Electronic Journal of Statistics
Vol. 10 (2016) 464–494
ISSN: 1935-7524
DOI: 10.1214/15-EJS1097

On the asymptotics of Z-estimators indexed by the objective functions

François Portier*

Université catholique de Louvain
e-mail: francois.portier@gmail.com

Abstract: We study the convergence of Z-estimators $\widehat{\theta}(\eta) \in \mathbb{R}^p$ for which the objective function depends on a parameter η that belongs to a Banach space $\mathcal{H}$. Our results include the uniform consistency over $\mathcal{H}$ and the weak convergence in the space of bounded $\mathbb{R}^p$-valued functions defined on $\mathcal{H}$. When η is a tuning parameter optimally selected at η_0, we provide conditions under which η_0 can be replaced by an estimated $\widehat{\eta}$ without affecting the asymptotic variance. Interestingly, these conditions are free from any rate of convergence of $\widehat{\eta}$ to η_0 but require the space described by $\widehat{\eta}$ to be not too large in terms of bracketing metric entropy. In particular, we show that Nadaraya-Watson estimators satisfy this entropy condition. We highlight several applications of our results and we study the case where η is the weight function in weighted regression.

MSC 2010 subject classifications: Primary 62F12, 62F35; secondary 62G20.
Keywords and phrases: Asymptotic theory, empirical process, semiparametric estimation, weighted regression, Z-estimation.

Received September 2015.

1. Introduction

Let P denote a probability measure defined on a measurable space $(\mathcal{Z}, \mathcal{A})$ and let $(Z_1, \ldots, Z_n)$ be independent and identically distributed random elements with law P. Given a measurable function $f : \mathcal{Z} \to \mathbb{R}$, we define

$$Pf = \int f dP, \qquad \mathbb{P}_n f = n^{-1} \sum_{i=1}^{n} f(Z_i), \qquad \mathbb{G}_n f = n^{1/2}(\mathbb{P}_n - P)f,$$

where $\mathbb{G}_n$ is called the empirical process. We consider the estimation of a Euclidean parameter $\theta_0 \in \Theta \subset \mathbb{R}^p$, when a collection of estimators, $\{\widehat{\theta}(\eta) : \eta \in \mathcal{H}\}$, is available. The index space $(\mathcal{H}, \|\cdot\|)$ is a Banach space. Suppose there exists $\eta_0 \in \mathcal{H}$ such that $\widehat{\theta}(\eta_0)$ is optimal, in some sense, within the collection. Typically, $\widehat{\theta}(\eta_0)$ might have the smallest asymptotic variance among the estimators of the collection. Such a situation arises in many fields of the statistics. For instance, η can be the cut-off parameter in Huber *robust regression*, or η might as well be

*This work has been supported by Fonds de la Recherche Scientifique (FNRS) A4/5 FC 2779/2014-2017 No. 22342320.

the weight function in *weighted least squares* (see equation (5) below and the next section for more details and examples). Unfortunately, η_0 is generally unknown since it depends on the distribution P. Usually, one is restricted to first estimate η_0 by, say, $\widehat{\eta}$ and then compute the estimator $\widehat{\theta}(\widehat{\eta})$, which should result in a not too bad approximation of θ_0. It turns out that, in many situations,

$$n^{1/2}(\widehat{\theta}(\widehat{\eta}) - \theta_0) \quad \text{has the same asymptotic law as} \quad n^{1/2}(\widehat{\theta}(\eta_0) - \theta_0), \tag{1}$$

meaning that, not only the rate of convergence but also the asymptotic variance are the same (see for instance Newey and McFadden (1994), page 2164, the reference therein, and van der Vaart (1998), page 61). This is all the more surprising since the accuracy of $\widehat{\eta}$ estimating η_0 does not matter provided its consistency.

A paradigm that encompasses the previous facts can be developed *via* the stochastic equicontinuity of the underlying empirical process $\mathbb{G}_n$ over the set of influence functions. Suppose that

$$\sup_{\eta \in \mathcal{H}} |n^{1/2}(\widehat{\theta}(\eta) - \theta_0) - \mathbb{G}_n \varphi_\eta| = o_P(1),$$

where $|\cdot|$ stands for the Euclidean norm and $\varphi_\eta : \mathcal{Z} \to \mathbb{R}^p$ is the so called influence function. It follows that, (1) holds whenever $\mathbb{G}_n(\varphi_{\widehat{\eta}} - \varphi_{\eta_0})$ goes to 0 in probability. This holds true if the process $\eta \mapsto \mathbb{G}_n(\varphi_\eta)$ is stochastically equicontinuous on $\mathcal{H}$, i.e., if for any $\epsilon > 0$,

$$\lim_{\delta \to 0} \limsup_{n \to +\infty} P\left(\sup_{\|\eta_1 - \eta_2\| < \delta} |\mathbb{G}_n(\varphi_{\eta_1} - \varphi_{\eta_2})| > \epsilon \right) = 0, \tag{2}$$

where the supremum is taken over η_1 and η_2 in $\mathcal{H}$, and if, in addition,

$$P(\widehat{\eta} \in \mathcal{H}) \xrightarrow{P} 1 \qquad \text{and} \qquad \|\widehat{\eta} - \eta_0\| \xrightarrow{P} 0. \tag{3}$$

Hence, to obtain (1), stochastic equicontinuity allows for relying on (3), a mild "no-rate" conditions on $\widehat{\eta}$. In fact, conditions (2) and (3) represent a trade-off we need to accomplish when selecting the norm $\|\cdot\|$. When one prefers to have $\|\cdot\|$ as weak as possible in order to prove (3), one needs the metric to be strong enough so that (2) can hold. Empirical process theory turns to be very useful to deal with this kind of problem. As it is summarized in van der Vaart and Wellner (1996), a natural choice for $\|\cdot\|$ is the $L_2(P)$-norm. Sufficient conditions for (2) then involve weak convergence of the empirical process $\eta \mapsto \mathbb{G}_n(\varphi_\eta)$ or, more restrictively, the metric entropy of the class of functions $\{\varphi_\eta : \eta \in \mathcal{H}\}$. Such an approach succeeded in deriving the asymptotics of specific semiparametric estimators (Akritas and Van Keilegom, 2001; van der Vaart and Wellner, 2007; Portier and Segers, 2015).

The main purpose of the paper is to establish conditions ensuring (1) holds, in the case when $\widehat{\theta}(\eta)$ is a Z-estimator for which the objective function depends

on some $\eta \in \mathcal{H}$. More formally, we consider θ_0 and $\widehat{\theta}(\eta)$ defined, respectively, as "zeros" of the maps

$$\theta \mapsto P\psi_\eta(\theta) \qquad \text{and} \qquad \theta \mapsto \mathbb{P}_n\psi_\eta(\theta), \tag{4}$$

where for each $\theta \in \Theta$ and $\eta \in \mathcal{H}$, $\psi_\eta(\theta)$ is an $\mathbb{R}^p$-valued measurable map defined on $\mathcal{Z}$. Since for every $\eta \in \mathcal{H}$, $P\psi_\eta(\theta_0) = 0$, we have several (possibly infinitely many) equations available and $\eta \in \mathcal{H}$ does not affect the limit, in probability, of the sequence $\widehat{\theta}(\eta)$. Hence η might better be understood as a tuning parameter rather than as a semiparametric nuisance parameter. In fact, the semiparametric models corresponding to (4) are subjected to an asymptotic orthogonality condition between θ and η.

In Newey (1994), semiparametric estimators are studied using pathwise derivatives along sub-models and the author underlines that, for such models, "different nonparametric estimators of the same functions should result in the same asymptotic variance" (Newey, 1994, page 1356). In Andrews (1994), the previous statement is formally demonstrated by relying on stochastic equicontinuity, as detailed in (2) and (3). In this paper, we provide new conditions on the map $(\theta, \eta) \mapsto \psi_\eta(\theta)$ and the estimators $\widehat{\eta}$ under which (1) holds. Despite considering slightly less general estimators than in Andrews (1994), our approach alleviates the regularity conditions imposed on the map $\theta \mapsto \psi_\eta(\theta)$. They are replaced by weaker regularity conditions dealing with the map $\theta \mapsto P\psi_\eta(\theta)$. In addition, the class of functions $\mathcal{H}$ is allowed to depend on n. We focus on *conditional moment restrictions models* in which η is a weight function. In this context, our approach results in a simple condition on the size of the bracketing metric entropy generated by $\widehat{\eta}$. In the case of *weighted linear regression*, when $\widehat{\eta}$ is a Nadaraya-Watson estimator, the previous condition is shown to be satisfied (see below for more details). Our result extends those of Ojeda (2008) and Portier and Segers (2015) on *local linear estimators*.

Our study is based on the weak convergence of $\{n^{1/2}(\widehat{\theta}(\eta) - \theta_0)\}_{\eta\in\mathcal{H}}$ in $\ell^\infty(\mathcal{H})^p$, the space of bounded $\mathbb{R}^p$-valued functions defined on $\mathcal{H}$. The tools we use in the proofs are reminiscent of the Z-estimation literature for which we mention some of the most relevant contributions. In the case where θ_0 is Euclidean, asymptotic normality is obtained in Huber (1967) and nonsmooth objective functions are considered in Pollard (1985). In the case where θ_0 is infinite dimensional, weak convergence is established in van der Vaart (1995). The presence of a nuisance parameter with possibly, slower than root n rates of convergence, is studied in Newey (1994) and nonsmooth objective functions are investigated in Chen et al. (2003). Relevant textbooks are Newey and McFadden (1994), van der Vaart and Wellner (1996), van der Vaart (1998), Kosorok (2008).

Among the different applications, we focus on *weighted linear regression* for *heteroscedastic models*. As this topic is quite well documented (see among others, Robinson (1987), Carroll et al. (1988) and the references therein), it allows for comparing our approach with the existing ones. Let $(Y_i, X_i)_{i=1,\dots,n}$ denote independently and identically distributed random variables with distribution P. The *weighted least squares estimator* is given by

$$\widehat{\beta}(w) = \operatorname{argmin}_{(\beta_1,\beta_2)\in\mathbb{R}^{1+q}} \sum_{i=1}^{n} (Y_i - \beta_1 + \beta_2^T X_i)^2 w(X_i), \tag{5}$$

where $w : \mathbb{R}^q \to \mathbb{R}$ is a measurable function. Among such a collection of estimators, there exists a member $\widehat{\beta}(w_0)$ with minimum variance (see Section 4.1 for details). Many studies have focused on the estimation of w_0. For instance, Carroll and Ruppert (1982) argues that a parametric estimation of w_0 can be performed, and Carroll (1982) and Robinson (1987) use different nonparametric estimators to approximate w_0. Usually, the estimators $\widehat{\beta}(\widehat{w})$ are shown to have minimal variance by relying on U-statistics-based decompositions. It involves relatively long and peculiar calculations depending on both $\widehat{w} : \mathbb{R}^q \to \mathbb{R}$ and the loss function. Our approach overpass this issue by providing high-level conditions on $\widehat{w}$ that are in some ways independent from the rest of the problem. To summarize, we require that $\widehat{w}(x) \to w_0(x)$ in probability, $dP(x)$-almost everywhere, and the existence of a function space $\mathcal{W}$, satisfying

$$P(\widehat{w} \in \mathcal{W}) \to 1 \qquad \text{and} \qquad \int_0^{+\infty} \sqrt{\log \mathcal{N}_{[]}\big(\epsilon, \mathcal{W}, L_r(P)\big)}\, d\epsilon < +\infty,$$

for some $r > 2$, where $\mathcal{N}_{[]}(\epsilon, \mathcal{W}, L_r(P))$ denotes the ϵ-bracketing number of the metric space $(\mathcal{W}, L_r(P))$ (van der Vaart and Wellner, 1996, Definition 2.1.6). As detailed in the paper, when w_0 is modelled parametrically, the previous conditions are fairly easy to verify. For nonparametric estimators of w_0, in particular for Nadaraya-Watson estimators, smoothness restrictions on the kernel function with respect to the dimension are appropriate to obtain, in the mean time, sufficiently sharp bounds on the bracketing numbers of $\mathcal{W}$ and that $\widehat{w}$ belongs to $\mathcal{W}$, with probability going to 1. In contrast to Carroll and Ruppert (1982) and Robinson (1987), the bandwidth sequence $(h_n)_{n\in\mathbb{N}}$ of the Nadaraya-Watson estimator is allowed to go to 0 as slowly as we wish but not too fast. It is required that $h_n \to 0$ and $nh_n^{2q+\delta} \to +\infty$, for some $\delta > 0$.

The paper is organised as follows. We describe in Section 2 some examples of estimators satisfying equation (4). Section 3 contains the theoretical background of the paper. We study the consistency of $\widehat{\theta}(\eta)$ (Section 3.1) and the weak convergence of $n^{1/2}(\widehat{\theta}(\eta) - \theta_0)$ in $\ell^\infty(\mathcal{H})^p$ (Section 3.2). Based on this, we establish conditions ensuring (1) (Section 3.3). In the end, we consider some weighted estimators for *conditional moment restrictions models* (Section 3.4). In Section 4, we focus on the metric entropy of estimators of the optimal weight function in *weighted linear regression*. We investigate different approaches, from the parametric to the fully nonparametric approach. In the later case, we study the Nadaraya-Watson estimator. In Section 5, we evaluate the finite sample performance of the methods by means of simulations.

2. Examples

As discussed in the introduction, the results of the paper allow to obtain (1) for estimators satisfying (4) that depend on a tuning parameter. This occurs at many levels of statistical theory. We raise several examples in the following.

Example 1. *(Least squares constrained estimation)* Given $\widehat{\theta}$, an arbitrary but consistent estimator of θ_0, the estimator $\widehat{\theta}_c$ is said to be a *least squares constrained estimator* if it minimizes $(\theta-\widehat{\theta})^T\Gamma(\theta-\widehat{\theta})$ over $\{\theta : g(\theta)=0\}$, for some function g, where Γ is lying over the set of symmetric positive definite matrices $\Gamma \geq b > 0$. Consequently, $\widehat{\theta}_c$ depends on the choice of Γ but since $|\widehat{\theta}_c-\widehat{\theta}|^2 \leq b^{-1}|\Gamma^{1/2}(\widehat{\theta}_c-\widehat{\theta})|^2 \leq b^{-1}|\Gamma^{1/2}(\theta_0-\widehat{\theta})|^2 \to 0$ in probability, the matrix Γ does not affect the consistency of $\widehat{\theta}_c$ estimating θ_0. It is well known that $\widehat{\theta}_c$ is a minimum variance estimator if Γ equals the inverse of the asymptotic variance of $\widehat{\theta}$ (Newey and McFadden, 1994, Section 5.2). Such a class is popular among econometricians and also known as *minimal distance estimator*.

In the above illustrative example, the use of the asymptotic equicontinuity of the process $\Gamma \mapsto n^{1/2}(\widehat{\theta}_c-\theta_0)$ is not really legitimate since we could obtain the asymptotics using more basic tools such as the Slutsky's lemma in Euclidean space. This is due of course to the Euclideanity of θ and Γ but also to the simplicity of the mapping $(\theta,\Gamma) \mapsto (\theta-\widehat{\theta})^T\Gamma(\theta-\widehat{\theta})$. Consequently, we highlight below more evolved examples in which either the tuning parameter is a function (Examples 2, 4 and 5) or the dependence structure between θ and η is more complicated than before (Example 3). To our knowledge, the asymptotics for the examples below are quite difficult to obtain.

Example 2. *(weighted linear regression)* This includes the estimators described by (5) but other losses than the square function might be used to adapt to the distribution of the noise. Examples are $L_r(\mathbb{P}_n)$-losses, Huber robust loss (see Example 3 for details), least absolute deviation and quantile losses. In a general framework covering every of the latter examples, a formula of the optimal weight function is established in Bates and White (1993).

Example 3. *(Huber cut-off)* Whereas weighted regression handles *heteroscedasticity* in the data, the cut-off in Huber *robust regression* carries out the adaptation to the distribution of the noise (Huber, 1967). The Huber objective function is defined as the continuous function that coincides with the identity on $[-c,c]$ (c is called the cut-off) and is constant elsewhere. A Z-estimator based on this function permits to handle heavy tails in the distribution of the noise. The choice of the cut-off might be done by minimizing the asymptotic variance.

Example 4. *(instrumental variable)* In Newey (1990), the class of *nonlinear instrumental variables* is defined through the *generalized method of moment*. The estimator $\widehat{\theta}$ depends on a so-called *matrix of instruments* W, and satisfies the equation $\sum_{i=1}^n W(\widetilde{Z}_i)\varphi(Z_i,\theta)=0$, where each $\widetilde{Z}_i$ is some set of coordinates of Z_i and φ is a given function. A formula for the optimal matrix of instruments is available.

Example 5. *(dimension reduction)* The method *sliced inverse regression* (Li, 1991) is based on estimating the subspace generated by the vectors $EX\psi(Y)$, when ψ varies in a given class of functions. Minimization the asymptotic variance leads to an expression of the optimal ψ_0 (Portier and Delyon, 2013).

3. Uniform *Z*-estimation theory

Define $(\mathcal{Z}_\infty, \mathcal{A}_\infty, P_\infty)$ as the probability space associated to the whole sequence $(Z_1, Z_2, \ldots)$. Random elements in $\ell^\infty(\mathcal{H})^p$, such as $\eta \mapsto \mathbb{G}_n \psi_\eta(\theta)$, are not necessarily measurable. To account for this, we work with the outer expectation E^o_∞ and the outer probability measure P^o_∞ (see the introduction of van der Vaart and Wellner (1996) for the definitions). Each convergence, in probability or in distribution, will be stated with respect to the outer probability. A class of functions $\mathcal{F}$ is said to be Glivenko-Cantelli if $\sup_{f\in\mathcal{F}} |(\mathbb{P}_n - P)f|$ goes to 0 in P^o_∞-probability. A class of functions $\mathcal{F}$ is said to be Donsker if $\mathbb{G}_n f$ converges weakly in $\ell^\infty(\mathcal{F})$ to a tight measurable element. Let $\|f\|_{L_2(P)} = \sqrt{Pf^2}$. A class $\mathcal{F}$ is Donsker if and only if it is totally bounded with respect to the $L_2(P)$-distance and if, for every $\epsilon > 0$,

$$\lim_{\delta\to 0} \limsup_{n\to+\infty} P^o_\infty \left(\sup_{\|f-g\|_{L_2(P)}<\delta} |\mathbb{G}_n(f-g)| > \epsilon \right) = 0, \tag{6}$$

where the supremum is taken over f and g in $\mathcal{F}$. The previous assertion follows from the characterization of tight sequences valued in the space of bounded functions (van der Vaart and Wellner, 1996, Theorem 1.5.7). We refer to the book van der Vaart and Wellner (1996) for a comprehensive study of the latter concepts.

For any element $A \in \mathbb{R}^{p\times q}$, let $|A|$ denote the Frobenius norm, i.e., $|A|^2 = \mathrm{tr}(A^T A)$. Note that if A is a vector, it coincides with the Euclidean norm. For $r > 0$ and f, a measurable function, let $\|f\|_{L_r(P)}$ denote the $L_r(P)$-norm of the function f.

For the sake of generality, we authorize, in Section 3.1 and 3.2, the parameter of interest θ_0 to depend on η. Hence we further assume that $\theta_0(\cdot)$ is an element of $\ell^\infty(\mathcal{H})^p$.

3.1. Uniform consistency

Before being possibly expressed as a Z-estimator, the parameter of interest θ_0 is often defined as an M-estimator, i.e., $\theta_0 \in \ell^\infty(\mathcal{H})^p$ is such that

$$\theta_0(\eta) = \mathrm{argmin}_{\theta\in\Theta} \, P m_\eta(\theta), \tag{7}$$

where $m_\eta(\theta) : \mathcal{Z} \to \mathbb{R}$ is a known real valued measurable function, for every $\theta \in \Theta$ and each $\eta \in \mathcal{H}$. The estimator of θ_0 is denoted by $\widehat{\theta}$, it depends on η since it satisfies

$$\widehat{\theta}(\eta) = \mathrm{argmin}_{\theta\in\Theta} \, \mathbb{P}_n m_\eta(\theta). \tag{8}$$

Both elements θ_0 and $\widehat{\theta}$ are $\mathbb{R}^p$-valued functions defined on $\mathcal{H}$. When dealing with consistency, considering M-estimators is more general but not more difficult than Z-estimators (see Remark 3). The following generalizes standard consistency theorems for M-estimators (van der Vaart, 1998, Theorem 5.7) to uniform consistency results.

Theorem 1. *Assume that (7) and (8) hold. Suppose that*

(a1) $\sup_{\eta\in\mathcal{H},\,\theta\in\Theta}|(\mathbb{P}_n-P)m_\eta(\theta)|\overset{P^o_\infty}{\to}0$.
(a2) For all $\delta>0$, *there exists* $\epsilon>0$ *such that*

$$\sup_{\eta\in\mathcal{H}}|\theta(\eta)-\theta_0(\eta)|\geq\delta\quad\Rightarrow\quad\sup_{\eta\in\mathcal{H}}P\{m_\eta(\theta(\eta))-m_\eta(\theta_0(\eta))\}\geq\epsilon.$$

then, we have $\sup_{\eta\in\mathcal{H}}|\widehat{\theta}(\eta)-\theta_0(\eta)|\overset{P^o_\infty}{\to}0$.

Proof. We follow the lines of the proof of Theorem 5.7 in van der Vaart (1998). Given $\delta>0$, assumption (a2) implies that there exists $\epsilon>0$ such that

$$P^o_\infty\left(\sup_{\eta\in\mathcal{H}}|\widehat{\theta}(\eta)-\theta_0(\eta)|\geq\delta\right)\leq P^o_\infty\left(\sup_{\eta\in\mathcal{H}}P\{m_\eta(\widehat{\theta}(\eta))-m_\eta(\theta_0(\eta))\}\geq\epsilon\right).$$

By definition, $\mathbb{P}_n\{m_\eta(\widehat{\theta}(\eta))-m_\eta(\theta_0(\eta))\}\leq 0$ for every $\eta\in\mathcal{H}$, then we know that

$$\begin{aligned}
&P\{m_\eta(\widehat{\theta}(\eta))-m_\eta(\theta_0(\eta))\}\\
&=(P-\mathbb{P}_n)\{m_\eta(\widehat{\theta}(\eta))\}+(\mathbb{P}_n-P)\{m_\eta(\theta_0(\eta))\}+\mathbb{P}_n\{m_\eta(\widehat{\theta}(\eta))-m_\eta(\theta_0(\eta))\}\\
&\leq(P-\mathbb{P}_n)\{m_\eta(\widehat{\theta}(\eta))\}+(\mathbb{P}_n-P)\{m_\eta(\theta_0(\eta))\}\\
&\leq 2\sup_{\theta\in\Theta,\,\eta\in\mathcal{H}}|(\mathbb{P}_n-P)\{m_\eta(\theta)\}|,
\end{aligned}$$

that goes to 0 in outer probability by (a1). □

Remark 1. *Condition (a1) requires the class* $\{m_\eta(\theta):\theta\in\Theta,\eta\in\mathcal{H}\}$ *to be Glivenko-Cantelli. It is enough to bound the uniform covering numbers or the bracketing numbers (van der Vaart and Wellner, 1996, Chapter 2.4). When* Θ *is unbounded, the Glivenko-Cantelli property may fail. Examples include* $L_r(\mathbb{P}_n)$*-losses in linear regression. In such situations, one may require the optimisation set* Θ *to be compact. Another possibility is to use, if available, special features of the functions* $\theta\mapsto m_\eta(\theta)$, $\eta\in\mathcal{H}$, *such as convexity (Newey, 1994, Theorem 2.7).*

Remark 2. *Condition (a2) is needed for the identifiability of the parameter* θ_0. *It says that when* $\theta(\cdot)$ *is not uniformly close to* $\theta_0(\cdot)$, *the objective function evaluated at* $\theta(\cdot)$ *is not uniformly small. Consequently, every sequence of functions* $\theta_n(\cdot)$ *such that* $\sup_{\eta\in\mathcal{H}}P\{m_\eta(\theta_n(\eta))-m_\eta(\theta_0(\eta))\}\to 0$ *as* $n\to+\infty$, *converges uniformly to* $\theta_0(\cdot)$. *It is a functional version of the so called "well-separated maximum" (Kosorok, 2008, page 252). It is stronger but often more convenient to verify*

(a2') $\inf_{\eta\in\mathcal{H}}\inf_{|\theta-\theta_0(\eta)|\geq\delta}P\{m_\eta(\theta)-m_\eta(\theta_0(\eta))\}>0$.

for every $\delta>0$. *This resembles to Van der Vaart's consistency conditions in van der Vaart (1998), Theorem 5.9. To show that (a2') implies (a2), suppose that* $\sup_{\eta\in\mathcal{H}}|\theta(\eta)-\theta_0(\eta)|\geq 2\delta$ *and write*

$$
\begin{aligned}
&P\{m_\eta(\theta(\eta)) - m_\eta(\theta_0(\eta))\} \\
&\geq \mathbb{1}_{\{|\theta(\eta)-\theta_0(\eta)|\geq\delta\}} P\{m_\eta(\theta(\eta)) - m_\eta(\theta_0(\eta))\} \\
&\geq \mathbb{1}_{\{|\theta(\eta)-\theta_0(\eta)|\geq\delta\}} \inf_{|\theta-\theta_0(\eta)|\geq\delta} P\{m_\eta(\theta) - m_\eta(\theta_0(\eta))\} \\
&\geq \mathbb{1}_{\{|\theta(\eta)-\theta_0(\eta)|\geq\delta\}} \inf_{\eta\in\mathcal{H}} \inf_{|\theta-\theta_0(\eta)|\geq\delta} P\{m_\eta(\theta) - m_\eta(\theta_0(\eta))\}.
\end{aligned}
$$

Conclude by taking the supremum over $\mathcal{H}$ in both side.

Remark 3. *Estimators defined through zeros of the map $\mathbb{P}_n\psi_\eta(\theta)$ are also minimizers of $|\mathbb{P}_n\psi_\eta(\theta)|$. Therefore they can be handle by Theorem 1. Let $\theta_0(\cdot)$ be such that $P\psi_\eta(\theta_0(\eta)) = 0$ for every $\eta \in \mathcal{H}$. If (a1) holds replacing m by ψ and if for all $\delta > 0$, there exists $\epsilon > 0$ such that*

$$
\sup_{\eta\in\mathcal{H}} |\theta(\eta) - \theta_0(\eta)| \geq \delta > 0 \quad \Rightarrow \quad \sup_{\eta\in\mathcal{H}} |P\psi_\eta(\theta(\eta))| \geq \epsilon > 0,
$$

then, the uniform convergence of zeros of $\mathbb{P}_n\psi_\eta(\theta)$ to the zero of $P\psi_\eta(\theta)$ is guaranteed. Given that m_η is differentiable, the associated M-estimator can be expressed as a Z-estimator with objective function $\nabla_\theta m_\eta$. Because a function can have several local minimums, the previous condition with $\nabla_\theta m_\eta$ is stronger than (a2). Consequently, for consistency purpose, M-estimators should not be expressed in terms of Z-estimators (see also Newey (1994), page 2117).

3.2. Weak convergence

We now consider the weak convergence properties of Z-estimators indexed by the objective functions. We assume further that $\theta_0 \in \ell^\infty(\mathcal{H})^p$ and satisfies the p-dimensional set of equations, for each $\eta \in \mathcal{H}$,

$$
P\psi_\eta(\theta_0(\eta)) = 0, \tag{9}
$$

where $\psi_\eta(\theta) : \mathcal{Z} \to \mathbb{R}^p$ is a known measurable function. The estimator of $\theta_0(\cdot)$ is denoted by $\widehat{\theta}(\cdot)$ and for each η, it holds that

$$
\mathbb{P}_n\psi_\eta(\widehat{\theta}(\eta)) = 0. \tag{10}
$$

Here we shall suppose that $\sup_{\eta\in\mathcal{H}} |\widehat{\theta}(\eta)-\theta_0(\eta)| = o_{P^o_\infty}(1)$, so that the functions ψ_η, $\eta \in \mathcal{H}$, are not intended to necessarily satisfy (a2). Indeed consistency of $\widehat{\theta}(\cdot)$ may have been established from other restrictions such as being a minimizer (see Remark 3).

We require some "uniform" Frechet differentiability for the map $\theta \mapsto P\psi_\eta(\theta)$, that is, there exists $A_\eta : \Theta \mapsto \mathbb{R}^{p\times p}$ such that, for all $\delta_n \to 0$,

$$
\sup_{0<|\theta-\widetilde{\theta}|\leq\delta_n,\, \eta\in\mathcal{H}} \left\{ \frac{\left| P\psi_\eta(\theta) - P\psi_\eta(\widetilde{\theta}) - A_\eta(\theta)(\theta-\widetilde{\theta}) \right|}{|\theta-\widetilde{\theta}|} \right\} \longrightarrow 0. \tag{11}
$$

Theorem 2. *Assume that (9) and (10) hold. Suppose that*

(a3) $\sup_{\eta\in\mathcal{H}} |\widehat{\theta}(\eta) - \theta_0(\eta)| \overset{P^o_\infty}{\to} 0$.
(a4) Let $\psi_{\eta,k}$ *denote the* k*-th coordinate of* ψ_η*. For all* $\epsilon > 0$*, there exists* $\delta > 0$ *such that* $|\theta - \widetilde{\theta}| < \delta$ *implies that* $\max_{k\in\{1,\dots,q\}} \sup_{\eta\in\mathcal{H}} \|\psi_{\eta,k}(\theta) - \psi_{\eta,k}(\widetilde{\theta})\|_{L_2(P)} < \epsilon$.
(a5) The matrix $B_\eta := A_\eta(\theta_0(\eta))$*, defined in (11), is bounded and invertible uniformly in* η.
(a6) There exists $\delta > 0$ *such that the class* $\Psi := \{z \mapsto \psi_\eta(\theta)(z) : |\theta - \theta_0(\eta)| < \delta,\ \eta \in \mathcal{H}\}$ *is* P*-Donsker.*

then, we have

$$\sup_{\eta\in\mathcal{H}} \left| n^{1/2}(\widehat{\theta}(\eta) - \theta_0(\eta)) + B_\eta^{-1}\mathbb{G}_n\psi_\eta(\theta_0(\eta)) \right| = o_{P^o_\infty}(1).$$

Consequently, $n^{1/2}(\widehat{\theta}(\eta) - \theta_0(\eta))$ *converges weakly to a tight zero-mean Gaussian element in* $\ell^\infty(\mathcal{H})$ *whose covariance function is given by*

$$(\eta_1, \eta_2) \mapsto B_{\eta_1}^{-1} P(\psi_{\eta_1}(\theta_0(\eta_1))\psi_{\eta_2}(\theta_0(\eta_2))^T) B_{\eta_2}^{-1}.$$

Proof. We follow a standard approach by first deriving the (uniform) rates of convergence and second computing the asymptotic distribution (van der Vaart, 1998, Theorem 5.21). From (a3) and (a4), we know that, for some nonrandom positive sequence $\delta_n \to 0$, the set

$$E_n = \left\{ \sup_{\eta\in\mathcal{H}} |\widehat{\theta}(\eta) - \theta_0(\eta)| < \delta_n,\ \max_{k\in\{1,\dots,p\}} \sup_{\eta\in\mathcal{H}} \|\psi_{\eta,k}(\widehat{\theta}(\eta)) - \psi_{\eta,k}(\theta(\eta))\|_{L_2(P)} < \delta_n \right\},$$

is such that $P^o_\infty(E_n) \to 1$. Because we are interested in showing convergence in probability, we can restrict attention to E_n. By definition of $\widehat{\theta}(\eta)$, we have

$$\begin{aligned} 0 &= n^{1/2}\{\mathbb{P}_n\psi_\eta(\widehat{\theta}(\eta)) - P\psi_\eta(\theta_0(\eta))\} \\ &= \mathbb{G}_n\{\psi_\eta(\widehat{\theta}(\eta)) - \psi_\eta(\theta_0(\eta))\} + \mathbb{G}_n\psi_\eta(\theta_0(\eta)) + n^{1/2}P\{\psi_\eta(\widehat{\theta}(\eta)) - \psi_\eta(\theta_0(\eta))\}. \end{aligned} \tag{12}$$

The first term is treated as follows. Under E_n, we have

$$\begin{aligned} |\mathbb{G}_n\{\psi_\eta(\widehat{\theta}(\eta)) - \psi_\eta(\theta_0(\eta))\}| &\le \sqrt{q} \max_{k\in\{1,\dots,p\}} |\mathbb{G}_n\{\psi_{\eta,k}(\widehat{\theta}(\eta)) - \psi_{\eta,k}(\theta_0(\eta))\}| \\ &\le \sqrt{p} \sup_{\|\psi - \widetilde{\psi}\|_{L_2(P)} < \delta_n} |\mathbb{G}_n\{\psi - \widetilde{\psi}\}|, \end{aligned}$$

where the supremum is over ψ and $\widetilde{\psi}$ in Ψ. Using (a6) together with equation (6), the first term in (12) goes to 0 in P^o_∞-probability. As a consequence, we have that $\mathbb{G}_n\psi_\eta(\theta_0(\eta)) + n^{1/2}P\{\psi_\eta(\widehat{\theta}(\eta)) - \psi_\eta(\theta_0(\eta))\} = o_{P^o_\infty}(1)$ or, equivalently, that

$$\mathbb{G}_n\{\psi_\eta(\theta_0(\eta))\} + B_\eta n^{1/2}(\widehat{\theta}(\eta) - \theta_0(\eta)) = a_n(\eta) + o_{P^o_\infty}(1), \tag{13}$$

where the $o_{P^o_\infty}(1)$ is uniform in $\eta \in \mathcal{H}$ and $a_n(\eta) = -n^{1/2}\{P\{\psi_\eta(\widehat{\theta}(\eta)) - \psi_\eta(\theta_0(\eta))\} - B_\eta n^{1/2}(\widehat{\theta}(\eta) - \theta_0(\eta))\}$. Using (a5), we have that

$$\begin{aligned} a_n(\eta) &\le |n^{1/2}(\widehat{\theta}(\eta) - \theta_0(\eta))| \sup_{0<|\theta-\widetilde{\theta}|\le\delta_n} \left\{ \frac{|P\psi_\eta(\theta) - P\psi_\eta(\widetilde{\theta}) - A_\eta(\theta)(\theta - \widetilde{\theta})|}{|\theta - \widetilde{\theta}|} \right\} \\ &\le \sup_{\eta\in\mathcal{H}} |n^{1/2}(\widehat{\theta}(\eta) - \theta_0(\eta))| \times o(1). \end{aligned} \tag{14}$$

Hence, because we know from (a6) that $\sup_{\eta\in\mathcal{H}} |\mathbb{G}_n\psi_\eta(\theta_0(\eta))| = O_{P^o_\infty}(1)$, and using (a5) again, in particular the full rank condition on B_η, we get

$$\sup_{\eta\in\mathcal{H}} |n^{1/2}(\widehat{\theta}(\eta) - \theta_0(\eta))| \le \sup_{\eta\in\mathcal{H}} \{|B_\eta n^{1/2}(\widehat{\theta}(\eta) - \theta_0(\eta))|\} \sup_{|u|=1} |B_\eta^{-1}u| = O_{P^o_\infty}(1).$$

Bringing the previous information in equation (14) gives that $\sup_{\eta\in\mathcal{H}} |a_n(\eta)| = o_{P^o_\infty}(1)$. Therefore equation (13) becomes

$$\sup_{\eta\in\mathcal{H}} \left| \mathbb{G}_n\psi_\eta(\theta_0(\eta)) + B_\eta n^{1/2}(\theta(\eta) - \theta_0(\eta)) \right| = o_{P^o_\infty}(1),$$

and the conclusion follows. □

Remark 4. *Weak convergence of M-estimators is in general more difficult to handle than weak convergence of Z-estimators (van der Vaart and Wellner, 1996, chapter 3.2). An interesting strategy is to focus on convex objective functions as developed in Pollard (1985). Unlike the approach taken in Theorem 2, this strategy handles non-smooth objective functions and it turns out to be useful to study least absolute deviation estimators. More recently, Kato (2009) considers convex objective functions that are indexed by real parameters. The main application deals with weak convergence of the quantile regression process.*

Remark 5. *All the examples in Section 2 focus on particular situations where θ_0 does not depend on η. Hence, $\{\psi_\eta, \eta \in \mathcal{H}\}$, represents a range of criterion functions available for estimating a single parameter θ_0. In this context, condition (a4) becomes*

(a4') For all $k \in \{1, \dots, p\}$, $\sup_{\eta\in\mathcal{H}} \|\psi_{\eta,k}(\theta) - \psi_{\eta,k}(\theta_0)\|_{L_2(P)} \to 0$, as $\theta \to \theta_0$.

and condition (a5) is reduced to

(a5') There exists $B_\eta \in \mathbb{R}^{p\times p}$, bounded and invertible, uniformly in η, such that

$$\sup_{\eta\in\mathcal{H}} |P\psi_\eta(\theta) - P\psi_\eta(\theta_0) - B_\eta(\theta - \theta_0)| = o(\theta - \theta_0).$$

3.3. Asymptotic equivalence

In this section, Theorem 2 is used to establish conditions for the asymptotics of $\widehat{\theta}(\widehat{\eta})$ estimating $\theta_0 \in \Theta$. Hence, we shall assume that for every $\eta \in \mathcal{H}$, $\theta_0(\eta) = \theta_0$,

as in the introduction and as in Remark 5. Let $\widehat{\eta}$ denote a consistent estimator of η_0. The next theorem asserts that, whatever the rate of convergence of $\widehat{\eta}$, $\widehat{\theta}(\eta_0)$ and $\widehat{\theta}(\widehat{\eta})$ have the same asymptotic behaviour. Consequently, whenever $\widehat{\theta}(\eta_0)$ is an efficient estimator of θ_0, $\widehat{\theta}(\widehat{\eta})$ is an efficient estimator of θ_0. For the sake of generality, we consider two cases: when the class $\mathcal{H}$ does not change with n and when it does.

Theorem 3. *Assume that (9), (10), (a3), (a4'), (a5') and (a6) hold. Suppose that*

(a7) For every $\eta \in \mathcal{H}$, $\theta_0(\eta) = \theta_0$.
(a8) There exist $\eta_0 \in \mathcal{H}$ and $\widehat{\eta}$ such that

(i) $P^o_\infty(\widehat{\eta} \in \mathcal{H}) \to 1$.

(ii) $\max_{k\in\{1,\dots,p\}} \|\psi_{\widehat{\eta},k}(\theta_0) - \psi_{\eta_0,k}(\theta_0))\|_{L_2(P)} \overset{P^o_\infty}{\to} 0$ and $|B_{\widehat{\eta}} - B_{\eta_0}| \overset{P^o_\infty}{\to} 0$.

then, $n^{1/2}(\widehat{\theta}(\widehat{\eta}) - \theta_0) = n^{1/2}(\widehat{\theta}(\eta_0) - \theta_0) + o_{P^o_\infty}(1)$.

Proof. Since

$$\widehat{\theta}(\widehat{\eta}) - \theta_0 = (\widehat{\theta}(\widehat{\eta}) - \widehat{\theta}(\eta_0)) + (\widehat{\theta}(\eta_0) - \theta_0),$$

we have to show that the first term of the right-hand side is neglectabe, i.e., $\widehat{\theta}(\widehat{\eta}) - \widehat{\theta}(\eta_0) = o_{P^o_\infty}(n^{-1/2})$. By (a8), for a certain nonrandom positive sequence $\delta_n \to 0$, the event

$$\left\{ \widehat{\eta} \in \mathcal{H}, \quad \max_{k\in\{1,\dots,p\}} \|\psi_{\widehat{\eta},k}(\theta_0) - \psi_{\eta_0,k}(\theta_0)\|_{L_2(P)} < \delta_n, \quad |B_{\widehat{\eta}} - B_{\eta_0}| < \delta_n \right\},$$

has probability going to 1. As we are concerned with convergence in probability, we can restrict attention to this event. Applying Theorem 2, we find

$$n^{1/2}(\widehat{\theta}(\widehat{\eta}) - \widehat{\theta}(\eta_0)) = \\ B^{-1}_{\widehat{\eta}} \mathbb{G}_n\{\psi_{\eta_0}(\theta_0) - \psi_{\widehat{\eta}}(\theta_0)\} + (B^{-1}_{\eta_0} - B^{-1}_{\widehat{\eta}})\mathbb{G}_n\psi_{\eta_0}(\theta_0) + o_{P^o_\infty}(1).$$

By (a5'), the second term in the right-hand side equals $B^{-1}_{\eta_0} - B^{-1}_{\widehat{\eta}} = B^{-1}_{\eta_0}(B_{\widehat{\eta}} - B_{\eta_0})B^{-1}_{\widehat{\eta}} = O(\delta_n)$ multiplied by a term which is bounded in probability, from (a6). To obtain the convergence in probability to 0 of the first term in the right-hand side, we follow a similar approach as in the proof of Theorem 2, i.e., we make use of (a6) to rely on the stochastic equicontinuity, as expressed in (6). □

Now we consider the case when the class $\mathcal{H}$ does change with n. We rely on results from van der Vaart and Wellner (2007) which considers empirical processes indexed by estimated functions. It requires to bound the ϵ-bracketing numbers of the class Ψ together with a Lindeberg condition on the class. Similar conditions can also be derived considering the covering numbers.

Theorem 4. *Let $\mathcal{H} := \mathcal{H}_n$ and assume that (9), (10), (a3), (a4'), (a5'), (a7) and (a8) hold (with $\mathcal{H}_n$ in place of $\mathcal{H}$). Suppose that*

(a6') *Let* $\Psi_{n,k} := \{z \mapsto \psi_{\eta,k}(\theta)(z) : |\theta - \theta_0(\eta)| < \delta, \eta \in \mathcal{H}_n\}$ *and* $\overline{\psi}_{n,k}$ *be a measurable envelope for the class* $\Psi_{n,k}$*, i.e.,* $|\psi(z)| \leq \overline{\psi}_{n,k}(z)$ *for every* $\psi \in \Psi_{n,k}$ *and* $z \in \mathcal{Z}$*. There exists* $s > 0$ *such that* $\max_{k\in\{1,\dots,p\}} P|\overline{\psi}_{n,k}|^{2+s} < +\infty$ *and, for every* $\delta_n \to 0$ *and every* $k \in \{1,\dots,p\}$*,*

$$\int_0^{\delta_n} \sqrt{\log \mathcal{N}_{[]}\big(\epsilon\|\overline{\psi}_{n,k}\|_{L_2(P)}, \Psi_{n,k}, L_2(P)\big)}d\epsilon \longrightarrow 0.$$

then, $n^{1/2}(\widehat{\theta}(\widehat{\eta}) - \theta_0) = n^{1/2}(\widehat{\theta}(\eta_0) - \theta_0) + o_{P^o_\infty}(1)$.

Proof. The proof is the same as the proof of Theorem 3 with one change. We no longer rely on the Donsker property to provide

$$\sup_{\eta\in\mathcal{H}_n} \left|\mathbb{G}_n\big\{\psi_\eta(\widehat{\theta}(\eta)) - \psi_\eta(\theta_0)\big\}\right| \xrightarrow{P^o_\infty} 0, \tag{15}$$

$$\left|\mathbb{G}_n\big\{\psi_{\widehat{\eta}}(\theta_0) - \psi_{\eta_0}(\theta_0)\big\}\right| \xrightarrow{P^o_\infty} 0, \tag{16}$$

respectively, in the proof of Theorems 2 and 3. We rely on Theorem 2.2 in van der Vaart and Wellner (2007), which asserts that (16) holds whenever, for every $k \in \{1,\dots,p\}$,

$$\begin{aligned}
&P(\psi_{\widehat{\eta},k}(\theta_0) - \psi_{\eta_0,k}(\theta_0))^2 \xrightarrow{P^o_\infty} 0,\\
&\int_0^{\delta_n} \sqrt{\log \mathcal{N}_{[]}\big(\epsilon\|\overline{\psi}_{n,k}\|_{L_2(P)}, \Psi_{n,k}, L_2(P)\big)}d\epsilon \xrightarrow{\delta_n\to 0} 0,\\
&P\overline{\psi}^2_{n,k} = O(1), \qquad P\overline{\psi}^2_{n,k}\mathbb{1}_{\{\overline{\psi}_{n,k}\geq\epsilon n^{1/2}\}} \to 0, \text{ for each } \epsilon > 0.
\end{aligned}$$

The first condition is (a8). The second is (a6'). The third is obtained from the Hölder inequality using the moment condition on $\overline{\psi}_{n,k}$ in (a6'). The same can be done to obtain (15). □

Remark 6. *Covering and bracketing numbers are classically employed to deal with weak convergence of empirical processes (van der Vaart and Wellner, 1996, Chapter 2.5). It often gives tractable conditions that can be verified in practice (see Chapter 2.6 and 2.7 in van der Vaart and Wellner (1996), and Section 5 of the present paper for applications to semiparametric estimators). In our approach, the entropy conditions allow moreover to consider classes that depends on n. It turns out to be important when treating weighted regression estimators in Section 5, in which* $\widehat{\eta}$ *is a Nadaraya-Watson estimator.*

3.4. Conditional moment restrictions

We now consider conditional moment restrictions models defined as follows. There exists $\beta_0 \in \mathbb{R}^p$ such that

$$E(\varphi(Z, \beta_0)|X) = 0, \tag{17}$$

where $X \in \mathcal{X}$ and $Z \in \mathcal{Z}$ are random variables with joint distribution P and φ is a known $\mathbb{R}^p$-valued function. Equation (17) implies that infinitely many (unconditional) equations are available to characterize β_0, that is, for every bounded measurable real function w defined on $\mathcal{X}$, one has

$$E(w(X)\varphi(Z,\beta_0)) = 0.$$

Let $(Z,X),(Z_1,X_1),(Z_2,X_2),\dots$ denote an independent and identically distributed sequence of random variables satisfying model (17). The probability measure associated to the sequence $(Z_1,X_1),(Z_2,X_2),\dots$ is still denoted by P_∞. The estimator $\widehat{\beta}(w)$ satisfies

$$n^{-1}\sum_{i=1}^{n} w(X_i)\varphi(Z_i,\widehat{\beta}(w)) = 0, \tag{18}$$

for every w in $\mathcal{W}$, a class of bounded real functions. Note that this framework includes Example 2 of Section 2. In the case where $\mathcal{W}$ is an $\mathbb{R}^{p\times p}$-valued class of functions, it includes Example 4 of Section 2. In the following theorem, $\mathcal{W}$ is a real-valued class of functions (see Remark 9 for $\mathbb{R}^{p\times p}$-valued classes of functions). The proof follows from an application of Theorems 3 and 4 to the particular case of a Z-estimator defined with the objective function $(\beta,w)\mapsto w(\cdot)\varphi(\cdot,\beta)$. As in the previous section, we start by considering the case when $\mathcal{W}$ is fixed.

Theorem 5. *Assume that (17) and (18) hold. Suppose that*

(b1) $\sup_{w\in\mathcal{W}}|\widehat{\beta}(w)-\beta_0| \overset{P_\infty^o}{\to} 0$.

(b2) Let φ_k *denote the* k*-th coordinate of* φ*. Whenever* $\beta\to\beta_0$*, we have* $\max_{k\in\{1,\dots,p\}}\|\varphi_k(Z,\beta)-\varphi_k(Z,\beta_0)\|_{L_2(P)}\to 0$.

(b3) There exist $\kappa>0$ *and* $B:\mathcal{X}\to\in\mathbb{R}^{p\times p}$ *such that* $E|B(X)|<+\infty$ *and*

$$E\left\{\left|E\left[\varphi(Z,\beta)-\varphi(Z,\beta_0)\mid X\right]-B(X)(\beta-\beta_0)\right|\right\}\le\kappa|\beta-\beta_0|^2,$$

and where $Ew(X)B(X)$ *is invertible, uniformly in* $w\in\mathcal{W}$.

(b4) Let $\mathcal{B}_0$ *be an open ball centred at* β_0 *and* $\overline{\varphi}_k(z)=\sup_{\beta\in\mathcal{B}_0}|\varphi_k(z,\beta)|$*. There exists* $s>0$ *such that* $\max_{k\in\{1,\dots,p\}}P|\overline{\varphi}_k|^{2+s}<+\infty$ *and the classes* $\Phi_k:=\{z\mapsto\varphi_k(z,\beta):\beta\in\mathcal{B}_0\}$*,* $k\in\{1,\dots,p\}$*, are* P*-Donsker. Moreover,* $\mathcal{W}$ *is uniformly bounded by* 1 *and the class* $\overline{\varphi}_k\mathcal{W}$ *is* P*-Donsker.*

(b5) There exist $\widehat{w}:\mathcal{X}\mapsto\mathbb{R}$ *(suitably measurable) and* $w_0:\mathcal{X}\mapsto\mathbb{R}$ *such that*

(i) $P_\infty^o(\widehat{w}\in\mathcal{W})\to 1$.

(ii) $|\widehat{w}(x)-w_0(x)| \overset{P_\infty^o}{\to} 0$*,* $dP(x)$*-almost everywhere.*

then, $n^{1/2}(\widehat{\beta}(\widehat{w})-\beta_0)=n^{1/2}(\widehat{\beta}(w_0)-\beta_0)+o_{P_\infty^o}(1)$.

Proof. We verify each condition of Theorem 3 for the map $\psi_\eta(\theta)$ given by $w(\cdot)\varphi(\cdot,\beta)$ in which β and w replace, respectively, θ and η. The space corresponding to $\mathcal{H}$ in Theorem 3 is here $\mathcal{W}$. Note first that (9), (10) and (a3) are implied by (17), (18) and (b1), respectively. Moreovoer (b2) implies (a4') and

(17) implies (a7). To complete the proof, we show that (a5'), (a6) and (a8) hold (see Remark 5 for (a4') and (a5')).

We start by showing that (b3) $\Rightarrow$ (a5') with B_η equal to $Ew(X)B(X)$. From (b3), the matrix $Ew(X)B(X)$ is invertible and bounded. Moreover by (b3), we have

$$\left|E\left\{w(X)(\varphi(Z,\beta)-\varphi(Z,\beta_0)-B(X)(\beta-\beta_0))\right\}\right| \leq \kappa|\beta-\beta_0|^2,$$

which implies (a5').

We now show that (b4) implies (a6), that is, we need to prove that the class $\{(x,z)\mapsto w(x)\varphi(z,\beta) : \beta\in B_0, w\in\mathcal{W}\}$ is P-Donsker. Consider the k-th coordinate class $\Phi_k\times\mathcal{W}=\{w(\cdot)\varphi_k(\cdot,\beta), \beta\in\mathcal{B}_0, w\in\mathcal{W}\}$. Because it is the product of two classes, Φ_k and $\mathcal{W}$, we can apply Corollary 2.10.13 in van der Vaart and Wellner (1996). Given two pairs $(\beta,\widetilde{\beta})$ and $(w,\widetilde{w})$, we check that, for every $x\in\mathcal{X}$ and $z\in\mathcal{Z}$,

$$\begin{aligned}&(w(x)\varphi_k(z,\beta)-\widetilde{w}(x)\varphi_k(z,\widetilde{\beta}))^2 \leq\\ &2(\varphi_k(z,\beta)-\varphi_k(z,\widetilde{\beta}))^2+2\sup_{\beta\in\mathcal{B}_0}|\varphi_k(z,\beta)|^2(w(x)-\widetilde{w}(x))^2.\end{aligned}$$

This corresponds to (2.10.12) in van der Vaart and Wellner (1996) with $L_{\alpha,1}=\sqrt{2}$ and $L_{\alpha,2}=\sqrt{2}\overline{\varphi}_k$. By assumption, the suitable classes, Φ_k and $\overline{\varphi}_k\mathcal{W}$, are P-Donsker. It remains to note that any member of the product class is square integrable as $P\overline{\varphi}_k^2<+\infty$. Therefore, we have that the class $\Phi_k\times\mathcal{W}$ is P-Donsker. Hence it is a tight sequence in $\ell^\infty(\mathcal{B}_0\times\mathcal{W})$. Since tightness of vector-valued random sequences is equivalent to tightness of each coordinate, the sequence $\{w(\cdot)\varphi(\cdot,\beta)), \beta\in\mathcal{B}_0, w\in\mathcal{W}\}$ is tight. Using the multivariate central limit theorem, we obtain the convergence in distribution of the finite dimensional distributions. From Theorem 1.5.4 in van der Vaart and Wellner (1996), the class $\{(x,z)\mapsto w(x)\varphi(z,\beta) : \beta\in B_0, w\in\mathcal{W}\}$ is P-Donsker. Note that the moments of order $2+s$ for $\overline{\varphi}_k$ have not been used yet.

It remains to show that (b5) implies (a8). Given $\epsilon>0$ and using that $|\int A(x)dx|\leq\int|A(x)|dx$, we have

$$\begin{aligned}&\left|\int(\widehat{w}(x)-w_0(x))B(x)dP(x)\right|\\ &\leq\int|B(x)|\,|\widehat{w}(x)-w_0(x)|dP(x)\\ &\leq\epsilon\int|B(x)|dP(x)+2\int|B(x)|\mathbb{1}_{\{|\widehat{w}(x)-w_0(x)|>\epsilon\}}dP(x).\end{aligned}$$

Taking the expectation, Fubini's theorem (measurability is here implicitly assumed) leads to

$$\begin{aligned}&E_\infty\left|\int(\widehat{w}(x)-w_0(x))B(x)dP(x)\right|\\ &\leq\epsilon\int|B(x)|dP(x)+2\int|B(x)|P_\infty(|\widehat{w}(x)-w_0(x)|>\epsilon)dP(x),\end{aligned}$$

the right-hand side goes to 0 by the Lebesgue dominated convergence theorem. Conclude choosing ϵ small. Using that $P|\overline{\varphi}_k|^{2+s} < +\infty$, the same analysis together with the Hölder inequality, leads to the fact that, for every $k \in \{1, \dots, p\}$, $E\varphi_k(Z, \beta_0)^2(\widehat{w}(X) - w_0(X))^2$ goes to 0 in P_∞-probability. □

Remark 7. *Condition (b3) is related to the regularity of the map defined as $\beta \mapsto E(\varphi(Z, \beta)|X)$. It is in general weaker than asking for the regularity of the map $\beta \mapsto \varphi(z, \beta)$. For instance, it permits to include the Huber loss function (defined in Example 3). Note also that, contrary to $\mathcal{W}$, the class of functions $\{z \mapsto \varphi(z, \beta) : \beta \in \mathcal{B}_0\}$ is not supposed to be bounded. This is important to have this flexibility in order to include examples such as weighted least squares.*

Remark 8. *Under the conditions of Theorem 5, the sequence $n^{1/2}(\widehat{\beta}(w) - \beta_0)$ converges weakly in $\ell^\infty(\mathcal{W})$ to a tight zero-mean Gaussian element whose covariance function is given by*

$$(w_1, w_2) \mapsto C_{w_1}^{-1} E\left(w_1(X)\varphi(Z, \beta_0)\varphi(Z, \beta_0)^T w_2(X)\right) C_{w_2}^{-1},$$

with $C_w = E(w(X)B(X))$.

Finally we treat the case when $\mathcal{W} := \mathcal{W}_n$ is changing with n by considering the bracketing numbers of the underlying classes.

Theorem 6. *Let $\mathcal{W} := \mathcal{W}_n$ and assume that (17), (18), (b1), (b2), (b3) and (b5) hold (with $\mathcal{W}_n$ in place of $\mathcal{W}$). Suppose that*

(b4') Let $\mathcal{B}_0$ be an open ball centred at β_0, $\Phi_k = \{z \mapsto \varphi_k(z, \beta) : \beta \in \mathcal{B}_0\}$, $\overline{\varphi}_k(z) = \sup_{\beta \in \mathcal{B}_0} |\varphi_k(z, \beta)|$. For all $k \in \{1, \dots, p\}$, there exists $s > 0$ such that $P|\overline{\varphi}_k|^{2+s} < +\infty$, and, for every sequence $\delta_n \to 0$ and $r = 2(2+s)/s$,

$$(i) \qquad \int_0^{+\infty} \sqrt{\log \mathcal{N}_{[]}\big(\epsilon \|\overline{\varphi}_k\|_{L_2(P)}, \Phi_k, L_2(P)\big)}\, d\epsilon < +\infty,$$

$$(ii) \qquad \int_0^{\delta_n} \sqrt{\log \mathcal{N}_{[]}\big(\epsilon, \mathcal{W}_n, L_r(P)\big)}\, d\epsilon \longrightarrow 0.$$

Moreover the functions in $\mathcal{W}_n$ are bounded by 1.

then, $n^{1/2}(\widehat{\beta}(\widehat{w}) - \beta_0) = n^{1/2}(\widehat{\beta}(w_0) - \beta_0) + o_{P_\infty^o}(1)$.

Proof. We apply Theorem 4 with Ψ_n equal to $\{z \mapsto w(x)\varphi(z, \beta) : \beta \in \mathcal{B}_0, w \in \mathcal{W}_n\}$. From the proof of Theorem 3, we have that (17), (18), (b1), (b2), (b3) and (b5) implies (9), (10), (a3), (a4'), (a5'), (a7) and (a8). We finish the proof by showing that (b4') is enough to get (a6') with $\Psi_{n,k}$ equal to $\Phi_k \times \mathcal{W}_n$.

Given $\epsilon > 0$, let $[\underline{\varphi}^{(i)}, \overline{\varphi}^{(i)}]$, $i = 1, \dots, n_1$, be $(\epsilon\|\overline{\varphi}_k\|_{L_2(P)}, L_2(P))$-brackets covering Φ_k and let $[\underline{w}^{(j)}, \overline{w}^{(j)}]$, $j = 1, \dots, n_2$, be $(\epsilon, L_r(P))$-brackets covering $\mathcal{W}_n$, with $r = 2(2+s)/s$. Because the function $z \mapsto xy$ attains its bounds on every rectangle at the edges of each rectangle, the brackets

$$[\min(g_{ij}), \max(g_{ij})], \quad i = 1, \dots, n_1,\ j = 1, \dots, n_2,$$

with $g_{ij} = (\underline{\varphi}^{(i)}\underline{w}^{(j)}, \underline{\varphi}^{(i)}\overline{w}^{(j)}, \overline{\varphi}^{(i)}\underline{w}^{(j)}, \overline{\varphi}^{(i)}\overline{w}^{(j)})$, cover the class $\Phi_k \times \mathcal{W}_n$. Moreover, we have

$$|\max(g_{ij}) - \min(g_{ij})| \leq |\overline{\varphi}^{(i)} - \underline{\varphi}^{(i)}| + |\overline{\varphi}_k||\overline{w}^{(j)} - \underline{w}^{(j)}|,$$

then, using Minkowski's, Hölder's and Jensen's inequalities (in this order), we get

$$\begin{aligned}\|\max(g_{ij}) - \min(g_{ij})\|_{L_2(P)} &\leq \epsilon\|\overline{\varphi}_k\|_{L_2(P)} + \|\overline{\varphi}_k(\overline{w}^{(j)} - \underline{w}^{(j)})\|_{L_2(P)} \\ &\leq \|\overline{\varphi}_k\|_{L_2(P)} + \|\overline{\varphi}_k\|_{L_{2+s}(P)}\,\|\overline{w}^{(j)} - \underline{w}^{(j)}\|_{L_r(P)} \\ &\leq 2\epsilon\|\overline{\varphi}_k\|_{L_{2+s}(P)},\end{aligned}$$

with $r = 2(2+s)/s$. Hence we have shown that, for every $\epsilon > 0$, $\mathcal{N}_{[]}\big(2\epsilon\|\overline{\varphi}_k\|_{L_{2+s}(P)}, \Phi_k \times \mathcal{W}_n, L_2(P)\big)$ is smaller than $\mathcal{N}_{[]}\big(\epsilon\|\overline{\varphi}_k\|_{L_2(P)}, \Phi_k, L_2(P)\big)$ times $\mathcal{N}_{[]}\big(\epsilon, \mathcal{W}_n, L_r(P)\big)$. This implies the integrability condition in (a6). □

Remark 9. *In Example 4, the class $\mathcal{W}$ is a matrix-valued class of functions. The statements in Theorems 5 and 6 only deal with the real-valued case. To be valid in the matrix-valued case, one needs to assume the same assumptions as in Theorems 5 and 6 but for each coordinate of the function class $\mathcal{W}$. The main reason for this is that the sum of two Donsker classes is Donsker.*

4. Application to weighted linear regression

In this section, we are interested in estimating $\beta_0 = (\beta_{01}, \beta_{02}) \in \mathbb{R}^{1+q}$, defined by the following model

$$E(Y|X) = \beta_{01} + \beta_{02}^T X, \tag{19}$$

where the conditional distribution of $Y - \beta_{01} - \beta_{02}^T X$ given $X \in \mathbb{R}^q$ is symmetric about 0. For the sake of clarity, we focus on the linear model and we assume that (Y, X) has a density with respect to the Lebesgue measure on $\mathbb{R} \times \mathcal{Q}$, with $\mathcal{Q} \subseteq \mathbb{R}^q$. Under classical regularity conditions, it is possible to include more general link functions in our analysis. We consider heteroscedasticity, i.e., when the conditional variance of the residual $Y - \beta_{01} - \beta_{02}^T X$ given X is not a constant. In this context, ordinary least squares are not efficient whereas weighted least squares might improve the estimation.

Let $(Y, X), (Y_1, X_1), (Y_2, X_2), \ldots$ denote an independent and identically distributed sequence of random variables satisfying model (19). Let P denote the distribution of (Y, X). The probability measure associated to the whole sequence $(Y_1, X_1), (Y_2, X_2), \ldots$ is still denoted by P_∞. The class of weighted estimators is given by $\widehat{\beta}(w)$, defined by

$$\widehat{\beta}(w) = \operatorname{argmin}_{(\beta_1,\beta_2)\in\mathbb{R}^{1+q}} n^{-1} \sum_{i=1}^{n} \rho\left(|Y_i - \beta_1 - \beta_2^T X_i|\right) w(X_i),$$

where $\rho : \mathbb{R}^+ \to \mathbb{R}^+$ is convex positive and differentiable and $w : \mathcal{Q} \to \mathbb{R}$ is called the weight function. Such a class of estimators is studied in Huber (1967), where a special attention is drawn on robustness properties associated to the choice of ρ. Note that when $\rho(x) = x^2$, we obtain weighted least squares, when $\rho(x) = x$, we get weighted median regression, $\rho(x) = (x^2/2)\mathbb{1}_{\{0 \le x \le c\}} + c(x - c/2)\mathbb{1}_{\{x > c\}}$ corresponds Huber's weighted robust regression (where c needs to be chosen in a proper way). Finally quantile regression estimators and $L_r(\mathbb{P}_n)$-losses estimators are as well included in this class.

We consider three approaches to estimate the optimal weight function w_0. Each approach is associated to a certain rate of convergence for $\widehat{w}$. The first one is parametric, i.e., w_0 is supposed to be in a given class of functions indexed by a Euclidean parameter. The second one is nonparametric, i.e., w_0 needs to satisfy some regularity conditions. The third one is called semiparametric and realizes a compromise between both previous approaches.

It is an exercise to verify each condition of Theorem 6. Here we focus on the special conditions dealing with the estimator $\widehat{w}$ of w_0, namely conditions (b4')(ii) and (b5). The other conditions are more classical and have been examined in different contexts (Newey and McFadden, 1994).

4.1. Minimum variance weights

A first question is to know, whether or not, such a class of estimators possesses a member with munimum variance. The answer is provided in Bates and White (1993) where the existence of a minimal variance estimator is studied. Basically, optimal members must satisfy the equation: "the variance of the score equals the Jacobian of the expected score" (as maximum likelihood estimators). Let $\epsilon = Y - \beta_{01} - \beta_{02}^T X$ denote the residual of the regression. The optimal weight function is

$$x \mapsto \frac{2\rho'(0) f_{\epsilon|X=x}(0) + E(g_{2,\beta_0}(Y,X)|X=x)}{E(g_{1,\beta_0}(Y,X)|X=x)},$$

where $g_{1,\beta}(y,x) = \rho'(|y - \beta_1 - \beta_2^T x|)^2$, $g_{2,\beta}(y,x) = \rho''(|y - \beta_1 - \beta_2^T x|)$ and $f_{\epsilon|X}$ is the conditional density of ϵ given X. In what follows, we consider the case where $\rho'(0) = 0$, so that w_0 simplifies to

$$w_0(x) = \frac{N_{\beta_0}(x)}{D_{\beta_0}(x)},$$

where $N_\beta(x) = E(g_{2,\beta}(Y,X)|X=x)f(x)$, $D_\beta = E(g_{1,\beta}(Y,X)|X=x)f(x)$ and f is the density of X. Concerning the examples cited above, this restriction only drops out quantile regression estimators.

A first estimator that can to be computed is $\widehat{\beta}^{(0)} = (\widehat{\beta}_1^{(0)}, \widehat{\beta}_2^{(0)})$, defined as $\widehat{\beta}(w)$ with constant weight function, $w(x) = 1$ for every $x \in \mathcal{Q}$. Even if $\widehat{\beta}^{(0)}$ is not efficient, it is well known that it is consistent for the estimation of β_0. Since w_0 depends on β_0, we use $\widehat{\beta}^{(0)}$ as a first-step estimator to carry on the estimation of w_0.

4.2. Parametric estimation of the weights

In this paragraph, we assume that $w_0(x) = w(x, \gamma_0)$, where γ_0 belongs to some Euclidean space. Typically, γ_0 is a vector containing β_0. Such a situation has been extensively studied (see Carroll et al. (1988) and the reference therein), and it has been shown under quite general conditions that γ_0 can be estimated consistently. Consequently, we assume in the next lines that there exists $\widehat{\gamma}$ such that $\widehat{\gamma} \to \gamma_0$, in P_∞^o-probability. The estimator of $w_0(x)$ is given by $w(x, \widehat{\gamma})$, for every $x \in \mathcal{Q}$, and β_0 is estimated by

$$\widehat{\beta} = \text{argmin}_{(\beta_1, \beta_2) \in \mathbb{R}^{1+q}} \, n^{-1} \sum_{i=1}^{n} \rho \left(|Y_i - \beta_1 - \beta_2^T X_i| \right) w(X_i, \widehat{\gamma}).$$

To verify (b4')(ii) and (b5), it is enough to ask the following Lipschitz condition. For every $x \in \mathcal{Q}$, we have

$$|w(x, \gamma) - w(x, \widetilde{\gamma})| \leq |\gamma - \widetilde{\gamma}|. \tag{20}$$

On the one hand, (b5) holds trivially with $\mathcal{W}_n$ equal to the class $\{x \mapsto w(x, \gamma) : |\gamma - \gamma_0| < \delta\}$, for every $\delta > 0$. On the other hand, (b4')(ii) is satisfied because the previous class has a $(\epsilon, \|\cdot\|_\infty)$-bracketing number of the same order as the $(\epsilon, |\cdot|)$-covering number of the Euclidean ball of radius δ (van der Vaart and Wellner, 1996, Theorem 2.7.11). Obviously, condition (20) is sufficient but not necessary. Another interesting example is $w(x, \beta, \gamma) = (1 + 1_{\{\beta_2^T x \leq \gamma\}})^{-1}$, reminiscent of a piecewise heteroscedastic model.

Within the context of linear regression given by (19), the parametric modelling of w_0 has serious drawbacks. Since a linear form is already assumed for the conditional mean, it is very restrictive, in addition, to parametrize the optimal weight function. It is even unnecessary as Theorem 6 does not ask for any rate of convergence estimating w_0. Finally, the definition of w_0, as a complicated quotient of conditional expectations, makes difficult for the analyst the setting of a plausible parametric family for w_0.

4.3. Nonparametric estimation of the weights

In this subsection, we consider the bracketing metric entropy generated by nonparametric Nadaraya-Watson estimators. The classical approach taken for local polynomial estimators relies on the asymptotic smoothness of such estimators (Ojeda, 2008; Portier and Segers, 2015). In the Nadaraya-Watson case, this smoothness approach can not succeed for compactly supported design. Due to the inconsistency of the Nadaraya-Watson estimator at boundary points, smoothness of the estimator is not inherited from the smoothness of the targeted function. Here, we handle the Nadaraya-Watson case by studying the bias and the variance separately. More precisely, we write the numerator $\widehat{N}(x)$ as $E_\infty[\widehat{N}(x)] + \widehat{\Delta}_N(x)$. It turns out that the class drawn by $x \mapsto E_\infty[\widehat{N}(x)]$, which is not random, has a smaller bracketing metric entropy than the function class

generated by $x \mapsto \widehat{\Delta}_N(x)$, which is included, as n increases and under reasonable conditions, in a smooth class of functions. The denominator is treated similarly.

For any differentiable function $f : \mathcal{Q} \subset \mathbb{R}^q \to \mathbb{R}$ and any $l = (l_1, \dots, l_q) \in \mathbb{N}^q$, let $f^{(l_1,\dots,l_q)}$ abbreviates $\frac{\partial^{|l|}}{\partial x_1^{l_1} \dots \partial x_q^{l_q}} f$, where $|l| = \sum_{j=1}^q l_j$. For $k \in \mathbb{N}$, $0 < \alpha \leq 1$ and $M > 0$, we say that $f \in \mathcal{C}_{k+\alpha,M}(\mathcal{Q})$ if, for every $|l| \leq k$, $f^{(l)}$ exists and is bounded by M on $\mathcal{Q}$ and, for every $|l| = k$ and every $(x, x') \in \mathcal{Q}^2$, we have

$$|f^{(l)}(x) - f^{(l)}(x')| \leq M|x - x'|^\alpha.$$

We define the estimator by

$$\widehat{w}(x) = \frac{\widehat{N}(x)}{\widehat{D}(x)},$$

where

$$\widehat{N}(x) = n^{-1} \sum_{i=1}^n g_{2,\widehat{\beta}^{(0)}}(Y_i, X_i) K_{h_n}(x - X_i),$$

$$\widehat{D}(x) = n^{-1} \sum_{i=1}^n g_{1,\widehat{\beta}^{(0)}}(Y_i, X_i) K_{h_n}(x - X_i),$$

$K_h(\cdot) = h^{-q} K(\cdot/h)$ and $(h_n)_{n \geq 1}$ is a sequence of bandwidths that goes to 0 as n goes to $+\infty$. We require the following set of assumptions.

(c1) The first step estimator is consistent, i.e., $\widehat{\beta}^{(0)} \overset{P^o_\infty}{\to} \beta_0$.

(c2) The density f of X is supported on a bounded convex set with nonempty interior $\mathcal{Q} \subset \mathbb{R}^q$ and there exists $b > 0$ such that $\inf_{x \in \mathcal{Q}} D_0(x) \geq b > 0$.

(c3) The map $x \mapsto D_0(x)$ is uniformly continuous on $\mathcal{Q}$. There exist $0 < \alpha_2 \leq 1$, $M_1 > 0$ and $\mathcal{B}_0 \subset \mathbb{R}^q$, an open ball centred at β_0, such that for any $x \in \mathbb{R}^q$, the maps $\beta \mapsto N_\beta(x)$ and $\beta \mapsto D_\beta(x)$ belongs to $\mathcal{C}_{\alpha_2,M_1}(\mathcal{B}_0)$. Moreover, the classes $\{(x, y) \mapsto g_{k,\beta}(y, x) \,:\, \beta \in \mathcal{B}_0\}$, $k = 1, 2$, are bounded measurable VC classes (we use the same terminology as in Giné and Guillou (2002), including the measurability requirements).

(c4) Let $K : \mathbb{R}^q \to \mathbb{R}$ be a bounded measurable function with compact support. There exists $h_0 > 0$ such that, for every $x \in \mathcal{Q}$ and $0 < h \leq h_0$,

$$\int K(u)du = 1, \qquad \int_{\{(\mathcal{Q}-x)/h\}} K(u)du \geq c > 0.$$

Moreover, there exists $k_1 \in \mathbb{N}$ such that for each $|l| \leq k_1 + 1$, the class

$$\left\{ K^{(l)}\left(\frac{x - \cdot}{h}\right) \,:\, h > 0,\, x \in \mathbb{R}^q \right\} \text{ is a bounded measurable } VC \text{ class.}$$

(c5) There exists $0 < \alpha_1 \leq 1$ such that, as $n \to +\infty$,

$$h_n \to 0, \qquad \frac{n h_n^{q+2(k_1+\alpha_1)}}{|\log(h_n)|} \to +\infty.$$

Let $\mathcal{V}(I)$ denote the set of all the functions that take their values in the set I. Define

$$\mathcal{F}_n = \mathcal{A}_{1,n}/\mathcal{A}_{2,n},$$

where

$$\begin{aligned}
\mathcal{A}_{1,n} &= \{\mathcal{C}_{k_1+\alpha_1,M_1}(\mathcal{Q}) + \mathcal{E}_{N,n}\} \cap \mathcal{V}[-M_2, M_2],\\
\mathcal{A}_{2,n} &= \{\mathcal{C}_{k_1+\alpha_1,M_1}(\mathcal{Q}) + \mathcal{E}_{D,n}\} \cap \mathcal{V}[cb/2, M_2],\\
\mathcal{E}_{N,n} &= \left\{x \mapsto \int N_\beta(x - h_n u)K(u)du \,:\, \beta \in \mathcal{B}_0\right\},\\
\mathcal{E}_{D,n} &= \left\{x \mapsto \int D_\beta(x - h_n u)K(u)du \,:\, \beta \in \mathcal{B}_0\right\},
\end{aligned}$$

and $M_2 = 2M_1 \int |K(u)|du$.

Theorem 7. *If (c1) to (c5) hold, we have*

$$\begin{aligned}
&P^o_\infty(\widehat{w} \in \mathcal{F}_n) \to 1,\\
&\log \mathcal{N}_{[\,]}(\epsilon, \mathcal{F}_n, \|\cdot\|_\infty) \le \mathit{const.}\epsilon^{-q/(k_1+\alpha_1)}, \qquad \mathit{for\ any}\ \epsilon > 0,
\end{aligned}$$

where const. depends on q, $\mathcal{Q}$, $k_1 + \alpha_1$, M_1, b and K.

Proof. By (c1), we have that $\widehat{\beta}^{(0)} \in \mathcal{B}_0$ with probability going to 1. Let

$$\begin{aligned}
\widehat{\Delta}_N(x) &= \widehat{N}(x) - E_\infty \widehat{N}(x),\\
\widehat{\Delta}_D(x) &= \widehat{D}(x) - E_\infty \widehat{D}(x).
\end{aligned}$$

We consider the following three steps.

(i) $P^o_\infty(\widehat{\Delta}_N \in \mathcal{C}_{k_1+\alpha_1,M_1}(\mathcal{Q})) \to 1$ and $P^o_\infty(\widehat{\Delta}_D \in \mathcal{C}_{k_1+\alpha_1,M_1}(\mathcal{Q})) \to 1$,
(ii) $P^o_\infty(\widehat{N} \in \mathcal{A}_{1,n}) \to 1$ and $P^o_\infty(\widehat{D} \in \mathcal{A}_{2,n}) \to 1$ (note that this is the first claim of the theorem).
(iii) Compute the bound on the bracketing numbers of $\mathcal{F}_n$.

Proof of (i). We make the proof for $\widehat{\Delta}_N$ since the treatment of $\widehat{\Delta}_D$ is similar. Given $l = (l_1, \ldots, l_q)$ such that $|l| \le k_1 + 1$, we have,

$$\partial^{(l)} E_\infty[\widehat{N}(x)] = h_n^{-(q+|l|)} E[g_{1,\widehat{\beta}^{(0)}}(Y, X)K^{(l)}(h_n^{-1}(x - X))].$$

Hence

$$\begin{aligned}
&\widehat{\Delta}_N^{(l)}(x) = \frac{1}{nh_n^{q+|l|}} \times\\
&\sum_{i=1}^n \left\{g_{1,\widehat{\beta}^{(0)}}(Y_i, X_i)K^{(l)}(h_n^{-1}(x - X_i)) - E[g_{1,\widehat{\beta}^{(0)}}(Y, X)K^{(l)}(h_n^{-1}(x - X))]\right\},
\end{aligned}$$

and, by (c3) and (c4), we can apply Lemma 8, stated in the Appendix, to get that

$$\sup_{x\in\mathcal{Q}} |\widehat{\Delta}_N^{(l)}(x)| = O_{P_\infty}\left(\sqrt{\frac{|\log(h_n)|}{nh_n^{q+2|l|}}}\right). \tag{21}$$

Then, for $1 \leq |l| \leq k_1$, we know that $\widehat{\Delta}_N^{(l)}$ goes to 0 uniformly over $\mathcal{Q}$, making the derivatives of $\widehat{\Delta}_N$ (with order smaller than or equal to k_1), bounded by M_1 with probability going to 1. Now we consider the Hölder property for $\widehat{\Delta}_N^{(l)}$ when $|l| = k_1$. For any $|x - x'| \leq h_n$, by the mean value theorem, we have that

$$\begin{aligned}|(\widehat{\Delta}_N^{(l)}(x) - \widehat{\Delta}_N^{(l)}(x'))(x-x')^{-\alpha_1}| &\leq |x-x'|^{1-\alpha_1} \sup_{z\in\mathcal{Q}} |\nabla_z \widehat{\Delta}_N^{(l)}(z)| \\ &\leq h_n^{1-\alpha_1} \sup_{z\in\mathcal{Q}} |\nabla_z \widehat{\Delta}_N^{(l)}(z)|,\end{aligned}$$

which is, in virtue of (21), equal to a $O_{P_\infty}\left(\sqrt{\frac{|\log(h_n)|}{nh_n^{q+2(k_1+\alpha_1)}}}\right) = o_P(1)$. For any $|x - x'| > h$, we have

$$|(\widehat{\Delta}_N^{(l)}(x) - \widehat{\Delta}_N^{(l)}(x'))(x-x')^{-\alpha_1}| \leq 2h_n^{-\alpha_1} \sup_{z\in\mathcal{Q}} |\widehat{\Delta}_N^{(l)}(z)|,$$

which has the same order as the previous term. As a consequence, for $|l| = k_1$, we have shown that

$$\sup_{x\neq x'} |(\widehat{\Delta}_N^{(l)}(x) - \widehat{\Delta}_N^{(l)}(x'))(x-x')^{-\alpha_1}| = o_{P_\infty}(1),$$

implying that $\widehat{\Delta}_N^{(l)}$ is α_1-Holder (with constant M_1) with probability going to 1.

Proof of (ii). For the first statement, using (i), it suffices to show that $\widehat{N}$ lies in $\mathcal{V}[-M_2, M_2]$ with probability going to 1. Lemma 8 and condition (c3) yield

$$\begin{aligned}|\widehat{N}(x)| &\leq |E_\infty \widehat{N}(x)| + \sup_{x\in\mathcal{Q}} |\widehat{\Delta}_N(x)| \\ &= |\int N_{\widehat{\beta}^{(0)}}(x - h_n u)K(u)du| + o_p(1) \\ &\leq M_1 \int |K(u)|du + o_p(1) \\ &\leq 2M_1 \int |K(u)|du, \qquad \text{with probability going to 1.}\end{aligned}$$

For the second statement, it suffices to show that $\widehat{D}$ lies in $\mathcal{V}[cb/2, M_2]$ with probability going to 1. To obtain the upper bound for this class, we mimic what has been done above to treat $\widehat{N}$. To obtain the lower bound, first write

$$E_\infty[\widehat{D}(x)] - (D_0 * K_{h_n})(x) = \int (D_{\widehat{\beta}^{(0)}}(x - h_n u) - D_{\beta_0}(x - h_n u))K(u)du,$$

by condition (c3), it goes to 0, uniformly over $x \in \mathcal{Q}$, in probability, . This yields

$$\begin{aligned}\widehat{D}(x) &= (D_0 * K_{h_n})(x) + E_\infty[\widehat{D}(x)] - (D_0 * K_{h_n})(x) + \widehat{\Delta}_D(x) \\ &\geq (D_0 * K_{h_n})(x) - \sup_{x\in\mathcal{Q}} |E_\infty[\widehat{D}(x)] - (D_0 * K_{h_n})(x)| - \sup_{x\in\mathcal{Q}} |\widehat{\Delta}_D(x)| \\ &= (D_0 * K_{h_n})(x) - o_{P_\infty}(1).\end{aligned}$$

Define $b(x,h) = \inf_{y\in\mathcal{Q},\,|y-x|\leq hA} D_0(y)$ and $M(x,h) = \sup_{y\in\mathcal{Q},\,|y-x|\leq hA} D_0(y)$ where A is such that, for every $u \in \mathbb{R}^q$, $K(u)\mathbb{I}_{\{|u|>A\}} = 0$ (A is finite because K is compactly supported). Note that, by the uniform continuity of D_0, $\sup_{x\in\mathcal{Q}} |M(x,h) - b(x,h)| \to 0$ as $h \to 0$, it follows that

$$\begin{aligned}&(D_0 * K_{h_n})(x) \\ &= \int D_0(x + h_n u)K(u)du \\ &\geq b(x,h_n) \int \mathbb{I}_{\{x+h_n u\in\mathcal{Q}\}}\{K(u)\}_+ du + M(x,h_n) \int \mathbb{I}_{\{x+h_n u\in\mathcal{Q}\}}\{K(u)\}_- du \\ &= b(x,h_n) \int \mathbb{I}_{\{x+h_n u\in\mathcal{Q}\}} K(u)du \\ &\qquad + (M(x,h_n) - b(x,h_n)) \int \mathbb{I}_{\{x+h_n u\in\mathcal{Q}\}}\{K(u)\}_- du \\ &\geq b(x,h_n) \int \mathbb{I}_{\{x+h_n u\in\mathcal{Q}\}} K(u)du - o(1) \\ &\geq b \int_{\{(\mathcal{Q}-x)/h_n\}} K(u)du - o(1),\end{aligned}$$

which is greater than $cb/2 > 0$, whenever n is large enough, by (c4).

Proof of (iii). We now bound the $(\epsilon, \|\cdot\|_\infty)$-bracketing number of $\mathcal{F}_n$. First, Corollary 2.7.2, page 157, in van der Vaart and Wellner (1996) states that, for every $\epsilon > 0$,

$$\log \mathcal{N}_{[\,]}(\epsilon, \mathcal{C}_{k_1+\alpha_1,M_1}(\mathcal{Q}), \|\cdot\|_\infty) \leq \text{const.}\epsilon^{-q/(k_1+\alpha_1)},$$

where const. depends only on q, $\mathcal{Q}$, $k_1 + \alpha_1$ and M_1. Second, by assumption (c3),

$$\begin{aligned}&\left|\int \left(N_\beta(x - h_n u) - N_{\beta'}(x - h_n u)\right) K(u)du\right| \leq |\beta - \beta'|^{\alpha_2} M_1 \int |K(u)|du, \\ &\left|\int \left(D_\beta(x - h_n u) - D_{\beta'}(x - h_n u)\right) K(u)du\right| \leq |\beta - \beta'|^{\alpha_2} M_1 \int |K(u)|du.\end{aligned}$$

This makes the classes $\mathcal{E}_{N,n}$ and $\mathcal{E}_{D,n}$ being α_2-Hölder in the index parameter. Hence, from Theorem 2.7.11 in van der Vaart and Wellner (1996), their $(\epsilon, \|\cdot\|_\infty)$-bracketing numbers are smaller than the $(\epsilon, |\cdot|)$-covering number of $\mathcal{B}_0$ (up to some multiplicative constant in the ϵ). It is also smaller than the $(\epsilon, \|\cdot\|_\infty)$-bracketing numbers of $\mathcal{C}_{k_1+\alpha_1,M_1}(\mathcal{Q})$, given previously. Therefore, since the class

$\mathcal{A}_{1,n}$ (resp. $\mathcal{A}_{2,n}$) coincides with the set $\mathcal{C}_{k_1+\alpha_1,M_1}(\mathcal{Q})$ plus $\mathcal{E}_{N,n}$ (resp. $\mathcal{E}_{D,n}$), we conclude that, for every $\epsilon > 0$,

$$\log \mathcal{N}_{[]}(\epsilon, \mathcal{A}_{j,n}, \|\cdot\|_\infty) \leq \text{const.}\epsilon^{-q/(k_1+\alpha_1)},$$

for $j = 1, 2$, where const. depends only on q, $\mathcal{Q}$, $k_1 + \alpha_1$, M_1 and K. Now we show that the previous bound is still valid for the class $\mathcal{F}_n = \mathcal{A}_{1,n}/\mathcal{A}_{2,n}$. Let $[\underline{N}_1, \overline{N}_1], \ldots, [\underline{N}_{n_1}, \overline{N}_{n_1}]$ (resp. $[\underline{D}_1, \overline{D}_1], \ldots, [\underline{D}_{n_2}, \overline{D}_{n_2}]$) be $(\epsilon, \|\cdot\|_\infty)$-brackets that cover $\mathcal{A}_{1,n}$ (resp. $\mathcal{A}_{2,n}$). By taking $\underline{D}_1 \vee b, \ldots, \underline{D}_{n_2} \vee b$ in place of $\underline{D}_1, \ldots, \underline{D}_{n_2}$, we can assume that the elements of the brackets of $\mathcal{A}_{2,n}$ are larger than or equal to b. By a similar argument, every brackets of $\mathcal{A}_{1,n}$ are bounded by M_2. Hence, for any $N \in \mathcal{A}_{1,n}$ and $D \in \mathcal{A}_{2,n}$, there exists $1 \leq i \leq n_1$ and $1 \leq j \leq n_2$, such that

$$\frac{\underline{N}_i}{\overline{D}_j} \leq \frac{N}{D} \leq \frac{\overline{N}_i}{\underline{D}_j},$$
$$\left\| \frac{\underline{N}_i}{\overline{D}_j} - \frac{\overline{N}_i}{\underline{D}_j} \right\|_\infty \leq \text{const.}\epsilon,$$

where const. is a constant that depends only on b and M_2. As a consequence we have exhibited a $(\text{const.}\epsilon, \|\cdot\|_\infty)$-bracketing with $n_1 n_2$ elements, yielding to the statement of the theorem. □

Remark 10. *On the one hand, no strong assumptions are imposed on the regularity of the targeted functions $x \mapsto N_0(x)$ and $x \mapsto D_0(x)$. Actually, we only require the uniform continuity of $x \mapsto D_0(x)$ to hold. The reason is that we do not use the consistency of $\widehat{N}(x)$ (resp. $\widehat{D}(x)$) estimating $N_0(x)$ (resp. $D_0(x)$). On the other hand, the kernel needs to be many times differentiable. Hence, our approach consists of approximating a function, non necessarily regular, by a smooth function. In this way, we control the bracketing metric entropy generated by the class of estimated functions. In light of Theorem 7, assumption (b4')(ii) is satisfied when $k_1 + \alpha_1 > q/2$ and when the bandwidth sequence $h_n \to 0$ such that $nh_n^{q+2(k_1+\alpha_1)}/|\log(h_n)| \to +\infty$. Provided the kernel function is smooth enough, one can put $k_1 + \alpha_1 = q/2 + \delta/4$, for some $\delta > 0$. Then, it suffices to choose the bandwidth such that $nh_n^{2q+\delta} \to +\infty$.*

Remark 11. *Another way to proceed is to consider the classes*

$$\mathcal{E}_N = \left\{ x \mapsto \int N_\beta(x - hu)K(u)du \,:\, \beta \in \mathcal{B}_0,\, h > 0 \right\},$$
$$\mathcal{E}_D = \left\{ x \mapsto \int D_\beta(x - hu)K(u)du \,:\, \beta \in \mathcal{B}_0,\, h > 0 \right\},$$

in place of $\mathcal{E}_{N,n}$ and $\mathcal{E}_{D,n}$. These classes are larger but they no longer depend on n. To calculate the bracketing entropy of the spaces $\mathcal{E}_N$ and $\mathcal{E}_D$, one might consider the $L_r(P)$-metric rather than the uniform metric because the latter would involve some difficulties at the boundary points.

Remark 12. *Examples of kernels that satisfy (c4) are given in Nolan and Pollard (1987), Lemma 22, see also Giné and Guillou (2002). An interesting fact is that* $\{\widetilde{K}((x-\cdot)/h) : h>0, x\in\mathbb{R}\}$ *is a uniformly bounded VC class of functions, when* $\widetilde{K}$ *has bounded variation. The assumption that* $\int_{\{(\mathcal{Q}-x)/h\}} K(u)du \geq c>0$, *for* $h_0 \geq h > 0$, *holds true if* $\mathcal{Q}$ *is a smooth surface, i.e., when the distance between* $x\in\mathbb{R}^q$ *and mathcalQ is a differentiable function of* x. *Note also that in the one-dimensional case, it is always verified. Moreover, this condition permits to include the case of non-smooth surfaces such as cubes.*

4.4. Semiparametric estimation of the weights

The nonparametric approach involves a smoothing in the space $\mathbb{R}^q$. It is well known that the smaller the dimension q, the better the estimation. Although it does not affect the asymptotic variance of $\widehat{\beta}(\widehat{w})$ (no specific rate of convergence of $\widehat{w}$ to w_0 is required in (b5)(ii)), it certainly influences the small sample size performances of the estimators.

There exist different ways to introduce a semiparametric procedure to estimate w_0. In the following, we rely on the single index approach. In our initial regression model (19), the conditional mean of Y given X depends only on $\beta_{02}^T X$. Given this, it is slightly stronger to ask that the conditional law of Y given X is equal to the conditional law of Y given $\beta_{02}^T X$, in other words, that

$$Y \perp X | \beta_{02}^T X, \tag{22}$$

or equivalently that,

$$E(g(Y)|X) = E(g(Y)|\beta_{02}^T X),$$

for every bounded measurable function g. Such an assumption has been introduced in Li (1991) to estimate the law of Y given X. Here (22) is introduced in a different spirit: since a linear regression model has already been imposed in (19), condition (22) appears as an additional mild requirement, that serves only the estimation of w_0. The calculation of semiparametric estimators of w_0 is done by using similar tools as in the previous section. In order to fully benefit from condition (22), we realize the smoothing in a low-dimensional subspace of $\mathbb{R}^q$. We define the estimator of w_0 by

$$\widehat{w}(x) = \frac{\widehat{N}_{\widehat{\beta}^{(0)}}(\widehat{\beta}_2^{(0)T} x)}{\widehat{D}_{\widehat{\beta}^{(0)}}(\widehat{\beta}_2^{(0)T} x)},$$

where, for every $t\in\mathbb{R}$,

$$\widehat{N}_\beta(t) = n^{-1}\sum_{i=1}^n g_{2,\beta}(Y_i, X_i) L_{h_n}(t - \beta_2^T X_i),$$

$$\widehat{D}_\beta(t) = n^{-1}\sum_{i=1}^n g_{1,\beta}(Y_i, X_i) L_{h_n}(t - \beta_2^T X_i),$$

with $L_h(\cdot) = h^{-1}L(\cdot/h)$. The proofs are more involved than in the nonparametric case notably because of the randomness of the space generated by $\widehat{\beta}_2^{(0)}$ on which the smoothing is realized.

5. Simulations

The asymptotic analysis conducted in the previous sections demonstrates that, in weighted linear regression, the estimation of w_0 does not matter provided its consistency, e.g., $\widehat{w}_1$ and $\widehat{w}_2$ might converge to w_0 with different rates, whereas $\widehat{\beta}(\widehat{w}_1)$ and $\widehat{\beta}(\widehat{w}_2)$ are asymptotically equivalent. Nevertheless when the sample size is not very large, differences might arise between the procedures. In the next we consider the three approaches investigated in the previous section, namely, parametric, nonparametric and semiparametric. Each of these procedures results in different rates of convergence of $\widehat{w}$ to w_0. Here the purpose is two folds. First to provide a clear picture of the small sample size performances of each method. Second, to analyse, from a practical point of view, the relaxation of the regularity conditions on w_0.

In what follows, we consider the following heteroscedastic linear regression model. Let (X, Y), $(X_i, Y_i)_{i=1,\dots,n}$ be independently and identically distributed random variables. Suppose that

$$Y = \beta_{01} + \beta_{02}^T X + \sigma_0(X)\epsilon,$$

where $(X, \epsilon) \in \mathbb{R}^{q+1}$ has a standard normal distribution and $(\beta_{01}, \beta_{02}) = (1, \dots, 1)/\sqrt{q+1}$. The weighted least square estimator is given by

$$\widehat{\beta}(w) = \widehat{\Sigma}(w)^{-1}\widehat{\gamma}(w), \tag{23}$$

with $\widehat{\Sigma}(w) = n^{-1}\sum_{i=1}^n \widetilde{X}_i\widetilde{X}_i^T w(X_i)$, $\widehat{\gamma}(w) = n^{-1}\sum_{i=1}^n \widetilde{X}_i Y_i w(X_i)$ and $\widetilde{X}_i^T = (1, X_i^T)$, for $i = 1, \dots, n$. Let $\widehat{\beta}^{(0)} = (\widehat{\beta}_1^{(0)}, \widehat{\beta}_2^{(0)})$, with $\widehat{\beta}_1^{(0)} \in \mathbb{R}$ and $\widehat{\beta}_2^{(0)} \in \mathbb{R}^q$, denote the coordinates of the first-step estimator with constant weights. For different sample sizes n and also several dimensions q, we consider two functions for σ_0: a smooth function and a not continuous function, respectively given by, for every $x \in \mathbb{R}^q$,

$$\sigma_{01}(x) = \frac{\beta_{02}^T x}{|\beta_{02}|} \qquad \text{and} \qquad \sigma_{02}(x) = \frac{1}{2} + 2 \cdot \mathbb{1}_{\{\beta_{02}^T x > 0\}}.$$

In each case, $k = 1, 2$, the optimal weight function $w_0 = 1/\sigma_{0k}^2$ is estimated by these methods:

(i) Parametric: $\widehat{w}_k$ is computed using $\widehat{\beta}^{(0)}$ in place of β_0 in the formula of w_{0k},
(ii) Nonparametric: Nadaraya-Watson procedure, $\widehat{w}_k$ is given by

$$\frac{\sum_{i=1}^n K_{h_n}(X_i - x)}{\sum_{i=1}^n (Y_i - \widehat{\beta}_1^{(0)} - \widehat{\beta}_2^{(0)T} X_i)^2 K_{h_n}(X_i - x)},$$

(iii) Semiparametric: Nadaraya-Watson procedure in a reduced sample-based space, $\widehat{w}_k$ is given by

$$\frac{\sum_{i=1}^n K_{h_n}(\widehat{A}(X_i - x))}{\sum_{i=1}^n (Y_i - \widehat{\beta}_1^{(0)} - \widehat{\beta}_2^{(0)T} X_i)^2 K_{h_n}(\widehat{A}(X_i - x))},$$

where $\widehat{A} = \widehat{P}_2^{(0)} + \epsilon I$ and $\widehat{P}_2^{(0)}$ denotes the orthogonal projector onto the linear space generated by $\widehat{\beta}_2^{(0)}$.

For (ii) and (iii), the kernel K is the Epanechnikov kernel given by $K(u) = c_q(1 - |u|^2)_+$, where c_q is such that $\int K(u)du = 1$. For $k = 1, 2$, $\widehat{w}_k$ is initially computed according to one of the method (i), (ii) or (iii), then the final estimator of β_0 is computed with $\widehat{\beta}(\widehat{w}_k)$ given by (23).

In practice, we find that choosing the bandwidth h_n by cross validation is reasonable. More precisely, considering the estimation of $\sigma_0^2(x)$ by $\widehat{\sigma}^2(x)$, it is defined by

$$h_{n,\mathrm{cv}} = \mathrm{argmin}_{h>0} \sum_{i=1}^n ((Y_i - \widehat{\beta}_1^{(0)} - \widehat{\beta}_2^{(0)T} X_i)^2 - \widehat{\sigma}^{2(-i)}(X_i))^2,$$

where $\widehat{\sigma}^{2(-i)}(x)$ is either the leave-one-out nonparametric estimator of $\sigma_0^2(x)$ given by (ii) or the leave-one-out semiparametric estimator of $\sigma_0^2(x)$ given by (iii). Such a data-driven algorithm for h_n has the advantage to select automatically the bandwidth without regard for the underlying dimension of the semi- and nonparametric estimators. In every examples, the semiparametric $h_{n,\mathrm{cv}}$ was smaller than the nonparametric $h_{n,\mathrm{cv}}$.

For the semiparametric method, the matrix $\widehat{A}$ denotes the orthogonal projector onto the space generated by $\widehat{\beta}_0$ perturbed by ϵ in the diagonal. This permits not to have a blind confidence in the first-step estimator $\widehat{\beta}_0$ accounting for variations of w_0 in the other directions. Hence ϵ is reasonably selected if ϵI has the same order as the error $\widehat{P}_2^{(0)} - P_2^{(0)}$, where $P_2^{(0)}$ is the orthogonal projector onto the linear space generated by $\beta_2^{(0)}$. We have

$$|\widehat{P}_2^{(0)} - P_2^{(0)}|_F^2 = 2\mathrm{trace}((I - P_2^{(0)})\widehat{P}_2^{(0)}) = \frac{2|(I - P_2^{(0)})\widehat{\beta}_2^{(0)}|^2}{|\widehat{\beta}_2^{(0)}|^2}.$$

The numerator is approximated by an estimator of the average value of its asymptotic law in the case where $\epsilon \perp\!\!\!\perp X$. It gives $2\widehat{\sigma}^2 n^{-1/2} \sum_{k=1}^q \widehat{\lambda}_k^2$ where $\widehat{\lambda}_k$ are the eigenvalues associated to the matrix $(I - \widehat{P}_2^{(0)})\widehat{\Sigma}_2^{-1}(I - \widehat{P}_2^{(0)})$, $\widehat{\sigma}^2 = n^{-1} \sum_{i=1}^n (Y_i - \widehat{\beta}_0 - \widehat{\beta}_0^T X_i)^2$ and Σ_2^{-1} denotes the $q \times q$ lower triangular block of the inverse of $n^{-1} \sum_{i=1}^n \widetilde{X}_i \widetilde{X}_i^T$. As a consequence, ϵ is given by

$$\epsilon = \sqrt{\frac{2\widehat{\sigma}^2 \sum_{k=1}^q \widehat{\lambda}_k^2}{nq|\widehat{\beta}|^2}},$$

where $\sqrt{q}$ appears as a normalizing constant.

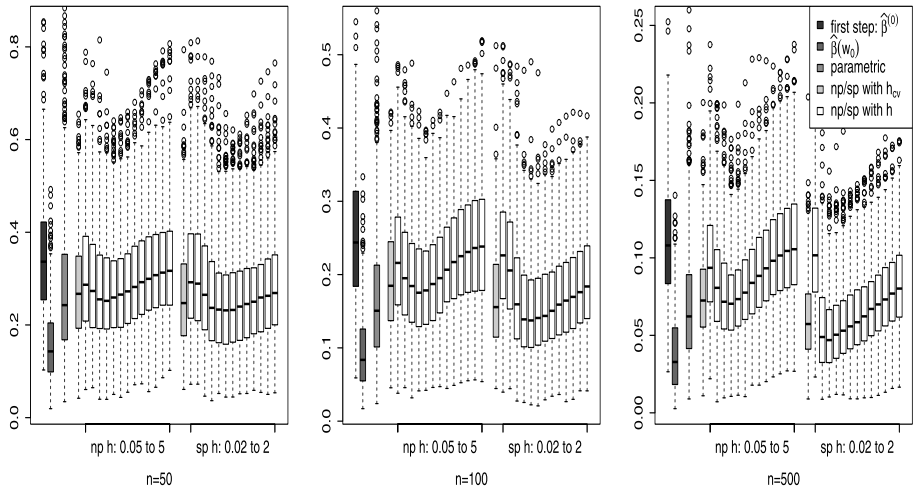

FIG 1. *Each boxplot is based on* 500 *estimates of* $|\widehat{\beta}-\beta_0|^2$ *when* $q=4$ *and* $\sigma_0(x)=\frac{\beta_{02}^T x}{|\beta_{02}|}$. *The parametric, nonparametric (np) and semiparametric (sp) approaches are respectively based on (i), (ii) and (iii).*

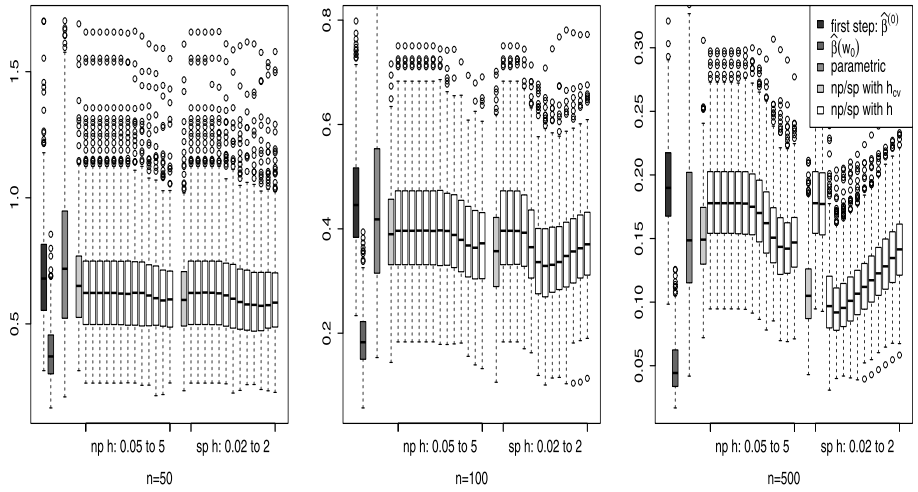

FIG 2. *Each boxplot is based on* 500 *estimates of* $|\widehat{\beta}-\beta_0|^2$ *when* $q=16$ *and* $\sigma_0(x)=\frac{\beta_{02}^T x}{|\beta_{02}|}$. *The parametric, nonparametric (np) and semiparametric (sp) approaches are respectively based on (i), (ii) and (iii).*

Figures 1 to 4 provide boxplots associated to the estimation error of each method, parametric (i), nonparametric (ii), and semiparamteric (iii), according to different values of $n = 50, 100, 500$, $q = 4, 16$ and $\sigma_0 = \sigma_{01}, \sigma_{02}$. We also consider the first-step estimator $\widehat{\beta}^{(0)}$ and a "reference estimator" computed with

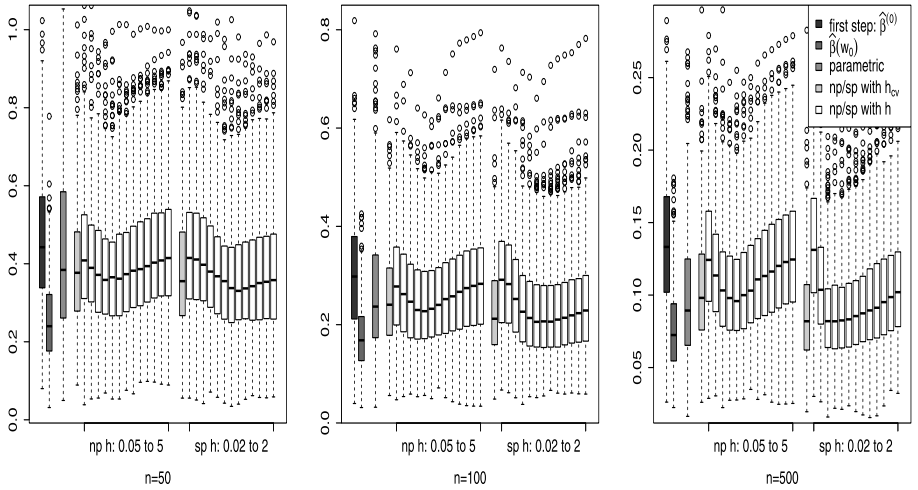

FIG 3. *Each boxplot is based on* 500 *estimates of* $|\widehat{\beta} - \beta_0|^2$ *when* $q = 4$ *and* $\sigma_0(x) = \frac{1}{2} + 2 \cdot \mathbb{1}_{\{\beta_{02}^T x > 0\}}$. *The parametric, nonparametric (np) and semiparametric (sp) approaches are respectively based on (i), (ii) and (iii).*

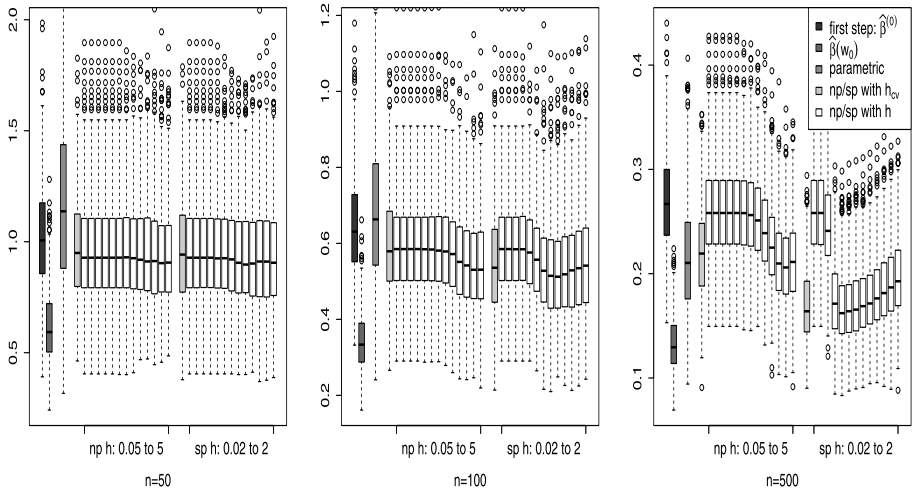

FIG 4. *Each boxplot is based on* 500 *estimates of* $|\widehat{\beta} - \beta_0|^2$ *when* $q = 16$ *and* $\sigma_0(x) = \frac{1}{2} + 2 \cdot \mathbb{1}_{\{\beta_{02}^T x > 0\}}$. *The parametric, nonparametric (np) and semiparametric (sp) approaches are respectively based on (i), (ii) and (iii).*

the unknown optimal weights, i.e., $\widehat{\beta}(w_0)$. In every case, the accuracy of each method lies between the first step estimator and the reference estimator. In agreement with Theorems 6 and 7, the gap between the reference estimator and the method (i), (ii), (iii) diminishes as n increases. Each method (i), (ii), (iii),

performs differently showing that their equivalence occurs only at very large sample size.

Among the three methods under evaluation (i), (ii), (iii), the clear winner is the semiparametric method (with selection of the bandwidth by cross validation). The fact that it over-rules the nonparametric estimator was somewhat predictable, but the difference in accuracy with the parametric method is surprising. In every situation, the variance and the mean of the error associated to the semiparametric approach are smaller than the variance and the mean of the others. Moreover, the nonparametric method performs as well as the parametric method even when the dimension is large. In fact, both approaches are similarly affected by the increase of the dimension. Finally, one sees that the choice of the bandwidth by cross validation works well for both methods nonparametric and semiparametric. In all cases, the estimator with $h_{n,\mathrm{cv}}$ performs similarly to the estimator with the optimal value of h_n.

Appendix: Concentration rates for kernel regression estimators

The following result follows from the formulation of the Talagrand inequality (Talagrand, 1994) given in Theorem 2.1 in Giné and Guillou (2002).

Lemma 8. *Let $(Y \in \mathbb{R}, X \in \mathbb{R}^q), (Y_1, X_1), (Y_2, X_2), \ldots$ denote a sequence of random variables independently and identically distributed such that X has a bounded density f. Let $\widetilde{K} : \mathbb{R}^q \to \mathbb{R}$ be a bounded square integrable measurable function and Ψ be a class of real-valued measurable functions defined on $\mathbb{R}^{q+1}$. If both classes Ψ and $\{\widetilde{K}((x-\cdot)/h) : x \in \mathbb{R}^q, h > 0\}$ are bounded measurable VC classes, then, for any sequence $h_n \to 0$ such that $nh_n^q/|\log(h_n)| \to +\infty$, we have, as $n \to +\infty$,*

$$\sup_{\psi\in\Psi,\, x\in\mathbb{R}^q} \left| \frac{1}{n}\sum_{i=1}^{n} \left\{\psi(Y_i, X_i)\widetilde{K}_{h_n}(x - X_i) - E[\psi(Y, X)\widetilde{K}_{h_n}(x - X)]\right\} \right| = O_{P_\infty}\left(\sqrt{\frac{|\log(h_n)|}{nh_n^q}}\right),$$

where $K_h(\cdot) = K(\cdot/h)/h^q$.

Proof. The empirical process to consider is indexed by the product class $\Psi \times \{\widetilde{K}((x-\cdot)/h_n), x \in \mathbb{R}^q\}$, which is uniformly bounded VC since the product of two uniformly bounded VC classes remains uniformly bounded VC. The variance satisfies

$$\begin{aligned}\mathrm{var}(\psi(Y, X)\widetilde{K}(h_n^{-1}(x - X)) &\le E\left(\psi(Y, X)^2\widetilde{K}(h_n^{-1}(x - X))^2\right)\\ &\le \|\psi\|_\infty^2\|f\|_\infty h_n^q \int \widetilde{K}(u)^2 du,\end{aligned}$$

and a uniform bound is given by $\|\psi\widetilde{K}\|_\infty \le \sup_{\psi\in\Psi}\|\psi\|_\infty\|\widetilde{K}\|_\infty$. The application of Theorem 2.1 in Giné and Guillou (2002) gives the specified bound. □

References

Akritas, M. G. and Van Keilegom, I. (2001). Non-parametric estimation of the residual distribution. *Scand. J. Statist. 28*(3), 549–567. MR1858417

Andrews, D. W. K. (1994). Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica 62*(1), 43–72. MR1258665

Bates, C. E. and White, H. (1993). Determination of estimators with minimum asymptotic covariance matrices. *Econometric Theory 9*(04), 633–648. MR1254303

Carroll, R. J. (1982). Adapting for heteroscedasticity in linear models. *Ann. Statist. 10*(4), 1224–1233. MR0673657

Carroll, R. J. and Ruppert, D. (1982). Robust estimation in heteroscedastic linear models. *Ann. Statist. 10*(2), 429–441. MR0653518

Carroll, R. J., Wu, C.-F. J., and Ruppert, D. (1988). The effect of estimating weights in weighted least squares. *J. Amer. Statist. Assoc. 83*(404), 1045–1054. MR0997580

Chen, X., Linton, O., and Van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica 71*(5), 1591–1608. MR2000259

Giné, E. and Guillou, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Ann. Inst. H. Poincaré Probab. Statist. 38*(6), 907–921. En l'honneur de J. Bretagnolle, D. Dacunha-Castelle, I. Ibragimov. MR1955344

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66), Vol. I: Statistics*, pp. 221–233. Univ. California Press, Berkeley, Calif. MR0216620

Kato, K. (2009). Asymptotics for argmin processes: Convexity arguments. *J. Multivariate Anal. 100*(8), 1816–1829. MR2535389

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference*. Springer Series in Statistics. New York: Springer. MR2724368

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc. 86*(414), 316–342. MR1137117

Newey, W. K. (1990). Efficient instrumental variables estimation of nonlinear models. *Econometrica 58*(4), 809–837. MR1064846

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica 62*(6), 1349–1382. MR1303237

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of econometrics, Vol. IV*, Volume 2 of *Handbooks in Econom.*, pp. 2111–2245. Elsevier. MR1315971

Nolan, D. and Pollard, D. (1987). U-processes: Rates of convergence. *Ann. Statist. 15*(2), 780–799. MR0888439

Ojeda, J. (2008). Hölder continuity properties of the local polynomial estimator. *Pre-publicaciones del Seminario Matemático García de Galdeano* no. 4, 1–21.

Pollard, D. (1985). New ways to prove central limit theorems. *Econometric Theory 1*(03), 295–313.

Portier, F. and Delyon, B. (2013). Optimal transformation: A new approach for covering the central subspace. *Journal of Multivariate Analysis 115*, 84–107. MR3004547

Portier, F. and Segers, J. (2015). On the weak convergence of the empirical conditional copula under a simplifying assumption. *arXiv preprint arXiv:1511.06544*.

Robinson, P. M. (1987). Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form. *Econometrica 55*(4), 875–891. MR0906567

Talagrand, M. (1994). Sharper bounds for Gaussian and empirical processes. *Ann. Probab. 22*(1), 28–76. MR1258865

van der Vaart, A. W. (1995). Efficiency of infinite-dimensional M-estimators. *Statist. Neerlandica 49*(1), 9–30. MR1333176

van der Vaart, A. W. (1998). *Asymptotic statistics*, Volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge: Cambridge University Press. MR1652247

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. New York: Springer-Verlag. With applications to statistics. MR1385671

van der Vaart, A. W. and Wellner, J. A. (2007). Empirical processes indexed by estimated functions. In *Asymptotics: Particles, processes and inverse problems*, Volume 55 of *IMS Lecture Notes Monogr. Ser.*, pp. 234–252. Beachwood, OH: Inst. Math. Statist. MR2459942