let

$$\mathcal{M}_\delta = \{m(., \theta, h_0) - m(., \theta_0, h_0) : \|\theta - \theta_0\| \leq \delta\}.$$

Finally, for any $p \in \mathbb{N}$, for any $f : \Theta \to I\!R$ and for any $\gamma = (\gamma_1, \ldots, \gamma_p) \in \Theta^p$ denote $\overline{f}_\gamma = (f(\gamma_1), \ldots, f(\gamma_p))^T$.

We introduce the following assumptions:

(C1) $r_n\|\widehat{\theta} - \theta_0\| = O_{P^*}(1)$ and $v_n d_{\mathcal{H}}(\widehat{h}, h_0) = O_{P^*}(1)$ for some sequences $r_n \to \infty$ and $v_n \to \infty$.

(C2) θ_0 belongs to the interior of Θ and $\Theta \subset (E, \|\cdot\|)$, where E is a finite dimensional Euclidean space (i.e. $E = \mathbb{R}^k$ for some k).

(C3) For all $\delta_2, \delta_3 > 0$,

$$\sup_{\|\theta - \theta_0\| \leq \frac{\delta_2}{r_n},\, d_{\mathcal{H}}(h, h_0) \leq \frac{\delta_3}{v_n}} \frac{|B_n(\theta, h) - B(\theta, h) - B_n(\theta, h_0) + B(\theta, h_0)|}{r_n^{-2} + |B_n(\theta, h)| + |B_n(\theta, h_0)| + |B(\theta, h)| + |B(\theta, h_0)|} = o_{P^*}(1).$$

(C4) For all $K, \eta > 0$,

$$\frac{r_n^4}{n}\mathbb{E}^*\Big[M_{\frac{K}{r_n}}^2\Big] = O(1) \quad \text{and} \quad \frac{r_n^4}{n}\mathbb{E}^*\Big[M_{\frac{K}{r_n}}^2 1_{\{r_n^2 M_{\frac{K}{r_n}} > \eta n\}}\Big] = o(1).$$

(C5) For all $K > 0$ and for any $\eta_n \to 0$,

$$\sup_{\|\gamma_1 - \gamma_2\| < \eta_n, \|\gamma_1\| \vee \|\gamma_2\| \leq K} \frac{r_n^4}{n}\mathbb{E}\Big[m\Big(Z, \theta_0 + \frac{\gamma_1}{r_n}, h_0\Big) - m\Big(Z, \theta_0 + \frac{\gamma_2}{r_n}, h_0\Big)\Big]^2 = o(1).$$

(C6) For all $z \in F$, the function $\theta \mapsto m(z, \theta, h_0(z, \theta))$ and almost all paths of the process $\theta \mapsto m(z, \theta, \widehat{h}(\theta, z))$ are uniformly (over θ) bounded on compact sets.

(C7) There exist $\beta_n = o_{P^*}(1)$, a random and linear function $W_n : E \to \mathbb{R}$, and a deterministic and bilinear function $V : E \times E \to \mathbb{R}$ such that for all $\theta \in \Theta$,

$$B(\theta, \widehat{h}) = W_n(\gamma_\theta) + V(\gamma_\theta, \gamma_\theta) + \beta_n \|\gamma_\theta\|^2 + o(\|\gamma_\theta\|^2)$$

and

$$B(\theta, h_0) = V(\gamma_\theta, \gamma_\theta) + o(\|\gamma_\theta\|^2),$$

where $\gamma_\theta = \theta - \theta_0$ and the notation $o(\|\gamma_\theta\|^2)$ means $\lim_{\|\gamma_\theta\| \to 0} \frac{o(\|\gamma_\theta\|^2)}{\|\gamma_\theta\|^2} = 0$.
Moreover, for any compact set $\mathcal{K}$ in E,

$$\exists \tau,\, \delta_1 > 0,\; r_n \sup_{\substack{\gamma \in E,\, \delta \le \delta_1, \\ \|\gamma\| \le \delta}} \left|\frac{W_n(\gamma)}{\delta^\tau}\right| = O_{P^*}(1) \text{ and } \sup_{\substack{\gamma, \gamma' \in \mathcal{K},\, \delta \le \delta_1, \\ \|\gamma - \gamma'\| \le \delta}} \frac{|V(\gamma, \gamma) - V(\gamma', \gamma')|}{\delta^\tau} < \infty.$$

(C8) For all $K > 0$, there exists $n_0 \in \mathbb{N}$ such that for all $n \ge n_0$,

$$M_n(\widehat{\theta}, \widehat{h}) \ge \sup_{\|\theta - \theta_0\| \le \frac{K}{r_n}} M_n(\theta, \widehat{h}) + o_{P^*}(r_n^{-2}).$$

(C9) There exists a deterministic continuous function Λ and a zero-mean Gaussian process $\mathbb{G}$ defined on E such that for all $p \in \mathbb{N}$ and for all $\gamma = (\gamma_1, \ldots, \gamma_p) \in E^p$,

$$r_n \overline{W_{n\gamma}} + r_n^2 \overline{B_n\Big(\theta_0 + \frac{\cdot}{r_n}, h_0\Big)}_\gamma \xrightarrow{\mathcal{L}} \overline{\Lambda}_\gamma + \overline{\mathbb{G}}_\gamma.$$

Moreover, $\mathbb{G}(\gamma) = \mathbb{G}(\gamma')$ a.s. implies that $\gamma = \gamma'$, and $P^*\big(\text{limsup}_{\|\gamma\| \to +\infty}(\Lambda_\gamma + \mathbb{G}_\gamma) < \sup_{\gamma \in E}(\Lambda_\gamma + \mathbb{G}_\gamma)\big) = 1$.

(C10) There exists a $\delta_0 > 0$ such that

$$\int_0^\infty \sup_{\delta \le \delta_0} \sqrt{\log\Big(N_{[]}(\epsilon \|M_\delta\|_{P^*,2},\, \mathcal{M}_\delta,\, \mathbb{L}^2(P))\Big)}\, d\epsilon < +\infty.$$

We will show below that $r_n(\widehat{\theta} - \theta_0)$ converges to the unique maximizer of the process $\gamma \mapsto \Lambda(\gamma) + \mathbb{G}(\gamma)$, where Λ and $\mathbb{G}$ are defined in (C9). However, let us first discuss the above assumptions.

Remark 3

(i) The first part of assumption (C1) can be obtained from Theorem 2. If in addition we assume that assumption (B2) holds with $\Phi_n \equiv \Phi$ not depending on n and

continuous, and if we take $r_n \to +\infty$ such that $r_n^2 \Phi(r_n^{-1}) = \sqrt{n}$, then assumptions (C4) and (C5) are implied by the following ones: there exists a $\delta_4 > 0$ such that for all $\delta \leq \delta_4$, $\mathbb{E}^*(M_\delta^2) \leq K\Phi^2(\delta)$ for some $K > 0$,

$$\lim_{\delta \to 0} \frac{\mathbb{E}^*[M_\delta^2 1_{\{M_\delta > \eta \delta^{-2}\Phi^2(\delta)\}}]}{\Phi^2(\delta)} = 0$$

for all $\eta > 0$, and

$$\lim_{\epsilon \to 0} \lim_{\delta \to 0} \sup_{\|\gamma_1 - \gamma_2\| \leq \epsilon,\, \|\gamma_1\| \vee \|\gamma_2\| \leq K} \frac{\mathbb{E}[m(Z, \theta_0 + \gamma_1 \delta, h_0) - m(Z, \theta_0 + \gamma_2 \delta, h_0)]^2}{\Phi^2(\delta)} = 0$$

for all $K > 0$, using the same arguments as in the proof on Theorem 3.2.10 in Van der Vaart and Wellner (1996).

(ii) Assumption (C6) ensures that for any compact $\mathcal{K} \subset E$ the processes $\gamma \mapsto r_n^2 B_n(\theta_0 + \frac{\gamma}{r_n}, \widehat{h})$ and $\gamma \mapsto r_n^2 B_n(\theta_0 + \frac{\gamma}{r_n}, h_0) + r_n W_n(\gamma)$ take values in $\ell^\infty(\mathcal{K})$ (assumption (C7) is also used for the second process). This assumption is not very restrictive. Moreover, because we deal with asymptotic results, we actually only require the latter properties for $n > n_{\mathcal{K}}$ (where $n_{\mathcal{K}}$ only depends on $\mathcal{K}$). It follows directly from (C1) and (C2) that there exists $n_{\mathcal{K}}$ such that $\theta_0 + \frac{\mathcal{K}}{r_n} \subset \Theta$ for $n > n_{\mathcal{K}}$. Hence it is only necessary to state (C6) on the compact set Θ.

(iii) Assumption (C3) is automatically fulfilled under the following slightly more restrictive (but common) assumption: for all $\delta_2, \delta_3 > 0$,

$$\sup_{\|\theta - \theta_0\| \leq \frac{\delta_2}{r_n}, d_{\mathcal{H}}(h, h_0) \leq \frac{\delta_3}{v_n}} |B_n(\theta, h) - B(\theta, h) - B_n(\theta, h_0) + B(\theta, h_0)| = o_{P^*}(r_n^{-2}).$$

The latter condition holds whenever the following one is fulfilled: there exists a function f and a constant $\delta_0 > 0$ such that for all $\delta_2, \delta_3 < \delta_0$,

$$r_n^2 f\Big(\frac{\delta_2}{r_n}, \frac{\delta_3}{v_n}\Big) = o(\sqrt{n}),$$

and

$$\mathbb{E}^*\Big[\sup_{\|\theta - \theta_0\| \leq \frac{\delta_2}{r_n}, d_{\mathcal{H}}(h, h_0) \leq \frac{\delta_3}{v_n}} |B_n(\theta, h) - B(\theta, h) - B_n(\theta, h_0) + B(\theta, h_0)|\Big] \leq \frac{f(\frac{\delta_2}{r_n}, \frac{\delta_3}{v_n})}{\sqrt{n}}.$$

This last bound may be obtained using arguments that are similar to those discussed in Remark 2(ii).

(iv) Now assume that assumptions (a)-(d) from Remark 2(v) hold. Following the same ideas as in this remark it is easy to show that (C7) is fulfilled with $E = \mathbb{R}^k$, $W_n(\gamma) = \langle \Gamma(\theta_0, \widehat{h}), \gamma \rangle$ and $V(\gamma, \gamma) = \frac{1}{2}\gamma^T \Lambda(\theta_0, h_0)\gamma$ whenever $\sup_{u \in \mathbb{R}^k, \|u\|=1} \|\Lambda(\theta_0, h_0)u\| < +\infty$. Moreover, in that case, if $\widehat{h}$ is computed from a dataset independent of $(Z_1, \ldots, Z_n)$, it is sufficient for (C9) to assume the weak convergence of each term $r_n \overline{W_{n\gamma}}$ and $r_n^2 \overline{B_n(\theta_0 + \frac{\cdot}{r_n}, h_0)}_\gamma$ separately. The convergence of the second term can be obtained as in the parametric case (see Theorem 3.2.10 in Van der Vaart and Wellner (1996)). Note also that if $r_n \Gamma(\theta_0, \widehat{h}) \to W$ in distribution, the marginals of the process $\gamma \mapsto \langle r_n \Gamma(\theta_0, \widehat{h}), \gamma \rangle$ tend in distribution to the marginals of $\gamma \mapsto \langle W, \gamma \rangle$. Furthermore, if $r_n = \sqrt{n}$, it is common to assume that $\Gamma(\theta_0, \widehat{h}) = n^{-1} \sum_{i=1}^n U_{i,n} + o_{P^*}(n^{-1/2})$, where the variables $U_{i,n}$ are independent and centered. The convergence then follows from Lindeberg's condition.

(v) Assumption (C8) allows to consider estimators $\widehat{\theta}$ that are approximations of the value that actually maximizes the map $\theta \mapsto M_n(\theta, \widehat{h})$.

(vi) Let $\mathcal{K}$ be an arbitrary compact subset in E. Assumption (C9) is used to derive the weak convergence (in the $\ell^\infty(\mathcal{K})$ sense) of the process $\gamma \mapsto r_n W_n(\gamma) + r_n^2 B_n(\theta_0 + \gamma r_n^{-1}, h_0)$ from the fact that it is asymptotically tight. If $r_n \sup_{\gamma \in \mathcal{K}, \gamma \neq 0} \|W_n(\gamma)\|\gamma\|^{-1}\| = o_{P^*}(1)$, we are in the same situation as in the parametric case and we obtain the convergence of the marginals whenever

$$\lim_{n \to \infty} \frac{r_n^4}{n} \mathbb{E}\Big\{ \Big[m\Big(Z, \theta_0 + \frac{\gamma_1}{r_n}, h_0\Big) - m\Big(Z, \theta_0 + \frac{\gamma_2}{r_n}, h_0\Big) \Big]^2 \Big\} = \mathbb{E}[(\mathbb{G}(\gamma_1) - \mathbb{G}(\gamma_2))^2]$$

for all γ_1, γ_2, by noting that the remaining term is a sum of an array of random variables that fulfill Lindeberg's condition (see Van der Vaart and Wellner (1996) p. 293-294). The last assumption on the process $\gamma \mapsto \Lambda_\gamma + \mathbb{G}_\gamma$ is used to ensure almost all sample paths have a supremum which is only related to their behaviour on compact sets. The dominant term of the deterministic part Γ is usually a negative definite quadratic form and hence exponential inequalities could lead to such result.

(vii) Finally, assumption (C10) is used to show that $\gamma \mapsto r_n^2 B_n(\theta_0 + \gamma r_n^{-1}, h_0)$ is asymptotically tight. It is the same as in the 'parametric case' where h_0 is known (see Theorem 3.2.10 in Van der Vaart and Wellner (1996)). The same holds true for assumptions (C4)-(C6). Van der Vaart and Wellner (1996) also give weaker conditions and alternatives for assumption (C10) based on covering numbers (see Theorems 2.11.22, 2.11.23 and 3.2.10).

We are now ready to state the main result of the paper about the asymptotic distribution of $\widehat{\theta} - \theta_0$. As before, we refer to the Appendix for the proof.

Theorem 3 *If assumptions (C1)-(C10) hold, then for all $K > 0$ the process $\gamma \mapsto r_n^2 B_n(\theta_0 + \frac{\gamma}{r_n}, \widehat{h})$ converges weakly to $\gamma \mapsto \Lambda(\gamma) + \mathbb{G}(\gamma)$ in $\ell^\infty(\mathcal{K})$ with $\mathcal{K} = \{\gamma \in E : \|\gamma\| \leq K\}$. Moreover, for any such $\mathcal{K}$ almost all paths of the limiting process have a unique maximizer γ_0 on $\mathcal{K}$. Assume now that γ_0 is measurable. Then, the random sequence $r_n(\widehat{\theta} - \theta_0)$ converges in distribution to γ_0.*

6 Examples

In this section we give several examples of situations in which existing theory on semiparametric estimators can not be applied, whereas Theorems 1–3 in this paper can be used to obtain the limiting distribution of the estimator. This will demonstrate the usefulness of the asymptotic results in this paper. We start with an example on classification with missing data, which we work out in full detail. The contexts of the other five examples in this section are shortly stated, but the verification of the conditions of Theorems 1–3 is omitted for obvious space restrictions. See also the paper by Xu, Sen and Ying (2014), who consider a Cox model for duration data containing a change point (threshold). Their model and estimator also fit in our general framework.

6.1 Classification with missing data

In this section we illustrate the theory, and in particular the verification of the assumptions, by means of an example coming from the area of classification with missing data.

Consider i.i.d. data $X_i = (X_{i1}, X_{i2})$ $(i = 1, \ldots, n)$ having the same distribution as $X = (X_1, X_2)$. We suppose that these data come in reality from two underlying populations. Let Y_i be j if observation i belongs to population j $(j = 0, 1)$, and let Y be the population indicator for the vector X. Based on these data, we wish to establish a classification rule for new observations, for which it will be unknown to which population they belong. The classification consists in regressing X_2 on X_1 via a parametric regression function $f_\theta(\cdot)$, and choosing θ by maximizing the criterion

$$P(Y = 1, X_2 \geq f_\theta(X_1)) + P(Y = 0, X_2 < f_\theta(X_1)). \tag{6}$$

Let θ_0 be the value of θ that maximizes (6) with respect to all $\theta \in \Theta$, where Θ is a compact subset of $\mathbb{R}^k$, whose interior contains θ_0.

We suppose now that some of the Y_i's are missing. Let Δ_i (respectively Δ) be 1 if Y_i (respectively Y) is observed, and 0 otherwise. Hence our data consist of i.i.d. vectors $Z_i = (X_i, Y_i\Delta_i, \Delta_i)$ $(i = 1, \ldots, n)$. We assume that the missing at random mechanism holds true, in the sense that

$$P(\Delta = 1|X_1, X_2, Y) = P(\Delta = 1|X_1) := p_0(X_1).$$

Note that (6) equals

$$E\Big[\frac{I(\Delta = 1)}{p_0(X_1)}\Big\{I(Y = 1, X_2 \geq f_\theta(X_1)) + I(Y = 0, X_2 < f_\theta(X_1))\Big\}\Big].$$

Hence, it is natural to define

$$m(Z, \theta, p) = \frac{I(\Delta = 1)}{p(X_1)}\Big\{I(Y = 1, X_2 \geq f_\theta(X_1)) + I(Y = 0, X_2 < f_\theta(X_1))\Big\}, \quad (7)$$

where the nuisance function $p(\cdot)$ belongs to a space $\mathcal{P}$ to be defined later, and where $Z = (X, Y\Delta, \Delta)$. Also, let

$$M(\theta, p) = E[m(Z, \theta, p)] \qquad \text{and} \qquad M_n(\theta, p) = n^{-1}\sum_{i=1}^{n} m(Z_i, \theta, p).$$

Finally, define the estimator $\widehat{\theta}$ of θ_0 by

$$\widehat{\theta} = \text{argmax}_{\theta\in\Theta}\, M_n(\theta, \widehat{p}), \quad (8)$$

where for any x_1,

$$\widehat{p}(x_1) = \sum_{i=1}^{n} \frac{k_h(x_1 - X_{i1})}{\sum_{j=1}^{n} k_h(x_1 - X_{j1})} I(\Delta_i = 1),$$

where k is a density function with support $[-1, 1]$, $k_h(u) = k(u/h)/h$ and $h = h_n$ is an appropriate bandwidth sequence.

We will now check the conditions of Theorems 1, 2 and 3. Suppose $d(\theta, \theta_0)$ is the Euclidean distance $\|\cdot\|$. Let $\mathcal{P}$ be the space of functions $p : R_{X_1} \to I\!R$ that are continuously differentiable, and for which $\sup_{x_1\in R_{X_1}} p(x_1) \leq M < \infty$, $\sup_{x_1\in R_{X_1}} |p'(x_1)| \leq M$ and $\inf_{x_1\in R_{X_1}} p(x_1) > \eta/2$, where $\eta = \inf_{x_1\in R_{X_1}} p_0(x_1)$ is supposed to be strictly positive, and where R_{X_1} is the support of X_1, which is supposed to be a compact subspace of $I\!R$. We equip the space $\mathcal{P}$ with the supremum norm : $d_{\mathcal{P}}(p_1, p_2) = \sup_{x_1\in R_{X_1}} |p_1(x_1) - p_2(x_1)|$ for any p_1, p_2.

First of all, (A1) is verified by construction of the estimator $\widehat{\theta}$. Condition (A2) is an identifiability condition, needed to ensure that θ_0 is unique, whereas (A3) holds true

provided the functions p_0 and k are continuously differentiable. Next, for (A4) is suffices by Remark 1(ii) to show that the class $\mathcal{F} = \{m(\cdot, \theta, p) : \theta \in \Theta, p \in \mathcal{P}\}$ is Glivenko-Cantelli. For this we will show that for all $\epsilon > 0$, the bracketing number $N_{[]}(\epsilon, \mathcal{F}, \mathbb{L}_1(P))$ is finite. First, note that it follows from Corollary 2.7.2 in Van der Vaart and Wellner (1996) that $N_{[]}(\epsilon, \mathcal{P}, \mathbb{L}_1(P)) \leq \exp(K\epsilon^{-1})$. In a similar way we can show that $N_{[]}(\epsilon, \{f_\theta : \theta \in \Theta\}, \mathbb{L}_1(P)) \leq \exp(K\epsilon^{-1})$, provided $x_1 \to f_\theta(x_1)$ is continuously differentiable and the derivatives are uniformly bounded over θ. From there it can be easily shown that the class

$$\mathcal{T} = \{(x_1, x_2) \to I(x_2 \geq f_\theta(x_1)) : \theta \in \Theta\}$$

satisfies $N_{[]}(\epsilon, \mathcal{T}, \mathbb{L}_1(P)) \leq \exp(K\epsilon^{-1})$, provided $\sup_{x_1,x_2} f_{X_2|X_1}(x_2|x_1) < \infty$. By combining the brackets for $\mathcal{P}$ and $\mathcal{T}$ we get that $N_{[]}(\epsilon, \mathcal{F}, \mathbb{L}_1(P)) \leq \exp(K\epsilon^{-1}) < \infty$ for some $K < \infty$. Finally, condition (A5) is straightforward, and hence the weak consistency of $\widehat{\theta}$ follows.

Next, we verify the B-conditions. Condition (B1) holds with $v_n^{-1} = K[(nh)^{-1/2}(\log n)^{1/2} + h]$. For (B2), it suffices by Remark 2(ii) to show that (4) and (5) hold true. Equation (5) holds true for $\Phi_n(\delta) = \delta^{1/2}$. Indeed the envelope $M_{\delta,\delta_1'}$ of the class $\mathcal{F}_{\delta,\delta_1'}$ can be taken equal to (for notational simplicity we suppose throughout that θ is one-dimensional)

$$M_{\delta,\delta_1'}(Z) = \frac{2}{\eta} I(f_{\theta_0}(X_1) - A\delta \leq X_2 \leq f_{\theta_0}(X_1) + A\delta),$$

where $A = \sup_{\theta,x_1} |\frac{\partial}{\partial\theta} f_\theta(x_1)|$, which we suppose to exist and to be finite. Hence, (5) is easily seen to hold provided X is absolutely continuous and $\sup_x f_X(x) < \infty$. For (4) note that

$$\begin{aligned} \|M_{\delta\delta_1'}\|_{L_2(P)}^2 &= \frac{4}{\eta^2} E\Big[F_{X_2|X_1}(f_{\theta_0}(X_1) + A\delta|X_1) - F_{X_2|X_1}(f_{\theta_0}(X_1) - A\delta|X_1)\Big] \\ &\geq \frac{8A\delta}{\eta^2} \inf{}^* f_{X_2|X_1}(x_2|x_1) =: K_1^2\delta, \end{aligned}$$

which we suppose to be strictly positive, where $\inf^*$ is the infimum over all (x_1, x_2) such that $|x_2 - f_{\theta_0}(x_1)| \leq A\delta$. It follows that $N_{[]}(\epsilon\|M_{\delta\delta_1'}\|_{L_2(P)}, \mathcal{F}_{\delta\delta_1'}, L_2(P))$ is bounded above by $N_{[]}(K_1\epsilon\delta^{1/2}, \mathcal{F}_{\delta\delta_1'}, L_2(P))$. We will first construct brackets for the set $\mathcal{G} := \{f_\theta : \theta \in \Theta, |\theta - \theta_0| \leq \delta\}$. Note that f_θ can be written as $f_\theta = [\frac{1}{\delta}(f_\theta - f_{\theta_0})]\delta + f_{\theta_0}$. If we assume that $x_1 \to \frac{\partial}{\partial\theta} f_\theta(x_1)$ is twice continuously differentiable in x_1 for all θ, it follows from Corollary 2.7.2 in Van der Vaart and Wellner (1996) that $r_\epsilon := N_{[]}(\epsilon^2, \mathcal{D}, L_2(P)) \leq \exp(K\varepsilon^{-1})$ with $\mathcal{D} = \{\frac{1}{\delta}(f_\theta - f_{\theta_0}) : \theta \in \Theta, |\theta - \theta_0| \leq \delta\}$. Let $d_1^L \leq d_1^U, \ldots, d_{r_\epsilon}^L \leq d_{r_\epsilon}^U$ be the ϵ^2-brackets for $\mathcal{D}$. It then easily follows that $N_{[]}(\epsilon^2\delta, \mathcal{G}, L_2(P)) = r_\epsilon$, and that the $\epsilon^2\delta$-brackets for $\mathcal{G}$ are given

by $g_j^L := d_j^L\delta + f_{\theta_0} \le d_j^U\delta + f_{\theta_0} =: g_j^U$. Moreover, $s_\epsilon := N_{[]}(\epsilon, \mathcal{P}, L_\infty(P)) \le \exp(K\epsilon^{-1})$. Let $h_1^L \le h_1^U, \ldots, h_{s_\epsilon}^L \le h_{s_\epsilon}^U$ be the ϵ-brackets for $\mathcal{P}$ defined in such a way that for $1 \le k \le s_\epsilon$, $\inf_{t\in R_{X_1}} h_k^L \le \eta$. We now claim that

$$N_{[]}(\epsilon \|M_{\delta\delta_1'}\|_{L_2(P)}, \mathcal{F}_{\delta\delta_1'}, L_2(P)) \le r_\epsilon s_\epsilon \le \exp(K\epsilon^{-1}). \tag{9}$$

Indeed, define for $1 \le j \le r_\epsilon$ and $1 \le k \le s_\epsilon$,

$$\begin{aligned} f_{jk}^L(Z) = \frac{I(\Delta=1)}{h_k^U(X_1)}\Big\{ & I(Y=1)\Big[I(X_2 \ge g_j^U(X_1)) - I(X_2 \ge f_{\theta_0}(X_1))\Big] \\ & + I(Y=0)\Big[I(X_2 < g_j^L(X_1)) - I(X_2 < f_{\theta_0}(X_1))\Big]\Big\}, \end{aligned}$$

and define in a similar way the upper bracket $f_{jk}^U(Z)$. Then,

$$\begin{aligned} & \|f_{jk}^U(Z) - f_{jk}^L(Z)\|_2^2 \\ & \le 4E\Big(\Big[\frac{1}{h_k^U(X_1)} - \frac{1}{h_k^L(X_1)}\Big]^2\Big[\Big|P(X_2 \le g_j^U(X_1)|X_1) - P(X_2 \le f_{\theta_0}(X_1)|X_1)\Big| \\ & \qquad + \Big|P(X_2 \le g_j^L(X_1)|X_1) - P(X_2 \le f_{\theta_0}(X_1)|X_1)\Big|\Big]\Big) \\ & \quad + \frac{4}{\eta^2}E\Big(P(X_2 \le g_j^U(X_1)|X_1) - P(X_2 \le g_j^L(X_1)|X_1)\Big) \\ & \le \frac{16\delta}{\eta^2}\sup_{x_1}|h_k^U(x_1) - h_k^L(x_1)|^2 \sup_{x_1,x_2} f_{X_2|X_1}(x_2|x_1)E\Big[|d_j^U(X_1)| + |d_j^L(X_1)|\Big] \\ & \quad + \frac{4}{\eta^2}\sup_{x_1,x_2} f_{X_2|X_1}(x_2|x_1)E\Big[g_j^U(X_1) - g_j^L(X_1)\Big] \\ & \le C\epsilon^2\delta, \end{aligned}$$

for some $0 < C < \infty$. Moreover, for each function in the class $\mathcal{F}_{\delta\delta_1'}$ there exists a bracket $[f_{jk}^L, f_{jk}^U]$ to which it belongs. This shows (9). It now follows that

$$\sup_{\delta\le\delta_0}\int_0^1 \sqrt{1 + \log N_{[]}(\epsilon\|M_{\delta\delta_1'}\|_{L_2(P)}, \mathcal{F}_{\delta\delta_1'}, L_2(P))}\, d\epsilon < \infty,$$

which shows (4) and hence (B2).

For (B3) we check conditions (b)-(d) of Remark 2(v). It is easily seen that (b) holds with

$$\Gamma(\theta_0, p) = E\Big[\frac{p_0(X_1)}{p(X_1)}\Big\{1 - 2P(Y=1|X_1,X_2)\Big\}f_{X_2|X_1}(f_{\theta_0}(X_1))\frac{\partial}{\partial\theta}f_{\theta_0}(X_1)\Big],$$

and

$$\begin{aligned} \Lambda(\theta_0, p) = E\Big[\frac{p_0(X_1)}{p(X_1)}\Big\{1 - 2P(Y=1|X_1,X_2)\Big\}\Big\{ & f'_{X_2|X_1}(f_{\theta_0}(X_1))\Big(\frac{\partial}{\partial\theta}f_{\theta_0}(X_1)\Big)^2 \\ & + f_{X_2|X_1}(f_{\theta_0}(X_1))\frac{\partial^2}{\partial\theta^2}f_{\theta_0}(X_1)\Big\}\Big], \end{aligned}$$

and provided the derivatives in $\Lambda(\theta_0, p)$ all exist. Next, by assuming that $\Gamma(\theta_0, p_0) = 0$ and that $\Lambda(\theta_0, p_0)$ is negative, and by noting that $\|\Gamma(\theta_0, \widehat{p})\| = O_P(r_n^{-1})$ if r_n satisfies $r_n[n^{-1/2} + h + (nh)^{-1}\log n)] = O(1)$, it follows that (c) and (d) are also valid. It remains to check condition (B4), which easily holds provided $r_n = O(n^{1/3})$. The two conditions on r_n and the fact that r_n should be chosen as large as possible, are reconcilable provided $nh^3 = O(1)$ and $(nh^{3/2})^{-1}(\log n)^{3/2} = O(1)$. Note that it is possible to weaken the first condition to $nh^6 = O(1)$ if we assume that $p_0(\cdot)$ is twice continuously differentiable. Note however that the rate v_n^{-1} of $\widehat{p}$ would then be $O((nh)^{-1/2}(\log n)^{1/2} + h^2)$, which is faster than the rate $r_n^{-1} = Kn^{-1/3}$ of $\widehat{\theta}$ provided $nh^3 \to \infty$. Hence, the latter case is of lower level of complexity than the case where p_0 is only once differentiable, and we therefore do not consider it further. To conclude, we have that

$$\widehat{\theta} - \theta_0 = O_{P^*}(n^{-1/3}).$$

Finally, we check the conditions needed for establishing the asymptotic distribution of $\widehat{\theta}$. Condition (C1) follows from Theorem 2 and condition (B1), whereas (C2) is immediately satisfied. For (C3) a similar proof as for condition (B2) can be given, which we omit for reasons of brevity. For (C4) and (C5), first note that the function $\Phi_n(\delta) = K\delta^{1/2}$ in condition (B2) is independent of n and continuous. Hence, (C4) and (C5) hold provided the three conditions stated in Remark 3(i) are verified. For the first one, we have that M_δ satisfies

$$|M_\delta(Z)| \leq \frac{2}{\eta} I\Big(f_{\theta_0}(X_1) - A\delta \leq X_2 \leq f_{\theta_0}(X_1) + A\delta\Big).$$

Hence, $E^*(M_\delta^2) \leq K\delta$ for some $K < \infty$. In a similar way the second and third condition can be proved, from which (C4) and (C5) follow. Next, (C6) is obviously satisfied since for fixed p and z, our function $m(z, \cdot, p)$ consists of indicator functions. Next, following Remarks 2(v) and 3(v), condition (C7) follows provided $|\Lambda(\theta_0, p_0)| < \infty$. By construction of the estimator $\widehat{\theta}$, condition (C8) holds true. For (C9), first note that $r_n W_n(\gamma) = r_n \Gamma(\theta_0, \widehat{p})\gamma = o_P(1)$ provided $nh^3 = o(1)$ and $(nh^{3/2})^{-1}(\log n)^{3/2} = o(1)$, using what has been shown already for (B3). Next,

$$\begin{aligned} &r_n^2 B_n(\theta_0 + \frac{\gamma}{r_n}, p_0) \qquad (10)\\ &= r_n^2\Big[M_n(\theta_0 + \frac{\gamma}{r_n}, p_0) - M_n(\theta_0, p_0) - M(\theta_0 + \frac{\gamma}{r_n}, p_0) + M(\theta_0, p_0)\Big] + \frac{1}{2}\Lambda(\theta_0, p_0)\gamma^2 + o(1), \end{aligned}$$

since $\Gamma(\theta_0, p_0) = 0$. Hence, $\Lambda(\gamma) = \frac{1}{2}\Lambda(\theta_0, p_0)\gamma^2$. The first terms on the right hand side of (10) are exactly the same as in the parametric case. Hence we can follow the same steps

and get the convergence of the marginals using Lindeberg condition and Remark 3 (vii) under some regularity assumptions on $f_{X_2|X_1}$ and $\theta \mapsto f_\theta$.Finally, condition (C10) can be proved in a similar way as (B2). The asymptotic distribution of $r_n(\widehat{\theta} - \theta_0)$ now follows from Theorem 3.

To illustrate the finite sample behavior of the estimator $\widehat{\theta}$ defined in (8), we carry out a small simulation study. Let $X_1 \sim U[0,1]$ and consider the model

$$X_2 = \max(\min(U + \varepsilon, 1), 0),$$

where $U \sim U[0,1]$, $\varepsilon \sim U[-r, r]$ for some small $r > 0$, and U, ε and X_1 are independent. Let $Y = I(U \geq f_\theta(X_1))$, where $f_\theta(x_1) = \theta x_1$ for some θ. Define

$$p(x_1) = P(\Delta = 1 | X_1 = x_1) = \alpha_0 + \alpha_1 (x_1 - 0.5)^2.$$

The data consist of $(X_{i1}, X_{i2}, Y_i\Delta_i, \Delta_i)$ $(i = 1, \ldots, n)$ from the above model. For the bandwidth we work with $h = c_h n^{-1/2}$, since this bandwidth satisfies the regularity conditions coming from the asymptotic theory. In Table 1 we show the bias and root mean squared error (RMSE) of the estimator $\widehat{\theta}$ for several values of n (150 and 300), r (0.1 and 0.2), α_0 (0.25, 0.50 and 0.75), and c_h (2, 3.5 and 5). The other parameters are set to $\theta = 1$ and $\alpha_1 = 1$. The results are based on 1000 Monte Carlo runs. The table shows that both the bias and the RMSE are quite small, and the results improve when n increases, α_0 increases or r decreases. Also, the table clearly shows that the estimation of θ is not very sensitive to the choice of the bandwidth h.

6.2 Partially linear binary choice model

Consider the following binary choice model with partially linear regression structure :

$$\begin{aligned} U &= X^T\beta + g(Z) - \varepsilon \\ Y &= I(U \geq 0), \end{aligned}$$

where the median of ε given X and Z is zero, X is of dimension $d \geq 1$, and Z is one-dimensional. The observations consist of the i.i.d. triplets (X_i, Z_i, Y_i), $i = 1, \ldots, n$, with the same distribution as (X, Z, Y). In the absence of the nonparametric function $g(\cdot)$, a semiparametric estimator of β, called the maximum score estimator, has been proposed by Manski (1975). The consistency of this estimator was proved by Manski (1985), while

			$c_h = 2.0$		$c_h = 3.5$		$c_h = 5.0$	
n	r	α_0	Bias	RMSE	Bias	RMSE	Bias	RMSE
150	0.1	0.25	.0346	.1321	.0375	.1320	.0385	.1310
		0.50	.0240	.0931	.0251	.0930	.0269	.0947
		0.75	.0218	.0741	.0230	.0733	.0222	.0734
	0.2	0.25	.0521	.1802	.0542	.1809	.0575	.1791
		0.50	.0539	.1452	.0549	.1445	.0566	.1453
		0.75	.0541	.1220	.0544	.1234	.0545	.1235
300	0.1	0.25	.0281	.0921	.0291	.0902	.0277	.0913
		0.50	.0208	.0651	.0219	.0658	.0217	.0658
		0.75	.0179	.0510	.0164	.0496	.0178	.0504
	0.2	0.25	.0525	.1403	.0554	.1397	.0588	.1383
		0.50	.0484	.1057	.0508	.1073	.0535	.1067
		0.75	.0491	.0925	.0506	.0931	.0506	.0936

Table 1: *Bias and root mean squared error (RMSE) of the estimator* $\widehat{\theta}$ *for several values of* n, r, α_0 *and* c_h.

Kim and Pollard (1990) showed that the estimator converges at $n^{1/3}$-rate, and Abrevaya and Huang (2005) proved the consistency of a certain bootstrap procedure.

For a given vector β we estimate the function g by means of an M-estimator of kernel type : $\widehat{g}_\beta(z) = \text{argmax}_a\, S_{n,\beta}(a|z)$, where

$$S_{n,\beta}(a|z) = (nh)^{-1} \sum_{i=1}^{n} \Big[2I(Y_i = 1) - 1\Big] I(X_i^T \beta + a \geq 0) k\Big(\frac{Z_i - z}{h}\Big),$$

where $h = h_n$ is a bandwidth sequence and k is a kernel function. This leads to the following estimator of β :

$$\widehat{\beta} = \text{argmax}_\beta\, M_n(\beta, \widehat{g}_\beta),$$

where

$$M_n(\beta, g) = n^{-1} \sum_{i=1}^{n} \Big[2I(Y_i = 1) - 1\Big] I(X_i^T \beta + g(Z_i) \geq 0).$$

It is clear that this criterion function can not be differentiated with respect to β, and hence it is an example of a situation which the existing theory on semiparametric estimation fails to cover, whereas the theory developed in this paper can be applied. We expect

that the estimator $\widehat{\beta}$ has a cube-root n convergence rate. We leave the verification of the conditions of Theorems 1–3 to the reader.

6.3 Hit rate model

Suppose we like to estimate

$$\theta = P\big(h(X^T\beta) \in A(Z)\big),$$

where $A(Z)$ is a random set, and β and h are unknown finite and infinite dimensional parameters. Applications where the estimation of θ is of interest can be found e.g. in Bliss (1997), who is interested in evaluating nonparametric yield curve fits.

Suppose that h is the density of $X^T\beta$. For a random sample (X_i, Z_i), $i = 1, \ldots, n$, define $\widehat{h}_\beta(z) = n^{-1}\sum_{i=1}^n k_h(X_i^T\beta - z)$, which is the classical kernel density estimator applied to the linear combination $X_i^T\beta$ $(i = 1, \ldots, n)$. Then, a natural estimator of β and θ is

$$(\widehat{\beta}, \widehat{\theta}) = \text{argmax}_{\beta,\theta}\, n^{-1}\sum_{i=1}^n \Big[I\big(\widehat{h}_\beta(X_i^T\beta) \in A(Z_i)\big) - \theta\Big]^2.$$

When $X^T\beta$ is replaced by X, this example has been studied by Chen, Linton and Van Keilegom (2003) using semiparametric theory for Z-estimators. However, the introduction of the β-vector makes the criterion function non-smooth, and hence the M-estimation problem can no longer be reduced to a Z-estimation problem.

6.4 Threshold (or change point) model

A popular model in statistics and econometrics is the following semiparametric threshold model :

$$Y = g_1(X)I(Z^T\beta \leq 0) + g_2(X)I(Z^T\beta > 0) + \varepsilon,$$

where X is one-dimensional, Z is possibly of higher dimension, Z may or may not contain X, $E(\varepsilon|X, Z) = 0$, $\text{Var}(\varepsilon|X, Z) < \infty$, and g_1 and g_2 are unknown but supposed to be smooth on the interior of their support. Given i.i.d. data $(X_1, Z_1, Y_1), \ldots, (X_n, Z_n, Y_n)$ from the above model, define

$$\widehat{g}_{1,\beta}(x) = \sum_{i=1}^n \frac{k_h(X_i - x)I(Z_i^T\beta \leq 0)}{\sum_{j=1}^n k_h(X_j - x)I(Z_j^T\beta \leq 0)} Y_i$$

and

$$\widehat{g}_{2,\beta}(x) = \sum_{i=1}^n \frac{k_h(X_i - x)I(Z_i^T\beta > 0)}{\sum_{j=1}^n k_h(X_j - x)I(Z_j^T\beta > 0)} Y_i,$$

where $k_h(\cdot) = k(\cdot/h)/h$. Then, the idea is to estimate the vector β by looking for the linear combination $Z^T\beta$ which leads to the largest difference between $\widehat{g}_{1,\beta}$ and $\widehat{g}_{2,\beta}$:

$$\widehat{\beta} = \text{argmax}_\beta \, n^{-1} \sum_{i=1}^{n} \Big(\widehat{g}_{1,\beta}(X_i) - \widehat{g}_{2,\beta}(X_i) \Big)^2 .$$

As in the previous examples, the non-differentiability of the criterion function with respect to β makes this an example of a case where our theory applies, contrary to existing theory.

6.5 Single index model with monotone link function

Another example where our theory can be applied comes from the context of single index regression models (see Ichimura, 1993) :

$$Y = g(X^T\beta) + \varepsilon,$$

where $E(\varepsilon|X) = 0$, $\text{Var}(\varepsilon|X) < \infty$ and we suppose that g is unknown but monotone. Given i.i.d. data $(X_1, Y_1), \ldots, (X_n, Y_n)$ from the above model, an estimator of the function g can be obtained by using e.g. the pool-adjacent-violators algorithm, leading to a non-smooth estimator $\widehat{g}_\beta$ of $g_\beta(z) = E[Y|X^T\beta = z]$. Next, an estimator of β can be found by applying the least-squares estimation method :

$$\widehat{\beta} = \text{argmax}_\beta \Big[- n^{-1} \sum_{i=1}^{n} \big(Y_i - \widehat{g}_\beta(X_i^T\beta) \big)^2 \Big].$$

Due to the non-smooth nature of $\widehat{g}_\beta(\cdot)$, this criterion function is not smooth in β. Hence, this is another example of a situation where the theory of this paper can help out.

6.6 Binary choice model with missing data

Reconsider a binary choice model as in Subsection 6.2, but suppose now that the regression function is linear :

$$\begin{aligned} U &= X^T\beta - \varepsilon \\ Y &= I(U \geq 0), \end{aligned}$$

where as before we suppose that ε has median zero, conditional on X. We suppose that Y is subject to the missing at random (MAR) mechanism, and the probability of observing Y depends on the value of X through a certain linear combination of the X-components :

$$P(\Delta = 1|X, Y) = P(\Delta = 1|X^T\gamma) := p(X^T\gamma),$$

where $\Delta = 1$ if Y is observed and 0 if it is missing. The data now consist of i.i.d. triplets $(X_i, Y_i\Delta_i, \Delta_i)$, $i = 1, \ldots, n$, from the above model. For estimating $p_\gamma(z) = P(\Delta = 1|X^T\gamma = z)$, we use a kernel estimator :

$$\widehat{p}_\gamma(z) = \sum_{i=1}^{n} \frac{k_h(X_i^T\gamma - z)}{\sum_{j=1}^{n} k_h(X_j^T\gamma - z)} I(\Delta_i = 1).$$

Next, let $(\widehat{\beta}, \widehat{\gamma}) = \text{argmax}_{\beta,\gamma}\, M_n(\beta, \gamma, \widehat{p}_\gamma)$, where

$$M_n(\beta, \gamma, p) = n^{-1} \sum_{i=1}^{n} \frac{I(\Delta_i = 1)}{p(X_i^T\gamma)} \Big[2I(Y_i = 1) - 1\Big] I(X_i^T\beta \geq 0).$$

It is clear that $M_n(\beta, \gamma, p)$ is smooth in γ but non-smooth in β, and hence existing theory cannot be applied here. As before, Theorems 1–3 can be applied to this criterion function to find the limiting distribution of $\widehat{\beta}$ and $\widehat{\theta}$.

More generally, from the above example it is clear that any parametric M-estimation problem with a non-smooth criterion function in which the response is missing at random can be turned into a semiparametric M-estimation problem by introducing the above propensity function. Hence, the theory of this paper applies to many more examples than the illustrative examples given here.

Appendix: Proofs

In this Appendix we give the proofs of the asymptotic results, namely we prove the consistency, the rate of convergence and the asymptotic distribution of our M-estimator $\widehat{\theta}$.

Proof of Theorem 1. Our aim is to show that

$$M(\theta_0, h_0) - M(\widehat{\theta}, h_0) = o_{P^*}(1). \tag{1}$$

Indeed, the result we want to obtain is a direct consequence of (1) and assumption (A2). It is easy to show that assumptions (A3) and (A4) imply that

$$\frac{|M_n(\widehat{\theta}, \widehat{h}) - M_n(\theta_0, \widehat{h}) - M(\widehat{\theta}, \widehat{h}) + M(\theta_0, \widehat{h})|}{1 + |M_n(\widehat{\theta}, \widehat{h}) - M_n(\theta_0, \widehat{h})| + |M(\widehat{\theta}, \widehat{h}) - M(\theta_0, \widehat{h})|} = o_{P^*}(1), \tag{2}$$

since $\widehat{\theta}$ belongs by construction to Θ. Consider the following decomposition:

$$\begin{aligned}
&M(\theta_0, h_0) - M(\widehat{\theta}, h_0) \\
&= M(\widehat{\theta}, \widehat{h}) - M(\widehat{\theta}, h_0) + M(\theta_0, h_0) - M(\theta_0, \widehat{h}) + M(\theta_0, \widehat{h}) - M(\widehat{\theta}, \widehat{h}) \\
&\leq M_n(\theta_0, \widehat{h}) - M_n(\widehat{\theta}, \widehat{h}) + 2 \sup_{\theta \in \Theta} |M(\theta, h_0) - M(\theta, \widehat{h})| \\
&\quad + |M_n(\widehat{\theta}, \widehat{h}) - M(\widehat{\theta}, \widehat{h}) - M_n(\theta_0, \widehat{h}) + M(\theta_0, \widehat{h})|.
\end{aligned}$$

This, together with (2) leads to the following inequality:

$$
\begin{aligned}
&(M(\theta_0, h_0) - M(\widehat{\theta}, h_0))(1 + o_{P^*}(1)) \\
&\leq (M_n(\theta_0, \widehat{h}) - M_n(\widehat{\theta}, \widehat{h}))(1 + o_{P^*}(1)) + 4 \sup_{\theta \in \Theta} |M(\theta, h_0) - M(\theta, \widehat{h})| + o_{P^*}(1).
\end{aligned}
$$

Now, the quantity $(1 + o_{P^*}(1))$ on the left hand side in the above inequality is positive on a set A_n whose outer probability tends to one when n tends to infinity. On A_n, a reformulation of the previous inequality gives:

$$
\begin{aligned}
&M(\theta_0, h_0) - M(\widehat{\theta}, h_0) \qquad (3) \\
&\leq (M_n(\theta_0, \widehat{h}) - M_n(\widehat{\theta}, \widehat{h}))(1 + o_{P^*}(1)) + 4 \sup_{\theta \in \Theta} |M(\theta, h_0) - M(\theta, \widehat{h})|(1 + o_{P^*}(1)) + o_{P^*}(1).
\end{aligned}
$$

Assumptions (A3) and (A5) imply that

$$
\sup_{\theta \in \Theta} |M(\theta, h_0) - M(\theta, \widehat{h})| = o_{P^*}(1), \qquad (4)
$$

and assumption (A1) gives that

$$
M_n(\theta_0, \widehat{h}) - M_n(\widehat{\theta}, \widehat{h}) \leq o_{P^*}(1). \qquad (5)
$$

It now follows directly from (3)-(5) that

$$
0 \leq M(\theta_0, h_0) - M(\widehat{\theta}, h_0) \leq o_{P^*}(1). \qquad \square
$$

Proof of Theorem 2. Let ξ_n be the $O_{P^*}(r_n^{-2})$-quantity involved in assumption (B4). We introduce the sets

$$
S_{j,n} = \Big\{ \theta : 2^{j-1} < r_n d(\theta, \theta_0) \leq 2^j \Big\},
$$

and observe that $\Theta \backslash \{\theta_0\} = \cup_{j=1}^{+\infty} S_{j,n}$. Our aim is to prove that for any $\epsilon > 0$ there exists $\tau_\epsilon > 0$ such that

$$
\mathbb{P}^*\big(r_n d(\widehat{\theta}, \theta_0) > \tau_\epsilon\big) < \epsilon \qquad (6)
$$

for n sufficiently large. From now on we work with an arbitrary fixed positive value of ϵ. For any δ, δ_1, M, K, $K' > 0$, we obtain the following bound using assumption (B4):

$$
\begin{aligned}
&\mathbb{P}^*\Big(r_n d(\widehat{\theta}, \theta_0) > 2^M\Big) \\
&\leq \sum_{j \geq M,\, 2^j \leq \delta r_n} \mathbb{P}^*\Big(\sup_{\theta \in S_{j,n}} [M_n(\theta, \widehat{h}) - M_n(\theta_0, \widehat{h})] \geq -K r_n^{-2},\, A_n \Big) \\
&\quad + \mathbb{P}^*\Big(2d(\widehat{\theta}, \theta_0) \geq \delta\Big) + \mathbb{P}^*\Big(r_n^2 |\xi_n| > K\Big) + \mathbb{P}^*\Big(r_n |W_n| > K'\Big) \\
&\quad + \mathbb{P}^*\Big(|\beta_n| > \frac{C}{2}\Big) + \mathbb{P}^*\Big(d_{\mathcal{H}}(\widehat{h}, h_0) > \frac{\delta_1}{v_n}\Big), \qquad (7)
\end{aligned}
$$

where $A_n = \{r_n|W_n| \leq K',\ |\beta_n| \leq \frac{C}{2},\ d_{\mathcal{H}}(\widehat{h}, h_0) \leq \frac{\delta_1}{v_n}\}$. Indeed, we can write

$$\begin{aligned}
&\mathbb{P}^*\Big(r_n d(\widehat{\theta}, \theta_0) > 2^M,\ 2d(\widehat{\theta}, \theta_0) < \delta,\ r_n^2|\xi_n| \leq K,\ A_n\Big) \\
&\leq \sum_{j \geq M, 2^j \leq \delta r_n} \mathbb{P}^*\Big(\widehat{\theta} \in S_{j,n},\ r_n^2|\xi_n| \leq K,\ A_n\Big) \\
&\leq \sum_{j \geq M, 2^j \leq \delta r_n} \mathbb{P}^*\Big(\sup_{\theta \in S_{j,n}} [M_n(\theta, \widehat{h}) - M_n(\theta_0, \widehat{h})] \geq \xi_n,\ r_n^2|\xi_n| \leq K,\ A_n\Big) \\
&\leq \sum_{j \geq M, 2^j \leq \delta r_n} \mathbb{P}^*\Big(\sup_{\theta \in S_{j,n}} [M_n(\theta, \widehat{h}) - M_n(\theta_0, \widehat{h})] \geq -Kr_n^2,\ A_n\Big).
\end{aligned}$$

Assumption (B1) implies that for all $\delta > 0$ there exists n_ϵ such that

$$\mathbb{P}^*(2d(\widehat{\theta}, \theta_0) \geq \delta) < \frac{\epsilon}{6} \tag{8}$$

for n larger than n_ϵ. Then, by definition of ξ_n and W_n and because of (B1), there exist three positive constants δ_1, K_ϵ and K'_ϵ such that

$$\begin{aligned}
&\mathbb{P}^*\Big(r_n^2|\xi_n| > K_\epsilon\Big) < \frac{\epsilon}{6}, \qquad \mathbb{P}^*\Big(r_n|W_n| > K'_\epsilon\Big) < \frac{\epsilon}{6}, \\
&\mathbb{P}^*\Big(|\beta_n| > \frac{C}{2}\Big) < \frac{\epsilon}{6}, \quad \text{and} \quad \mathbb{P}^*\Big(d_{\mathcal{H}}(\widehat{h}, h_0) > \frac{\delta_1}{v_n}\Big) < \frac{\epsilon}{6}
\end{aligned} \tag{9}$$

for n larger than some $n_1 \in \mathbb{N}$. We fix $\delta < \delta_0$ and suppose that $n \geq \max(n_0, n_1, n_\epsilon)$ to get that assumptions (B2) and (B3) are fulfilled on all $S_{j,n}$ such that $2^j \leq \delta r_n$.

Now, it follows directly from assumption (B3) that for each fixed j such that $2^j \leq \delta r_n$ one has for all $\theta \in S_{n,j}$:

$$\begin{aligned}
&M_n(\theta, \widehat{h}) - M_n(\theta_0, \widehat{h}) \\
&\leq M(\theta, \widehat{h}) - M(\theta_0, \widehat{h}) + \sup_{d(\theta, \theta_0) \leq \frac{2^j}{r_n}} |M_n(\theta, \widehat{h}) - M_n(\theta_0, \widehat{h}) - M(\theta, \widehat{h}) + M(\theta_0, \widehat{h})| \\
&\leq |W_n|\frac{2^j}{r_n} - (C - \beta_n)\frac{2^{2j-2}}{r_n^2} + \sup_{d(\theta, \theta_0) \leq \frac{2^j}{r_n}} |M_n(\theta, \widehat{h}) - M_n(\theta_0, \widehat{h}) - M(\theta, \widehat{h}) + M(\theta_0, \widehat{h})|.
\end{aligned}$$

Consequently, we obtain the following inequality:

$$\begin{aligned}
&\mathbb{P}^*\Big(\sup_{\theta \in S_{j,n}} [M_n(\theta, \widehat{h}) - M_n(\theta_0, \widehat{h})] \geq -K_\epsilon r_n^{-2},\ A_n\Big) \\
&\leq \mathbb{P}^*\Big(\sup_{d(\theta, \theta_0) \leq \frac{2^j}{r_n},\, d_{\mathcal{H}}(\widehat{h}, h_0) \leq \frac{\delta_1}{v_n}} |M_n(\theta, h) - M_n(\theta_0, h) - M(\theta, h) + M(\theta_0, h)| \\
&\qquad \geq \frac{2^{2j-2}}{r_n^2}\big(\frac{C}{2} - K'_\epsilon 2^{2-j} - K_\epsilon 2^{2-2j}\big)\Big).
\end{aligned}$$

Now, there exists M_ϵ such that for all $j \geq M_\epsilon$ one gets

$$\frac{C}{2} - K'_\epsilon 2^{2-j} - K_\epsilon 2^{2-2j} \geq \frac{C}{4}.$$

Consequently, if $M \geq M_\epsilon$, using assumption (B2) and Chebychev's inequality we have that

$$\begin{aligned}
&\sum_{j\geq M,\, 2^j \leq \delta r_n} \mathbb{P}^*\Big(\Big\{ \sup_{\theta \in S_{j,n}} [M_n(\theta, \widehat{h}) - M_n(\theta_0, \widehat{h})] \geq -K_\epsilon r_n^{-2} \Big\} \cap A_n\Big) \\
&\leq \sum_{j\geq M,\, 2^j \leq \delta r_n} \mathbb{P}^*\Big(\sup_{d(\theta,\theta_0)\leq \frac{2^j}{r_n},\, d_{\mathcal{H}}(\widehat{h},h_0)\leq \frac{\delta_1}{v_n}} |M_n(\theta, \widehat{h}) - M_n(\theta_0, \widehat{h}) - M(\theta, \widehat{h}) + M(\theta_0, \widehat{h})| \geq \frac{C 2^{2j-2}}{4 r_n^2}\Big) \\
&\leq \frac{4K r_n^2}{C\sqrt{n}} \sum_{j\geq M,\, 2^j \leq \delta r_n} \frac{\Phi_n(\frac{2^j}{r_n})}{2^{2j-2}} \\
&\leq \frac{4K r_n^2}{C\sqrt{n}} \sum_{j\geq M,\, 2^j \leq \delta r_n} \frac{2^{j\alpha}\Phi_n(\frac{1}{r_n})}{2^{2j-2}} \\
&\leq \frac{16K}{C} \sum_{j\geq M} 2^{j(\alpha-2)}.
\end{aligned}$$

Finally, since $\alpha < 2$, the series $\sum_{j\geq M} 2^{j(\alpha-2)}$ converges and hence there exists $M'_\epsilon \geq M_\epsilon$ such that

$$\frac{16K}{C} \sum_{j\geq M'_\epsilon} 2^{j(\alpha-2)} \leq \frac{\epsilon}{6}.$$

This finishes the proof showing (6) with $\tau_\epsilon = 2^{M'_\epsilon}$. □

Proof of Theorem 3. The first step of the proof consists in showing the weak convergence of the process $\gamma \mapsto r_n^2 B_n(\theta_0 + \frac{\gamma}{r_n}, \widehat{h})$. This is shown in Lemma 4 (given below).

The remainder of the proof is based on somewhat similar arguments as those used to state the Argmax theorem in Van der Vaart and Wellner (1996). First note that E is a σ-compact metric space since $E = \cup_{i=1}^{\infty} \mathcal{K}_i$ with $\mathcal{K}_i = \{\gamma \in E : \|\gamma\| \leq a_i\}$ for any positive sequence $(a_i)_{i\in\mathbb{N}^*}$ tending to infinity.

Then deduce from assumption (C9) together with Lemmas 5 and 6 given below, that almost all paths of the limiting process $\gamma \mapsto \Lambda(\gamma) + \mathbb{G}(\gamma)$ attain their supremum at an unique point γ_0, following similar ideas to what is done in the parametric case (see Theorem 3.2.10 in Van der Vaart and Wellner (1996)). Assume now that γ_0 is measurable. The weak convergence of $r_n(\widehat{\theta} - \theta_0)$ to γ_0 is equivalent to the next statement (Portmanteau's theorem) :

$$\text{limsup}_{n\to\infty} \mathbb{P}^*\Big(r_n(\widehat{\theta} - \theta_0) \in C\Big) \leq \mathbb{P}\Big(\gamma_0 \in C\Big), \text{ for every closed set } C.$$

Let C be an arbitrary closed subset of E and fix $\epsilon > 0$. The random variable γ_0 is tight because it takes values in E, which is σ-compact. Combining this tightness and the first part of (C1), it is possible to find $K_\epsilon > 0$ and hence a compact set $\mathcal{K}_\epsilon := \{\gamma : \|\gamma\| \leq K_\epsilon\}$ such that

$$\mathbb{P}^*\Big(\gamma_0 \notin \mathcal{K}_\epsilon\Big) \leq \frac{\epsilon}{2}, \text{ and } \mathbb{P}^*\Big(r_n(\widehat{\theta} - \theta_0) \notin \mathcal{K}_\epsilon\Big) \leq \frac{\epsilon}{2}. \tag{10}$$

It follows easily from (10) that

$$\begin{aligned}
&\text{limsup}_{n\to\infty}\mathbb{P}^*\Big(r_n(\widehat{\theta} - \theta_0) \in C\Big) \\
&\leq \mathbb{P}^*\Big(r_n(\widehat{\theta} - \theta_0) \in C \cap \mathcal{K}_\epsilon,\, \gamma_0 \in \mathcal{K}_\epsilon\Big) + \text{limsup}_{n\to\infty}\mathbb{P}^*\Big(\{r_n(\widehat{\theta} - \theta_0) \notin \mathcal{K}_\epsilon\} \cup \{\gamma_0 \notin \mathcal{K}_\epsilon\}\Big) \\
&\leq \mathbb{P}^*\Big(r_n(\widehat{\theta} - \theta_0) \in C \cap \mathcal{K}_\epsilon,\, \gamma_0 \in \mathcal{K}_\epsilon\Big) + \epsilon.
\end{aligned} \tag{11}$$

Now using Lemma 4 and assumption (C8) we obtain

$$\begin{aligned}
&\text{limsup}_{n\to\infty}\mathbb{P}^*\Big(r_n(\widehat{\theta} - \theta_0) \in C \cap \mathcal{K}_\epsilon,\, \gamma_0 \in \mathcal{K}_\epsilon\Big) \\
&\leq \text{limsup}_{n\to\infty}\mathbb{P}^*\Big(\sup_{\gamma\in C\cap\mathcal{K}_\epsilon} r_n^2 B_n\Big(\theta_0 + \frac{\gamma}{r_n}, \widehat{h}\Big) \geq \sup_{\gamma\in\mathcal{K}_\epsilon} r_n^2 B_n\Big(\theta_0 + \frac{\gamma}{r_n}, \widehat{h}\Big) + o_{P^*}(1),\, \gamma_0 \in \mathcal{K}_\epsilon\Big) \\
&\leq \mathbb{P}^*\Big(\sup_{\gamma\in C\cap\mathcal{K}_\epsilon} (\Lambda + \mathbb{G})(\gamma) \geq \sup_{\gamma\in\mathcal{K}_\epsilon} (\Lambda + \mathbb{G})(\gamma),\, \gamma_0 \in \mathcal{K}_\epsilon\Big),
\end{aligned} \tag{12}$$

by Slutsky's lemma and Portmanteau's theorem. On the other hand, for every open set G containing γ_0, we have:

$$(\Lambda + \mathbb{G})(\gamma_0) > \sup_{\gamma\in G^c\cap\mathcal{K}_\epsilon} (\Lambda + \mathbb{G})(\gamma).$$

This together with (12) leads to

$$\text{limsup}_{n\to\infty}\mathbb{P}^*\Big(r_n(\widehat{\theta} - \theta_0) \in C \cap \mathcal{K}_\epsilon,\, \gamma_0 \in \mathcal{K}_\epsilon\Big) \leq \mathbb{P}^*\Big(\gamma_0 \in C\Big). \tag{13}$$

Consequently, it follows from (11) that for all $\epsilon > 0$,

$$\text{limsup}_{n\to\infty}\mathbb{P}^*\Big(r_n(\widehat{\theta} - \theta_0) \in C\Big) \leq \mathbb{P}^*\Big(\gamma_0 \in C\Big) + \epsilon. \tag{14}$$

Since the right hand side of (14) holds for all $\epsilon > 0$, it also holds for $\epsilon = 0$. The result now follows from Portmanteau's theorem. □

We end this section with three lemmas that were needed in the proof of Theorem 3.

Lemma 4 *For all $K > 0$, let $\mathcal{K} = \{\gamma \in E : \|\gamma\| \leq K\}$ be a compact subset of E. Then, under the assumptions of Theorem 3, for any such $\mathcal{K}$, the process $\gamma \mapsto r_n^2 B_n(\theta_0 + \frac{\gamma}{r_n}, \widehat{h})$ converges weakly to the process $\gamma \mapsto \Lambda(\gamma) + \mathbb{G}(\gamma)$ in $\ell^\infty(\mathcal{K})$. Moreover, almost all paths of the limiting process are continuous (uniformly on every compact $\mathcal{K}$) with respect to $\|\cdot\|$.*

Proof. The weak convergence of the process $\gamma \mapsto r_n^2 B_n(\theta_0 + \frac{\gamma}{r_n}, \widehat{h})$ in $\ell^\infty(\mathcal{K})$ follows directly from Slutsky's theorem and Lemmas 5 and 6. On the other hand, $\|\cdot\|$ makes $\mathcal{K}$ totally bounded (since it is compact) and $\gamma \mapsto r_n^2 B_n(\theta_0 + \frac{\gamma}{r_n}, h_0) + r_n W_n(\gamma)$ is asymptotically uniformly $\|\cdot\|$-equicontinuous in probability, asymptotically tight, and it converges weakly to $\gamma \mapsto \Lambda(\gamma) + \mathbb{G}(\gamma)$ in $\ell^\infty(\mathcal{K})$ (see proof of Lemma 6). Thus almost all paths of the limiting process are uniformly $\|\cdot\|$-continuous on $\mathcal{K}$ (see Theorem 1.5.7 in Van der Vaart and Wellner (1996)). Moreover, because E may be covered by a countable sequence of such compact sets, almost all paths of the limiting process are $\|\cdot\|$-continuous on E. □

Lemma 5 *Let $\mathcal{K} = \{\gamma \in E : \|\gamma\| \leq K\}$. Then, under the assumptions of Theorem 3, for all $\gamma \in \mathcal{K}$, there exist $\xi_{0,n}, \xi_{1,n}, \xi_{2,n}$, such that $\sup_{\gamma\in\mathcal{K}} |\xi_{j,n}| = o_{P^*}(1), j = 0, 1, 2,$ and*

$$r_n^2 B_n\Big(\theta_0 + \frac{\gamma}{r_n}, \widehat{h}\Big)(1 + \xi_{0,n}) = \Big[r_n^2 B_n\Big(\theta_0 + \frac{\gamma}{r_n}, h_0\Big) + r_n W_n(\gamma)\Big](1 + \xi_{1,n}) + \xi_{2,n}.$$

Proof. Let us introduce the following notations :

$$\begin{aligned}
\alpha_{0,n}(\gamma) &= \frac{B_n(\theta, h) - B(\theta, h) - B_n(\theta, h_0) + B(\theta, h_0)}{r_n^{-2} + |B_n(\theta, h)| + |B_n(\theta, h_0)| + |B(\theta, h)| + |B(\theta, h_0)|},\\
s_{n,h}(\gamma) &= \operatorname{sign}\Big[B_n\Big(\theta_0 + \frac{\gamma}{r_n}, h\Big)\Big],\\
s_h(\gamma) &= \operatorname{sign}\Big[B\Big(\theta_0 + \frac{\gamma}{r_n}, h\Big)\Big],
\end{aligned}$$

with $\theta = \theta_0 + \gamma/r_n$.

Because the compact $\mathcal{K}$ is bounded and θ_0 belongs to the interior of Θ, there exists $n_{\mathcal{K}}$ such that for all $n \geq n_{\mathcal{K}}$ and for all $\gamma \in \mathcal{K}$, the quantity $\theta_0 + \frac{\gamma}{r_n}$ is in Θ. Then, for all $\gamma \in \mathcal{K}$ entails that

$$\begin{aligned}
&B_n\Big(\theta_0 + \frac{\gamma}{r_n}, \widehat{h}\Big)\\
&= B_n\Big(\theta_0 + \frac{\gamma}{r_n}, h_0\Big) + B\Big(\theta_0 + \frac{\gamma}{r_n}, \widehat{h}\Big) - B\Big(\theta_0 + \frac{\gamma}{r_n}, h_0\Big)\\
&\quad + \alpha_{0,n}(\gamma)\Big(r_n^{-2} + \Big|B_n\Big(\theta_0 + \frac{\gamma}{r_n}, \widehat{h}\Big)\Big| + \Big|B_n\Big(\theta_0 + \frac{\gamma}{r_n}, h_0\Big)\Big|\\
&\qquad + \Big|B\Big(\theta_0 + \frac{\gamma}{r_n}, \widehat{h}\Big)\Big| + \Big|B\Big(\theta_0 + \frac{\gamma}{r_n}, h_0\Big)\Big|\Big).
\end{aligned} \tag{15}$$

This can be reformulated as

$$\begin{aligned}
&r_n^2 B_n\Big(\theta_0 + \frac{\gamma}{r_n}, \widehat{h}\Big)\Big(1 - \alpha_{0,n}(\gamma) s_{n,\widehat{h}}(\gamma)\Big)\\
&= r_n^2 B_n\Big(\theta_0 + \frac{\gamma}{r_n}, h_0\Big)\Big(1 + \alpha_{0,n}(\gamma) s_{n,h_0}(\gamma)\Big) + r_n^2 B\Big(\theta_0 + \frac{\gamma}{r_n}, \widehat{h}\Big)\Big(1 + \alpha_{0,n}(\gamma) s_{\widehat{h}}(\gamma)\Big)\\
&\quad - r_n^2 B\Big(\theta_0 + \frac{\gamma}{r_n}, h_0\Big)\Big(1 - \alpha_{0,n}(\gamma) s_{h_0}(\gamma)\Big) + \alpha_{0,n}(\gamma).
\end{aligned} \tag{16}$$

Then use assumptions (C1) and (C7) to get

$$r_n^2\Big[B\Big(\theta_0+\frac{\gamma}{r_n},\widehat{h}\Big)-B\Big(\theta_0+\frac{\gamma}{r_n},h_0\Big)\Big] = r_nW_n(\gamma)+\beta_n\|\gamma\|^2+r_n^2o\Big(\frac{\|\gamma\|^2}{r_n^2}\Big) := r_nW_n(\gamma)+\alpha_{1,n}(\gamma). \quad (17)$$

Combining (16) and (17) we obtain

$$\begin{aligned} &r_n^2B_n\Big(\theta_0+\frac{\gamma}{r_n},\widehat{h}\Big)(1+\xi_{0,n}(\gamma)) \\ &= \Big[r_n^2B_n\Big(\theta_0+\frac{\gamma}{r_n},h_0\Big)+r_nW_n(\gamma)\Big](1+\xi_{1,n}(\gamma))+\xi_{2,n}(\gamma), \end{aligned} \quad (18)$$

with

$$\begin{aligned} \xi_{0,n}(\gamma) &= -\alpha_{0,n}(\gamma)s_{n,\widehat{h}}(\gamma), \\ \xi_{1,n}(\gamma) &= \alpha_{0,n}(\gamma)s_{n,h_0}(\gamma), \\ \xi_{2,n}(\gamma) &= \alpha_{0,n}(\gamma)\Big[1+\Big(V(\gamma,\gamma)+r_n^2o\Big(\frac{\|\gamma\|^2}{r_n^2}\Big)\Big)(s_{\widehat{h}}+s_{h_0})(\gamma) \\ &\quad +\Big(r_nW_n(\gamma)+\alpha_{1,n}(\gamma)\Big)(s_{\widehat{h}}-s_{n,h_0})(\gamma)\Big]+\alpha_{1,n}(\gamma)(1+\xi_{1,n}(\gamma)). \end{aligned}$$

It can be easily shown that $\sup_{\gamma\in\mathcal{K}}|\xi_{j,n}(\gamma)| = o_{P^*}(1)$ for $j=0,1,2$ using assumptions (C3) and (C7). □

Lemma 6 *Let $\mathcal{K}=\{\gamma\in E : \|\gamma\|\leq K\}$. Then, under the assumptions of Theorem 3, the process $\gamma\mapsto r_n^2B_n(\theta_0+\frac{\gamma}{r_n},h_0)+r_nW_n(\gamma)$ is asymptotically tight, asymptotically uniformly equicontinuous with respect to $\|\cdot\|$ on $\mathcal{K}$, and it converges weakly to the process $\gamma\mapsto\Lambda(\gamma)+\mathbb{G}(\gamma)$ in $\ell^\infty(\mathcal{K})$.*

Proof. The main idea of this proof consists in writing the process $T_n:\gamma\mapsto r_n^2B_n(\theta_0+\frac{\gamma}{r_n},h_0)+r_nW_n(\gamma)$ as the sum of two processes $T_{1,n}:\gamma\mapsto r_n^2(B_n(\theta_0+\frac{\gamma}{r_n},h_0)-B(\theta_0+\frac{\gamma}{r_n},h_0))$ and $T_{2,n}:\gamma\mapsto r_n^2B(\theta_0+\frac{\gamma}{r_n},h_0)+r_nW_n(\gamma)$ and studying separately the properties of $T_{1,n}$ and $T_{2,n}$. However, in some specific cases it could be possible to state the weak convergence of T_n without this decomposition. Let us first note that assumption (C7) implies that for n sufficiently large (only depending on $\mathcal{K}$) so that $\theta_0+\frac{\mathcal{K}}{r_n}\subset\Theta$, the processes $T_{1,n}$ and $T_{2,n}$ take values in $\ell^\infty(\mathcal{K})$.

The process $T_{1,n}$ does not depend on the estimation of the nuisance parameter. Hence, following similar ideas as in the parametric case we get from assumptions (C4), (C5) and (C10) the asymptotic uniform equicontinuity of $T_{1,n}$ with respect to $\|\cdot\|$ on $\mathcal{K}$ (as a sub-product of the proof of Theorem 2.11.9 in Van der Vaart and Wellner (1996)). On the

other hand, for n large enough, $\theta + \frac{\gamma}{r_n} \in \Theta$ (see the proof of Lemma 5). Assume now that n is large enough and use assumption (C7) to conclude that for all $0 < \delta \leq \delta_1$,

$$\begin{aligned}
&\sup_{\gamma,\gamma' \in \mathcal{K}, \|\gamma - \gamma'\| \leq \delta} |T_{2,n}(\gamma) - T_{2,n}(\gamma')| \\
&= \sup_{\gamma,\gamma' \in \mathcal{K}, \|\gamma - \gamma'\| \leq \delta} \Big| W_n(\gamma - \gamma') + V(\gamma, \gamma) - V(\gamma', \gamma') + r_n^2 \Big(o\Big(\frac{\|\gamma\|^2}{r_n^2} \Big) + o\Big(\frac{\|\gamma'\|^2}{r_n^2} \Big) \Big) \Big| \\
&\leq \delta^\tau \Big(r_n \sup_{\gamma \in E, \, \delta \leq \delta_1, \, \|\gamma\| \leq \delta} \Big| \frac{W_n(\gamma)}{\delta^\tau} \Big| + \sup_{\gamma,\gamma' \in E, \, \delta \leq \delta_1, \, \|\gamma - \gamma'\| \leq \delta} \frac{|V(\gamma, \gamma) - V(\gamma', \gamma')|}{\delta^\tau} \Big) + b_n \\
&:= \delta^\tau \alpha_n + b_n, \qquad (19)
\end{aligned}$$

where $b_n \leq \sup_{\gamma,\gamma' \in \mathcal{K}} |r_n^2 (o(\frac{\|\gamma\|^2}{r_n^2}) + o(\frac{\|\gamma'\|^2}{r_n^2}))| \to 0$ as n tends to infinity, and $\alpha_n = O_{P^*}(1)$ uniformly over $\delta \leq \delta_1$. Let ϵ and η be arbitrary positive constants. It is clear that, for any $0 < \delta \leq \delta_1$ and any positive constant K, (19) leads to

$$\begin{aligned}
&\text{limsup}_{n \to \infty} \mathbb{P}^* \Big(\sup_{\gamma,\gamma' \in \mathcal{K}, \|\gamma - \gamma'\| \leq \delta} |T_{2,n}(\gamma) - T_{2,n}(\gamma')| > \epsilon \Big) \\
&\leq \text{limsup}_{n \to \infty} \mathbb{P}^* \Big(\delta^\tau \alpha_n + b_n > \epsilon, \alpha_n \leq K, |b_n| < \frac{\epsilon}{2} \Big) + \text{limsup}_{n \to \infty} \mathbb{P}^* \Big(\alpha_n > K \Big) \\
&\leq \text{limsup}_{n \to \infty} \mathbb{P}^* \Big(\delta^\tau > \frac{\epsilon}{2K} \Big) + \text{limsup}_{n \to \infty} \mathbb{P}^* \Big(\alpha_n > K \Big).
\end{aligned}$$

Finally choose K_η such that the last term is smaller than η, and take $\delta \leq \delta_1 \wedge (\frac{\epsilon}{2K_\eta})^{\frac{1}{\tau}}$. It then follows that $T_{2,n}$ is asymptotically uniformly equicontinuous in probability with respect to $\|\cdot\|$ on $\mathcal{K}$.

Hence, the same is also true for the process T_n, since it is the sum of two such processes. The asymptotic tightness and hence the weak convergence of T_n to $\Lambda + \mathbb{G}$ in $\ell^\infty(\mathcal{K})$ now follows from Theorems 1.5.7 and 1.5.4 in Van der Vaart and Wellner (1996), together with assumption (C9) and the fact that $\mathcal{K}$ is totally bounded with respect to the $\|\cdot\|$-norm (since it is compact). Moreover, using Addendum 1.5.8 in the same book, almost all paths of the limiting process on $\mathcal{K}$ are uniformly continuous with respect to $\|\cdot\|$. □

Acknowledgments. The authors would like to thank Xiaohong Chen, Guang Cheng, Oliver Linton, Michael Kosorok, Bin Nan, Bodhi Sen and Jon Wellner for stimulating discussions and helpful comments that improved the quality of the paper.

References

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

Bliss, R. (1997). Testing term structure estimation methods. *Advances in Futures and Options Research*, **9**, 197-231.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. In: Heckman, J.J. and Leamer, E.E. (Eds.), *Handbook of Econometrics*, 6B, Chapter 76, Elsevier.

Chen, X. and Fan, Y. (2006). Estimation of copula-based semiparametric time series models. *Journal of Econometrics*, **130**, 307-335.

Chen, X. and Liao, Z. (2012). On limiting distributions of sieve M-estimators of irregular functionals. Unpublished manuscript.

Chen, X., Linton, O. and Van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, **71**, 1591-1608.

Chen X. and Pouzo, D. (2009). Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *Journal of Econometrics*, **152**, 46-60.

Cheng, G. and Shang, Z. (2015). Joint asymptotics for semi-nonparametric regression models under partially linear structure. *Annals of Statistics* (to appear).

Ding, Y. and Nan, B. (2011). A sieve M-theorem for bundled parameters in semiparametric models, with application to the efficient estimation in a linear model for censored data. *Annals of Statistics*, **39**, 3032-3061.

Escanciano, J., Jacho-Chavez, D. and Lewbel, A. (2012). Identification and estimation of semiparametric two step models (under revision for *Quantitative Economics*).

Escanciano, J., Jacho-Chavez, D. and Lewbel, A. (2014). Uniform convergence of weighted sums of non- and semi-parametric residuals for estimation and testing. *Journal of Econometrics*, **178**, 426-443.

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

Groeneboom, P., Jongbloed, G. and Wellner, J.A. (2001). Estimation of a convex function: characterizations and asymptotic theory. *Annals of Statistics*, **29**, 1653-1698.

Groeneboom, P. and Wellner, J.A. (1992). *Information bounds and nonparametric maximum likelihood estimation.* Birkhäuser, Basel.

Horowitz, J. (2009). *Semiparametric and Nonparametric Methods in Econometrics.* Springer, New York.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single index models. *Journal of Econometrics*, **58**, 71-120.

Ichimura, H. and Lee, S. (2010). Characterization of the asymptotic distribution of semiparametric M-estimators. *Journal of Econometrics*, **159**, 252-266.

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

Kosorok, M.R. (2008). *Introduction to empirical processes and semiparametric inference.* Springer, New York.

Kristensen, D. and Salanié (2013). Higher order improvements of approximate estimators. Unpublished manuscript.

Linton, O., Sperlich, S. and Van Keilegom, I. (2008). Estimation of a semiparametric transformation model. *Annals of Statistics*, **36**, 686-718.

Ma, S. and Kosorok, M.R. (2005). Robust semiparametric M-estimation and the weighted bootstrap. *Journal of Multivariate Analysis*, **96**, 190-217.

Mammen, E., Rothe, C. and Schienle, M. (2011). Semiparametric estimation with generated covariates. Unpublished manuscript.

Manski, C.F. (1975). The maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics*, **3**, 205-228.

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

Mohammadi, L. and Van de Geer, S. (2005). Asymptotics in empirical risk minimization. *Journal of Machine Learning Research*, **6**, 2027-2047.

Radchenko, P. (2008). Mixed-rates asymptotics. *Annals of Statistics*, **36**, 287-309.

Van de Geer, S.A. (2000). *Empirical processes in M-estimation.* Cambridge University Press, New York.

Van der Vaart, A.W. and Wellner, J.A. (1996). *Weak convergence and empirical processes: with applications in statistics.* Springer, New York.

Van der Vaart, A.W. and Wellner, J.A. (2007). Empirical processes indexed by estimated functions. *IMS Lecture Notes-Monograph Series*, **55**, 234-252.

Xu, G., Sen, B. and Ying, Z. (2014). Bootstrapping a change-point Cox model for survival data. *Electr. J. Statist.*, **8**, 1345-1379.