INSTITUT NATIONAL DE LA STATISTIQUE ET DES ETUDES ECONOMIQUES
Série des Documents de Travail du CREST
(Centre de Recherche en Economie et Statistique)

n° 2004-17

On Semiparametric $M-$estimation in Single-Index Regression

M. DELECROIX[1]
M. HRISTACHE[2]
V. PATILEA[3]

Les documents de travail ne reflètent pas la position de l'INSEE et n'engagent que leurs auteurs.

Working papers do not reflect the position of INSEE but only the views of the authors.

[1] CREST-ENSAI. Email : delecroi@ensai.fr

[2] ENSAI and CREST, Campus de Ker Lann, Rue Blaise Pascal, BP 37203, 35172 Bruz Cedex, France. Email : hristach@ensai.fr

[3] CREST-ENSAI, Campus de Ker Lann, Rue Blaise Pascal, BP 37203, 35172 Bruz Cedex, France. Email : valentin.patilea@ensai.fr
Part of the work for this paper was accomplished while this author was at LEO, Université d'Orléans.

On semiparametric $M-$estimation in single-index regression

Michel Delecroix[*], Marian Hristache[†] and Valentin Patilea [‡]

August 24, 2004

[*] CREST-ENSAI. Email: delecroi@ensai.fr

[†] ENSAI and CREST, Campus de Ker Lann, Rue Blaise Pascal, BP 37203, 35172 Bruz Cedex, France. Email: hristach@ensai.fr

[‡] CREST-ENSAI, Campus de Ker Lann, Rue Blaise Pascal, BP 37203, 35172 Bruz Cedex, France. Email: valentin.patilea@ensai.fr. Part of the work for this paper was accomplished while this author was at LEO, Université d'Orléans.

Abstract

In this paper we analyze a large class of semiparametric $M-$estimators for single-index models, including semiparametric quasi-likelihood and semiparametric maximum likelihood estimators. Some possible applications to robustness are also mentioned. The definition of these estimators involves a kernel regression estimator for which a bandwidth rule is necessary. Given the semiparametric $M-$estimation problem, we propose a natural bandwidth choice by joint maximization of the $M-$estimation criterion with respect to the parameter of interest and the bandwidth. In this way we extend a methodology first introduced by Härdle, Hall and Ichimura (1993) for semiparametric least-squares. We prove asymptotic normality for our semiparametric estimator. We derive the asymptotic equivalence between our bandwidth and the optimal bandwidth obtained through weighted cross-validation. Empirical evidence obtained from simulations suggests that our bandwidth improves the higher order asymptotics of the semiparametric $M-$estimator when it replaces the usual bandwidth chosen by cross-validation.

Key words: semiparametric $M-$estimator, single-index model, bandwidth selection, cross-validation, $U-$processes, semiparametric quasi-likelihood, robustness.

MSC 2000: 62G05, 62G08, 62G20, 62J12.

Résumé

Dans ce papier nous étudions une classe de $M-$estimateurs semi-paramétriques dans les modèles à direction révélatrice unique. Cette classe contient en particulier les estimateurs semi-paramétriques de quasi-vraisemblance et du maximum de vraisemblance. La technique utilisée fait appel à un estimateur préliminaire à noyau, pour lequel un choix de fenêtre est nécessaire. Nous proposons un choix "naturel" résultant de la maximisation du critère de $M-$estimation conjointement par rapport au paramètre d'intérêt et à cette fenêtre. Nous étendons ainsi une méthodologie considérée par Härdle, Hall et Ichimura (1993) dans le cadre des moindrés carrés semi-paramétriques. Nous montrons la normalité asymptotique de notre estimateur semi-paramétrique. Nous montrons que la largeur de la fenêtre proposée est asymptotiquement équivalente à celle obtenue par validation croisée pondérée. Des résultats empiriques obtenus par simulations suggèrent que notre fenêtre améliore le comportement asymptotique d'ordre supérieur de notre estimateur semi-paramétrique quand elle remplace la fenêtre usuelle choisie par validation croisée.

Mots clefs: $M-$estimateur semi-paramétrique, modèle à direction révélatrice unique, choix de la fenêtre, $U-$processus, quasi-vraisemblance semi-paramétrique, robustesse.

1 Introduction

Consider the problem of estimating a regression function $m(x) = E(Y|X = x)$ from independent copies $\left(Y_1, X_1^T\right)^T, \ldots, \left(Y_n, X_n^T\right)^T$ of a random vector $\left(Y, X^T\right)^T \in \mathbb{R}^{d+1}$. In GLM (generalized linear models; e.g., McCullagh and Nelder (1989)) it is assumed that $m(x) = r_0(x\theta_0)$ with r_0 known. Here, $x\theta$ is a notation for $x^T\theta$ when x, $\theta \in \mathbb{R}^d$. The function r_0 is the inverse of the so-called link function. Moreover, the conditional density $f_{Y|X=x}$ of Y given $X = x$ belongs to the linear exponential family, that is

$$f_{Y|X=x}(y) = \exp\left[B(r_0(x\theta_0)) + C(r_0(x\theta_0))\,y + D(y)\right],$$

where B, C and D are known functions.

A natural extension of GLM is provided by the semiparametric single-index models (SIM), where one only assumes the existence of some $\theta_0 \in \mathbb{R}^d$ (unique up to a scale normalization factor) such that

$$E(Y \mid X) = E(Y \mid X\theta_0), \tag{1.1}$$

that is $m(x) = r_0(x\theta_0)$, with unknown r_0. Since the regression $r_0(t) = E(Y \mid X\theta_0 = t)$ depends on θ_0, hereafter, we shall write r_{θ_0} instead of r_0. In SIM framework, both θ_0 and r_{θ_0} are to be estimated. Numerous semiparametric approaches for root-n consistent estimation of θ_0 have been proposed : $M-$estimation [e.g., Ichimura (1993), Sherman (1994b), Delecroix and Hristache (1999), Xia and Li (1999), Xia, Tong and Li (1999)], direct (average derivative based) estimation [e.g., Powel, Stock and Stoker (1989), Härdle and Stoker (1989), Hristache, Juditsky and Spokoiny (2001), Hristache, Juditsky, Polzehl and Spokoiny (2001)], iterative methods [e.g., Weisberg and Welsh (1994), Chiou and Müller (1998), Bonneu and Gba (1998), Xia and Härdle (2002)].

Typically, the semiparametric M-estimators mentioned above can be written as

$$\widehat{\theta} = \arg\max_{\theta} \frac{1}{n}\sum_{i=1}^{n} \psi\left(Y_i, \widehat{r}_{\theta,h}^{\,i}(X_i\theta)\right)\tau_n(X_i), \tag{1.2}$$

where $\widehat{r}_{\theta,h}^{\,i}(t)$ is, for instance, the leave-one-out Nadaraya-Watson estimator (with bandwidth h) of $r_\theta(t) = E(Y \mid X\theta = t)$, $-\psi$ is a contrast function and $\tau_n(\cdot)$ is a so-called trimming function introduced to guard against small values for the denominators appearing in $\widehat{r}_{\theta,h}^{\,i}(t)$. Finally, the regression function $m(x)$ is estimated by $\widehat{r}_{\widehat{\theta},h}\left(x\widehat{\theta}\right)$. Other smoothers, such as local polynomials and splines, can replace the Nadaraya-Watson estimator.

In order to estimate θ_0 and $r_{\theta_0}(\cdot\,\theta_0)$, two smoothing parameters seem to be necessary. First, after choosing a primary bandwidth h, the estimator $\widehat{\theta}$ is computed as in (1.2). Afterwards, $r_{\theta_0}(x\theta_0)$ is estimated by $\widehat{r}_{\widehat{\theta},h^*}\left(x\widehat{\theta}\right)$, a kernel estimator, with bandwidth h^*, of the expectation of Y given $x\widehat{\theta}$. The rates of decay for the two bandwidths should verify some conditions. When $\psi(y,r) = -(y-r)^2$, Härdle, Hall and Ichimura (1993) defined more directly

$$\left(\widehat{\theta}, \widehat{h}\right) = \arg\max_{\theta,h} \frac{1}{n}\sum_{i=1}^{n} \psi\left(Y_i, \widehat{r}_{\theta,h}^{\,i}(X_i\theta)\right) I_A(X_i)\ . \tag{1.3}$$

Here, the trimming function is $I_A(\cdot)$, the indicator function of the set A, and A is fixed, bounded and strictly included in the support of X. The regression $r_{\theta_0}(\cdot\,\theta_0)$ can be then estimated by $\widehat{r}_{\widehat{\theta},\widehat{h}}\left(\cdot\,\widehat{\theta}\right)$.

In this paper we consider a class of semiparametric $M-$estimators defined by a general function ψ. Moreover, we provide an automatic and natural choice of the smoothing parameter h used to define the estimator $\widehat{\theta}$. This bandwidth has also some optimal properties for the nonparametric regression. In particular, it is of order $n^{-1/5}$. To achieve these goals we extend Härdle, Hall and Ichimura's idea, that is, given a function ψ, we maximize the semiparametric $M-$estimation criterion (1.2) simultaneously in θ and h. For simplicity we use a leave-one-out Nadaraya-Watson estimation of the regression function, although this approach could be applied for other smoothers like, for instance, local polynomials. Our proofs allow for discrete covariates and do not require a preliminary (or pilot) estimator of θ_0 having a suitable rate of convergence in probability $O_P(n^{-\delta})$, $\delta > 0$.

The methodology we propose allows to build efficient estimators of θ_0 under suitable additional model assumptions. Moreover, it can be extended and applied to a multi-index framework, that is when there exists $\theta_0^1, ..., \theta_0^p \in \mathbb{R}^d$, $p < d$, such that

$$E(Y \mid X) = E\left(Y \mid X\theta_0^1, ..., X\theta_0^p\right)$$

[see Ichimura and Lee (1991) and Picone and Butler (2000)]. Finally, if the probabilistic results on $U-$processes we use in the proofs could be extended to non-i.i.d. data, our theoretical results could be adapted easily to such a case.

The paper is organized as follows. Existing results on semiparametric $M-$estimation are reviewed in section 2. Moreover, the gaps our paper aims to fulfill are clearly described. The methodology we use for the theoretical results is depicted in section 3. As in Härdle, Hall and Ichimura (1993), the basic idea is to show that joint maximization in θ and h is asymptotically equivalent to separate maximization of a purely parametric term with respect to θ and of a purely nonparametric term with respect to h. In this way we derive the asymptotic normality of $\widehat{\theta}$, while for $\widehat{h}$ we obtain an asymptotic equivalence with a theoretical "optimal" bandwidth maximizing the quantity

$$\frac{1}{n}\sum_{i=1}^{n} \psi\left(Y_i, \hat{r}^i_{\theta_0,h}(X_i\theta_0)\right) I_{\left\{x:\, f_{\theta_0}(x\theta_0)\geq c\right\}}(X_i),$$

where f_{θ_0} is the density of $X\theta_0$ and c is some positive constant. We call this quantity a $\psi-CV$ (cross-validation) function. When $\psi(y,r) = -(y-r)^2$, the usual cross-validation function from nonparametric smoothing is recovered up to a change of sign (Clark (1975)). In general, we show that maximizing the $\psi - CV$ function is asymptotically equivalent to minimizing a weighted (mean-squared) cross-validation function. Chiou and Müller (1998, 1999) provide empirical evidence supporting the idea of choosing the bandwidth using other criteria than the usual cross-validation function. Their nonparametric quasi-likelihood criterion is closely related to a $\psi - CV$. Our theoretical results are stated in section 4. Section 5 contains some empirical evidence. It is shown that other functions ψ than the usual $\psi(y,r) = -(y-r)^2$ may provide $M-$estimators $\widehat{\theta}$ with better performances. The choice of ψ acts on the performances of $\widehat{\theta}$ in two ways, through the asymptotic variance and through the optimal choice of h based on the $\psi - CV$ function. The two effects are discussed. Some comments and conclusions end the paper. The assumptions and the technical proofs are provided in the appendices.

Let us end this introduction noticing that it is not clear, a priori, whether an optimal bandwidth for the regression function is also optimal for the estimation of the parameter θ. As pointed out by a referee, to find the optimal bandwidth for θ is of theoretical interest but quite difficult since it involves higher order asymptotic expansions of the semiparametric estimator. This refinement lies beyond the scope of our paper.

2 Motivations

2.1 Possible choices of ψ

Flexibility in the choice of the function $\psi(y, r)$ could be helpful, for instance, when the interest is focused on efficiency, goodness-of-fit or robustness. Sherman (1994b) and Delecroix and Hristache (1999) seem to be the only papers on semiparametric $M-$estimation allowing ψ to belong to a large class of functions.

Apart some technical aspects, our theoretical findings are based on two conditions ensuring that joint maximization in θ and h as in (1.3) is asymptotically equivalent to splitting the criterion into two parts, one purely parametric and another one purely nonparametric, and maximizing separately with respect to θ and h, respectively. These conditions are

$$E\left[\partial_2\psi\left(Y,\ r_{\theta_0}\left(X\theta_0\right)\right) \mid X\right] = 0 \tag{2.1}$$

and

$$E\left[\partial_\theta\partial_2\psi\left(Y,\ r_{\theta_0}\left(X\theta_0\right)\right) \mid X\theta_0\right] = 0, \tag{2.2}$$

where ∂_2 denotes the derivative with respect to the second argument of ψ and ∂_θ is the derivative with respect to all occurrences of θ, that is, given y and x,

$$\partial_\theta\partial_2\psi\left(y,\ r_{\theta_0}\left(x\theta_0\right)\right) = \frac{\partial}{\partial\theta}\,\partial_2\psi\left(y,\ r_\theta\left(x\theta\right)\right)\Big|_{\theta=\theta_0}$$

(see also Sherman (1994b) for similar conditions). In the SIM framework, the two orthogonality assumptions can be satisfied by at least two important types of hypothesis: i) assumptions on ψ without any reference to the conditional distribution of Y given X; and ii) assumptions on ψ combined with some conditions on the conditional law. This brings us to at least three cases where our approach applies, provided that the single-index assumption holds.

Example 1 (quasi-likelihood). Consider a SIM without additional distributional assumptions on the conditional law of Y given X. In this case, the first condition is equivalent to ψ given by linear exponential families (cf. Delecroix and Hristache (1999); see also Gouriéroux, Monfort and Trognon (1984)). We have

$$\begin{aligned}\theta_0 &= \arg\max_\theta E[B\left(r_\theta\left(X\theta\right)\right) + C\left(r_\theta\left(X\theta\right)\right)Y + D\left(Y\right)] \qquad (2.3)\\ &= \arg\max_\theta E[B\left(r_\theta\left(X\theta\right)\right) + C\left(r_\theta\left(X\theta\right)\right)Y],\end{aligned}$$

where B, C satisfy the identity $B'(r) + C'(r)r \equiv 0$ and $C' > 0$. In other words, θ_0 maximizes the function $\theta \mapsto E\left[Q\left(Y, r_\theta\left(X\theta\right)\right)\right]$, where Q is the quasi-(log-)likelihood function

$$Q\left(y, r\right) = \int_y^r C'\left(s\right)\ \left(y - s\right)\ ds.$$

Thus $\partial_2 \psi\left(Y, r\right) = \partial_2 Q\left(y, r\right) = C'(r)\ (y - r)$ from which (2.1) follows. The second condition (2.2) is a consequence of the equation

$$E\left[\partial_{22}^2 \psi\left(Y,\ r_{\theta_0}\left(X\theta_0\right)\right) \mid X\right] = E\left[\partial_{22}^2 \psi\left(Y,\ r_{\theta_0}\left(X\theta_0\right)\right) \mid X\theta_0\right]$$

and of the identity

$$E\left[\partial_\theta r_{\theta_0}\left(X\theta_0\right) \mid X\theta_0\right] = E\left[r'_{\theta_0}\left(X\theta_0\right)\left(X - E\left[X \mid X\theta_0\right]\right) \mid X\theta_0\right],$$

where $r'_{\theta_0}(\cdot)$ is the derivative of $r_{\theta_0}(\cdot)$. This last identity is always true under the SIM assumption (see Newey (1994), p. 1358). The semiparametric least-squares corresponds to $B(r) = -r^2$ and $C(r) = 2r$. Taking $B(r) = -r$ and $C(r) = \ln r$ yields the Poisson pseudo-maximum likelihood method.

More generally, one can allow the functions B, C and D to depend on a nuisance parameter η that is supposed fixed when writing (2.3) (e.g., negative binomial and gamma pseudo-log-likelihood functions). This additional parameter could be used to specify second order moments of Y given X (see Gouriéroux, Monfort and Trognon (1984), section 5).

Example 2 (maximum likelihood). Assume that θ_0 is unique such that the true conditional density of Y given X depends on X only through $r_{\theta_0}\left(X\theta_0\right)$, that is $f_{Y|X=x}(y) = f_0(y; r_{\theta_0}\left(x\theta_0\right))$ with f_0 known. Moreover, the marginal law of X does not depend on θ. If $\psi = \log f_0$, conditions (2.1) and (2.2) are then direct consequences of the model assumptions. This choice of ψ yields the semiparametric maximum likelihood estimator. Examples are : GLM with unknown link function [e.g., Huh and Park (2002), Carroll, Fan, Gijbels and Wand (1997), Chen (1995), Klein and Spady (1993)], quadratic exponential families like the normal $N(r, r^2)$ [e.g., Xia, Tong and Li (2002) with Gaussian residuals]. The corresponding estimator is efficient in the semiparametric sense.

Example 3 (robustness). In robust statistics one usually considers $\psi(y, r) = -\rho(y - r)$ where a) ρ is symmetric; b) the conditional law of $Y - r_{\theta_0}\left(X\theta_0\right)$ given X is symmetric; and c) the conditional law of Y given X depends only on $X\theta_0$ (this is the case, for instance, if the errors $Y - r_{\theta_0}\left(X\theta_0\right)$ are independent of the regressors X). An example of function ρ is the Tukey biweight function

$$\rho_c(t) = \min\left(\frac{t^2}{2} - \frac{t^4}{2c^2} + \frac{t^6}{6c^4}, \frac{c^2}{6}\right),$$

a smooth Huber-type function (see Fraiman, Yohai and Zamar (2001) for a larger class of such smooth functions). For other important examples of $M-$estimators in robust statistics we refer to Hampel, Ronchetti, Rousseeuw and Stahel (1986). Another example is provided by the so-called $\alpha-$estimator defined by $\rho\left(t\right) = 1 - e^{-\alpha t^2}$ (cf. Vajda (1989)).

2.2 Bandwidth range and pilot estimation

A major critique for most of existing asymptotic results on semiparametric $M-$estimation in index-models is related to the domain from which the bandwidths of the regression estimator has to be chosen. Moreover, quite often there is no explicit rule on how to choose the bandwidth in practice. Another important critique is the need of a preliminary estimator approaching θ_0 at a suitable rate $O_P\left(n^{-\delta}\right)$, $\delta > 0$. Table 1 contains a (non-exhaustive) list of papers on semiparametric $M-$estimation together with their assumptions on pilot estimation and bandwidth range. For instance, Härdle, Hall and Ichimura (1993) constrained θ in a $O_P(n^{-1/2})$ neighborhood of θ_0 and h of order $n^{-1/5}$. However, it is not obvious that there exists a $\sqrt{n}-$consistent semiparametric estimator when h is of order $n^{-1/5}$. Another example which may raise questions is the iterative strategy of Xia, Tong and Li (1999), page 837. Their idea is to replace the joint maximization in θ and h by a scheme where θ and h are updated iteratively by separate maximization with respect to one of them when the other is fixed. The parameter θ is taken in a cone Θ_n that shrinks to θ_0 sufficiently fast and h is restricted to a domain of bandwidths of order $n^{-1/5}$. Such a scheme makes sense only if one proves that for a bandwidth of order $n^{-1/5}$ there exists a semiparametric $M-$estimator with a rate compatible with Θ_n.

Table 1. Overview of some assumptions in semiparametric $M-$ estimation for Single-Index Models. Let $\Theta_{n,\delta}$ denote a $O_P(n^{-\delta})$ neighborhood of θ_0 and let $H_{n,\alpha}$ stand for a range of bandwidths of order $O(n^{-\alpha})$. The regression $r_{\theta_0}(t)$ is assumed k times differentiable.

	$\Theta_{n,\delta}$	$H_{n,\alpha}$	k	continuous X
Ichimura (1993)	$O(1)$	$\alpha \in \left(\frac{1}{8}, \frac{1}{7}\right)$	3	no
Härdle, Hall and Ichimura (1993)	$O_P\left(n^{-1/2}\right)$	$\alpha = \frac{1}{5}$	2	yes
Klein and Spady (1993)	$O_P\left(n^{-1/3}\right)$	$\alpha \in \left(\frac{1}{8}, \frac{1}{6}\right)$	1	no
Sherman (1994b)	$o_P\left(1\right)$	$\alpha \in \left(\frac{1}{7}, \frac{1}{6}\right)$	7	no
Carroll, Fan, Gijbels and Wand (1997)	$O_P\left(n^{-1/2}\right)$	$\alpha \in \left(\frac{1}{4}, \frac{1}{2}\right)$	2	yes
Bonneu and Gba (1998)	$O_P\left(n^{-1/2}\right)$	$\alpha \in \left(\frac{1}{8}, \frac{1}{7}\right)$	2	no
Delecroix and Hristache (1999)	$O(1)$	$\alpha \in \left(\frac{1}{8}, \frac{1}{7}\right)$	3	no
Xia and Li (1999)	$\delta \in \left(\frac{3}{10}, \frac{1}{2}\right)$	$\alpha = \frac{1}{5}$	2	yes
Xia, Tong and Li (1999)	$\delta \in \left(\frac{3}{10}, \frac{1}{2}\right)$	$\alpha = \frac{1}{5}$	2	yes
Xia, Tong and Li (2002)	$O(1)$	$\alpha \in \left(\frac{1}{6}, \frac{1}{4}\right)$	4	yes
Xia and Härdle (2002)	$O(1)$	$\alpha \in \left(\frac{1}{6}, \frac{1}{4}\right)$	3	yes
Weisberg and Welsh (1994)	$O(1)$	$\alpha \in \left(\frac{1}{6}, \frac{1}{4}\right)$	4	yes

In section 3 we provide new and useful insights on preliminary estimates and bandwidth range. Our bandwidth should decrease slower than $n^{-1/4}$ and faster than $n^{-1/8}$. If the explanatory variables are bounded, we only need a consistent in probability pilot estimator. For unbounded X but satisfying a suitable moment condition, we only require a consistent preliminary estimator with a rate of convergence in probability faster than $1/\ln n$. Moreover, we indicate how to build an estimator with this rate.

When a function $\psi(y,r)$ other then $-(y-r)^2$ is preferred for defining the $M-$estimator $\widehat{\theta}$, it seems natural to choose h in a way compatible with ψ. For instance, one may prefer a Poisson pseudo-maximum likelihood method ($\psi(y,r) = -r + y \ln r$) for count data regression. In this case, a $\psi - CV$ type criterion should be used for choosing h. Such a choice is equivalent to using a quasi-deviance criterion (McCullagh and Nelder (1989); see also the nonparametric quasi-likelihood deviance used by Chiou and Müller (1998)). The use of a general $\psi - CV$ for choosing h is also supported by the empirical findings reported in Chiou and Müller (1998). Some additional empirical evidence is provided in section 5.

Finally, when selecting the smoothing parameter h in a semiparametric $M-$estimation procedure, one may look for a bandwidth with the optimal rate of decay $n^{-1/(2k+1)}$, where

k is the number of derivatives required for the regression (Stone (1982)). From this point of view, the popular rate $n^{-1/5}$ is justified only if $k = 2$. Our bandwidth is shown to be of order $n^{-1/5}$. For this, only second order derivatives for the regression function are needed.

3 Methodology

To ensure the estimability of the parameter θ, let us fix its first component to 1 and identify θ with its last $d-1$ components. More precisely, from now on θ will be a vector of $\mathbb{R}^{d-1}$ and $x\theta$, with $x \in \mathbb{R}^d$, denotes the matrix product $(1, \theta^T)x$. Accordingly, the parameter set Θ is a subset of $\mathbb{R}^{d-1}$. Finally, without loss of generality, assume that $\psi(\cdot,\cdot) \le 0$.

Given $1/8 < \beta_1 < \beta_2 < 1/4$ and the constants $c_1, c_2 > 0$, define

$$\mathcal{H}_n = \left\{h : \; c_1 n^{-\beta_2} \le h \le c_2 n^{-\beta_1}\right\} \tag{3.1}$$

and take $h_n \in \mathcal{H}_n$, $n \ge 1$. Let θ_n, $n \ge 1$ be a preliminary consistent estimator of θ_0. Define the semiparametric $M-$estimator

$$\left(\widehat{\theta}, \widehat{h}\right) = \underset{\theta\in\Theta,\, h\in\mathcal{H}_n}{\arg\max} \; \frac{1}{n}\sum_{i=1}^{n} \psi\left(Y_i, \hat{r}^i_{\theta,h}\left(X_i\theta\right)\right) I_{\left\{x:\, \widehat{f}^i_{\theta_n,h_n}(x\theta_n)\ge c\right\}}(X_i), \tag{3.2}$$

where $c > 0$ and

$$\hat{r}^i_{\theta,h}(t) = \frac{\frac{1}{n-1}\sum\limits_{j\neq i} Y_j \; K_h\left(t - X_j\theta\right)}{\frac{1}{n-1}\sum\limits_{j\neq i} K_h\left(t - X_j\theta\right)} =: \frac{\widehat{\gamma}^i_{\theta,h}(t)}{\widehat{f}^i_{\theta,h}(t)}$$

denotes the leave-one-out version of the Nadaraya-Watson estimator of the regression function

$$r_\theta(t) = E\left(Y|X\theta = t\right) =: \frac{\gamma_\theta(t)}{f_\theta(t)},$$

with f_θ the density of $X\theta$. The function $K(\cdot)$ is a (second order) kernel function and $K_h(\cdot)$ stands for $K(\cdot/h)/h$.

Trimming is designed to keep $\widehat{f}^i_{\theta,h}$ away from zero and thus to stabilize computations. On the other hand, trimming is usually required for analyzing the asymptotic properties of the nonparametric regression estimator and of the optimal bandwidth. The practical purpose of trimming recommends a data-driven device like $I_{\left\{x:\, \widehat{f}^i_{\theta,h}(x\theta)\ge c\right\}}(\cdot)$. However, to ensure consistency with such trimming, one should require that θ_0 is the maximizer of the map

$$\theta \mapsto E\left[\psi\left(Y, r_\theta\left(X\theta\right)\right) I_{\{x:\, f_\theta(x\theta)\ge c\}}(X)\right], \qquad \theta\in\Theta.$$

Meanwhile, a trimming function like $I_{\left\{x:\, f_{\theta_0}(x\theta_0)\ge c\right\}}(\cdot)$ is easier to manipulate in theory. Our trimming procedure aims to reduce this gap. It is simple and easy to implement in applications since it only consists of a checking of the observations before starting the optimization. In practice, quite often one may take $I_{\left\{z:\, \widehat{f}^i_{\theta_n,h_n}(x'\theta_n)\ge c\right\}}(\cdot) \equiv 1$. On the other hand, in a certain sense, our trimming is asymptotically equivalent to the fixed trimming $I_{\left\{x:\, f_{\theta_0}(x\theta_0)\ge c\right\}}(\cdot)$ which renders the proofs quite transparent. We prove this equivalence under two types of assumptions: either i) X is bounded and $\theta_n - \theta_0 = o_P(1)$, or ii)

$E\left[\exp\left(\lambda\left|X\right|\right)\right] < \infty$, for some $\lambda > 0$, and $\theta_n - \theta_0 = o_P\left(1/\ln n\right)$. Preliminary consistent estimates can be obtained by $M-$estimation with a fixed trimming $I_B(\cdot)$ where B is a subset of $\mathbb{R}^d$ such that $f_\theta\left(x\theta\right) \geq c > 0$, $x \in B$, $\theta \in \Theta$. In particular, it can be shown that this preliminary estimator is $o_P\left(1/\ln n\right)$ for a range of bandwidths $\left[n^{-(1/2-\varepsilon)}, n^{-\varepsilon}\right]$, $0 < \varepsilon < 1/2$. For a complete proof, see Appendix E.

Let us point out that a minor modification of the arguments used in Appendix E yields the consistency in probability for the estimator $\widehat{\theta}$ defined in (3.2) when X is bounded. Therefore, for the asymptotic results, in the maximization problem (3.2) we can replace the parameter set Θ by a sequence of neighborhoods Θ_n, $n \geq 1$ shrinking to θ_0. For technical reasons, when X is unbounded, we have to define the estimator (3.2) with Θ replaced by shrinking neighborhoods Θ_n, $n \geq 1$. In practice, there is no difference between the cases X bounded and X unbounded and therefore $(\widehat{\theta}, \widehat{h})$ can be always computed by maximization over Θ.

Define $A = \{x : f_{\theta_0}\left(x\theta_0\right) \geq c\} \subset \mathbb{R}^d$ and $A^\delta = \{x : \left|f_{\theta_0}\left(x\theta_0\right) - c\right| \leq \delta\}$, $\delta > 0$. By little algebra, for all $\theta \in \Theta_n$, $h \in \mathcal{H}_n$ and i,

$$\left|I_{\left\{x:\, \widehat{f}^i_{\theta,h}(x\theta)\geq c\right\}}(X_i) - I_A(X_i)\right| \leq I_{A^\delta}(X_i) + I_{(\delta,\infty)}(Z_n),$$

where

$$Z_n = \max_{1\leq i\leq n} \sup_{\theta\in\Theta_n,\, h\in\mathcal{H}_n} \left|\widehat{f}^i_{\theta,h}\left(X_i\theta\right) - f_{\theta_0}\left(X_i\theta_0\right)\right|.$$

Let

$$\widehat{S}\left(\theta, h; \widetilde{A}\right) = \frac{1}{n}\sum_{i=1}^n \psi\left(Y_i, \hat{r}^i_{\theta,h}\left(X_i\theta\right)\right)\ I_{\widetilde{A}}\left(X_i\right)$$

with $\widetilde{A} = A$ or A^δ. Since $\psi\left(\cdot,\cdot\right) \leq 0$, we have

$$\left|\frac{1}{n}\sum_{i=1}^n \psi\left(Y_i, \hat{r}^i_{\theta,h}\left(X_i\theta\right)\right)\ I_{\left\{x:\, \widehat{f}^i_{\theta_n,h_n}(x\theta_n)\geq c\right\}}(X_i) - \widehat{S}\left(\theta, h; A\right)\right|$$
$$\leq -\widehat{S}\left(\theta, h; A^\delta\right) - \frac{I_{(\delta,\infty)}(Z_n)}{n}\sum_{i=1}^n \psi\left(Y_i, \hat{r}^i_{\theta,h}\left(X_i\theta\right)\right).$$

We show that $\widehat{S}\left(\theta, h; A^\delta\right) = o_P(\widehat{S}\left(\theta, h; A\right))$, uniformly over $\Theta_n \times \mathcal{H}_n$, provided that $\delta \to 0$ and $P\left(f_{\theta_0}\left(X\theta_0\right) = c\right) = 0$. On the other hand, we prove that $P\left(Z_n > \delta\right) \to 0$, provided that $\delta \to 0$ slowly enough (see Lemma B.2 in the appendix). All this proves that, modulo arbitrarily small corrections, $\left(\widehat{\theta}, \widehat{h}\right)$ can be defined as the maximizer of $\widehat{S}\left(\theta, h; A\right)$ over $\Theta_n \times \mathcal{H}_n$. Hereafter, we simply write $\widehat{S}\left(\theta, h\right)$ instead of $\widehat{S}\left(\theta, h; A\right)$ and we consider

$$\left(\widehat{\theta}, \widehat{h}\right) = \underset{\theta\in\Theta_n,\, h\in\mathcal{H}_n}{\arg\max}\ \widehat{S}\left(\theta, h\right),$$

with Θ_n, $n \geq 1$ shrinking to θ_0 and $\mathcal{H}_n$ defined in (3.1).

Next, the basic idea is that the semiparametric criterion $\widehat{S}\left(\theta, h\right)$ can be split into a purely parametric part $\widetilde{S}\left(\theta\right)$, a purely nonparametric one $T(h)$ and a reminder term

$R(\theta, h)$, where

$$\begin{aligned}
\widetilde{S}(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\psi\left(Y_i, r_\theta\left(X_i\theta\right)\right)\ I_A\left(X_i\right) - \frac{1}{n}\sum_{i=1}^{n}\psi\left(Y_i, r_{\theta_0}\left(X_i\theta_0\right)\right)\ I_A\left(X_i\right),\\
T(h) &= \frac{1}{n}\sum_{i=1}^{n}\psi\left(Y_i, \hat{r}^i_{\theta_0,h}\left(X_i\theta_0\right)\right)\ I_A\left(X_i\right),\\
R(\theta,h) &= \frac{1}{n}\sum_{i=1}^{n}\left[\psi\left(Y_i, \hat{r}^i_{\theta,h}\left(X_i\theta\right)\right) - \psi\left(Y_i, r_\theta\left(X_i\theta\right)\right)\right]\ I_A\left(X_i\right)\\
&\quad -\frac{1}{n}\sum_{i=1}^{n}\left[\psi\left(Y_i, \hat{r}^i_{\theta_0,h}\left(X_i\theta_0\right)\right) - \psi\left(Y_i, r_{\theta_0}\left(X_i\theta_0\right)\right)\right]\ I_A\left(X_i\right)
\end{aligned}$$

(see also Härdle, Hall and Ichimura (1993) for a slightly different splitting). In view of this general purpose decomposition, the simultaneous optimization of $\widehat{S}(\theta, h)$ is equivalent to separately optimizing $\widetilde{S}(\theta)$ with respect to θ and $T(h)$ with respect to h, provided that $R(\theta, h)$ is sufficiently small. To prove that $R(\theta, h)$ is indeed negligible we show in Proposition D.1 in Appendix D that

$$\begin{aligned}
R(\theta,h) &= \left\{O_P\left(h^4\right) + O_P\left(\frac{1}{nh^2}\right) + O_P\left(\frac{1}{n\sqrt{n}h^4}\right) + o_P\left(\frac{1}{\sqrt{n}}\right)\right. \qquad (3.3)\\
&\quad \left. + O_P\left(|\theta-\theta_0|\right)\left[O\left(h^2\right) + O_P\left(\frac{1}{h^2\sqrt{n}}\right)\right]\right\} \times O_P\left(|\theta-\theta_0|\right),
\end{aligned}$$

as $n \to \infty$, uniformly in $h \in \left[n^{-(1/2-\varepsilon)}, n^{-\varepsilon}\right]$, with $0 < \varepsilon < 1/2$, and uniformly in $\theta \in \Theta_n$; herein, $|\cdot|$ denotes the Euclidean metric. The key ingredients for proving this identity are the cornerstone conditions (2.1) and (2.2), the definition of the trimming set A and results on the rates of convergence for *degenerate* $U-$processes (Sherman (1994a)).

Taylor expansion yields

$$\widetilde{S}(\theta) = O_P\left(\frac{|\theta-\theta_0|}{\sqrt{n}}\right) + O_P\left(|\theta-\theta_0|^2\right)$$

and therefore, to ensure that $R(\theta, h)$ is negligible with respect to $\widetilde{S}(\theta)$, it suffices to constrain h in the range $\mathcal{H}_n$ defined in (3.1). Even if other decompositions of $\widehat{S}(\theta, h)$ may be used, a careful inspection of our proofs suggests that a bandwidth range like $\mathcal{H}_n$ is the largest for which we may deduce $\sqrt{n}-$consistency for $\widehat{\theta}$ in our framework.

3.1 Asymptotic distribution for $\widehat{\theta}$

In view of (3.3) deduce that

$$R(\theta,h) = o_P\left(\frac{|\theta-\theta_0|}{\sqrt{n}}\right) + o_P\left(|\theta-\theta_0|^2\right),$$

uniformly in $h \in \mathcal{H}_n$ and $\theta \in \Theta_n$. Use this and Taylor expansion to write

$$\begin{aligned}
\widehat{S}(\theta,h) &= \frac{1}{\sqrt{n}}\left(\theta-\theta_0\right)^T V_n - \frac{1}{2}\left(\theta-\theta_0\right)^T\ W_n\ \left(\theta-\theta_0\right) \qquad (3.4)\\
&\quad + o_P\left(\frac{|\theta-\theta_0|}{\sqrt{n}}\right) + o_P\left(|\theta-\theta_0|^2\right) + \{\text{terms not depending on } \theta\},
\end{aligned}$$

uniformly over Θ_n, where

$$V_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \partial_\theta \psi \left(Y_i, r_{\theta_0} \left(X_i \theta_0 \right) \right) I_A \left(X_i \right), \qquad W_n = -\frac{1}{n} \sum_{i=1}^{n} \partial^2_{\theta\theta} \psi \left(Y_i, r_{\theta_0} \left(X_i \theta_0 \right) \right) I_A \left(X_i \right)$$

(here, $\partial_\theta \psi$ is a vector in $\mathbb{R}^{d-1}$, while $\partial^2_{\theta\theta}\psi$ is a $(d-1) \times (d-1)$ matrix). From the assumptions we shall impose below, the vector V_n converges in distribution to $\mathcal{N}(0, M_0)$ and $W_n \to W_0$, almost surely, where

$$M_0 = E \left[\partial_\theta \psi \left(Y_i, r_{\theta_0} \left(X_i \theta_0 \right) \right) \ \partial_\theta \psi \left(Y_i, r_{\theta_0} \left(X_i \theta_0 \right) \right)^T I_A \left(X_i \right) \right],$$

and

$$W_0 = -E \left[\partial^2_{\theta\theta} \psi \left(Y_i, r_{\theta_0} \left(X_i \theta_0 \right) \right) \ I_A \left(X_i \right) \right].$$

Intuitively, $\widehat{\theta}$ has the same asymptotic distribution as the maximizer of the quadratic form (3.4). More precisely, apply Theorems 1 and 2 of Sherman (1994a) to deduce first, the $\sqrt{n}-$consistency of $\widehat{\theta}$ and next, the asymptotic normality

$$\sqrt{n} \left(\widehat{\theta} - \theta_0 \right) \xrightarrow{\mathcal{D}} \mathcal{N} \left(0, W_0^{-1} M_0 W_0^{-1} \right).$$

3.2 Order of $T(h)$ and behavior of $\widehat{h}$

By Taylor expansion we can write

$$T(h) = T_0 + T_1(h) + T_2(h) + \{\text{negligible terms}\},$$

where T_0 is independent of h,

$$T_1(h) = \frac{1}{n} \sum_{i=1}^{n} \ \partial_2 \psi \left(Y_i, r_{\theta_0} \left(X_i \theta_0 \right) \right) \left[\hat{r}^i_{\theta_0,h} \left(X_i \theta_0 \right) - r_{\theta_0} \left(X_i \theta_0 \right) \right] \ I_A \left(X_i \right)$$

and

$$T_2(h) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \ \partial^2_{22} \psi \left(Y_i, r_{\theta_0} \left(X_i \theta_0 \right) \right) \left[\hat{r}^i_{\theta_0,h} \left(X_i \theta_0 \right) - r_{\theta_0} \left(X_i \theta_0 \right) \right]^2 \ I_A \left(X_i \right).$$

Using condition (2.1) and rates of convergence for degenerate $U-$processes, we deduce from Lemma C.1, Appendix C, that, uniformly over $\mathcal{H}_n$,

$$T_2(h) = -C_1 h^4 - C_2/nh + o_P(h^4 + 1/nh),$$

with C_1, C_2 some constants defined in section 4 below; our Lemma C.1 is a refinement of a well-known result from nonparametric regression (e.g., Härdle and Marron (1985)). We also show that $T_1(h) = o_P(T_2(h))$, uniformly over $\mathcal{H}_n$ (see Lemma C.2).

Note that $R(\theta, h) = o_P(T_2(h))$, uniformly in θ in $O_P(n^{-1/2})$ neighborhoods of θ_0 and $h \in \mathcal{H}_n$. Since $\widehat{\theta}$ was shown to be $\sqrt{n}-$consistent, deduce that $\widehat{h}$ is asymptotically equivalent to the maximizer of $T_2(h)$. More precisely, $\widehat{h}/h_n^{opt} \to 1$, in probability, where $h_n^{opt} = (C_2/4C_1)^{1/5} n^{-1/5}$. As a by-product of these results, we obtain the asymptotic equivalence between the $\psi - CV$ function $T(h)$ and $T_2(h)$.

Remark that in the quasi-likelihood framework (see Example 1 in section 2) where $\psi(y,r) = B(r) + C(r)y + D(y)$ and $\partial^2_{22}\psi(y,r) = C''(r)(y-r) - C'(r)$, the function $-T_2(h)$ is asymptotically equivalent to the weighted cross-validation function

$$\mathcal{X}^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\,C'\left(r_{\theta_0}(X_i\theta_0)\right)\left[\hat{r}^{i}_{\theta_0,h}(X_i\theta_0) - r_{\theta_0}(X_i\theta_0)\right]^2\; I_A(X_i)\,.$$

(Recall that $[C'(r)]^{-1}$ is the variance of the conditional law given by $\exp\psi(y,r)$). In other words, choosing h by optimizing a $\psi - CV$ criterion is asymptotically equivalent to choosing it as the minimizer of a kind of Pearson chi-squared statistics $\mathcal{X}^2$. In particular, this explains why in practice Pearson and deviance-based bandwidth choices are about the same (cf. Chiou and Müller (1998), page 1382).

To conclude this section, note that the arguments above reduce maximization of $\widehat{S}(\theta,h)$ over $\Theta_n \times \mathcal{H}_n$ to maximization with respect to θ in a $O_P(n^{-1/2})$ neighborhood of θ_0 and h of order $O_P(n^{-1/5})$. Indeed, up to asymptotically negligible adjustments, we can write

$$\begin{aligned}
\max_{\theta\in\Theta_n,\, h\in\mathcal{H}_n}\widehat{S}(\theta,h) &= \max_{h\in\mathcal{H}_n}\left\{\max_{\theta\in\Theta_n}\left[\widetilde{S}(\theta) + R(\theta,h)\right] + T(h)\right\}\\
&= \max_{h\in\mathcal{H}_n}\left\{\max_{|\theta-\theta_0| = O_P\left(n^{-1/2}\right)}\left[\widetilde{S}(\theta) + R(\theta,h)\right] + T(h)\right\}\\
&= \max_{h\in\mathcal{H}_n}\max_{|\theta-\theta_0| = O_P\left(n^{-1/2}\right)}\widehat{S}(\theta,h) = \max_{|\theta-\theta_0| = O_P\left(n^{-1/2}\right)}\max_{h\in\mathcal{H}_n}\widehat{S}(\theta,h)\\
&= \max_{|\theta-\theta_0| = O_P\left(n^{-1/2}\right)}\left\{\max_{h\in\mathcal{H}_n}\left[T(h) + R(\theta,h)\right] + \widetilde{S}(\theta)\right\}\\
&= \max_{|\theta-\theta_0| = O_P\left(n^{-1/2}\right)}\left\{\max_{h = O_P\left(n^{-1/5}\right)}\left[T(h) + R(\theta,h)\right] + \widetilde{S}(\theta)\right\}\\
&= \max_{|\theta-\theta_0| = O_P\left(n^{-1/2}\right)}\max_{h = O_P\left(n^{-1/5}\right)}\widehat{S}(\theta,h)\,.
\end{aligned}$$

Hence, one of our contributions is to prove, and no longer to assume, as in Härdle, Hall and Ichimura (1993), that the rates of $\widehat{\theta}$ and $\widehat{h}$ are indeed $n^{-1/2}$ and $n^{-1/5}$, respectively.

4 The main results

Assume that the parameter set $\Theta \in \mathbb{R}^{d-1}$ is compact with nonvoid interior. Define

$$\begin{aligned}
C_1 &= \frac{K_1^2}{4}\,E\left\{\frac{1}{2}\,\partial^2_{22}\psi\left(r_{\theta_0}(X\theta_0), r_{\theta_0}(X\theta_0)\right)\right.\\
&\qquad\left.\times\left[r''_{\theta_0}(X\theta_0) + \frac{2\,r'_{\theta_0}(X\theta_0)\;f'_{\theta_0}(X\theta_0)}{f_{\theta_0}(X\theta_0)}\right]^2\; I_A(X)\right\}, \qquad (4.1)\\
C_2 &= K_2\,E\left\{\frac{1}{2}\,\partial^2_{22}\psi\left(r_{\theta_0}(X\theta_0), r_{\theta_0}(X\theta_0)\right)\frac{1}{f_{\theta_0}(X\theta_0)}\,v_{\theta_0}(X\theta_0)\; I_A(X)\right\}
\end{aligned}$$

and

$$h_n^{opt} = \arg\max_h\left(C_1h^4 + C_2n^{-1}h^{-1}\right) = (C_2/4C_1)^{1/5}\,n^{-1/5}.$$

Theorem 4.1 *Suppose that the assumptions of Appendix A hold and X is bounded. If $(\widehat{\theta},\widehat{h})$ is defined as in (3.2), then $\widehat{h}/h_n^{opt} \to 1$, in probability, and*

$$\sqrt{n}\left(\widehat{\theta}-\theta_0\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, W_0^{-1}M_0W_0^{-1}\right).$$

If X is unbounded, consider a sequence of real numbers $\{d_n\}$ such that $d_n \ln n \to 0$ and let $\Theta_n = \{\theta : |\theta-\theta_0| \le d_n\}$, $n \ge 1$. The conclusions remain true if Θ is replaced by Θ_n in the definition of $(\widehat{\theta},\widehat{h})$.

Proof. First, consider that $(\widehat{\theta},\widehat{h})$ is defined as in (3.2) with Θ replaced by Θ_n. We can decompose $\widehat{S}\left(\theta, h; A^\delta\right)$ in the same way as $\widehat{S}\left(\theta,h\right)$ and obtain the same order, uniformly over $\Theta_n \times \mathcal{H}_n$ and uniformly in $\delta \in [0,\delta_0]$, for some small δ_0 (apply also Lemma B.4b)). The constants appearing in the dominating terms of the decomposition vanishes as $\delta \to 0$, provided that $P\left(f_{\theta_0}\left(X\theta_0\right)=c\right) = 0$. Consequently, $\widehat{S}\left(\theta,h;A^\delta\right) = o_P(\widehat{S}\left(\theta,h;A\right))$, uniformly over $\Theta_n \times \mathcal{H}_n$, if $\delta \to 0$. Next, the conclusions follow from Lemma B.2, Corollary 3.1, Proposition D.1 and the arguments in section 3.

For the case where X is bounded, the same arguments as in Appendix E yield the consistency in probability for the estimator $\widehat{\theta}$ when $(\widehat{\theta},\widehat{h})$ is defined as in (3.2), that is maximizing over $\Theta \times \mathcal{H}_n$. This means that $\Theta \times \mathcal{H}_n$ can be replaced by $\Theta_n \times \mathcal{H}_n$ where the diameter d_n of Θ_n tends to zero. Finally, use the Remark following Lemma B.3 to complete the proof. ■

For the nonparametric part we have the following usual result (see Härdle and Stoker (1989)). The proof is omitted.

Theorem 4.2 *Assume that the conditions of Theorem 4.1 are fulfilled. Then, for any t such that $f_{\theta_0}(t) > 0$,*

$$\sqrt{n\widehat{h}}\left(\widehat{r}_{\widehat{\theta},\widehat{h}}(t) - r_{\theta_0}(t) - \widehat{h}^2\beta(t)\right) \xrightarrow{\mathcal{D}} N\left(0,\ K_2 v_{\theta_0}(t) f_{\theta_0}(t)^{-1}\right)$$

where $\beta(t) = (K_1/2)\left[r_{\theta_0}''(t) + 2r_{\theta_0}'(t)f_{\theta_0}'(t)f_{\theta_0}(t)^{-1}\right]$.

Note that, for any x such that $f_{\theta_0}(x\theta_0) > 0$,

$$\sqrt{n\widehat{h}}\left(\widehat{r}_{\widehat{\theta},\widehat{h}}\left(x\widehat{\theta}\right) - r_{\theta_0}(x\theta_0) - \widehat{h}^2\beta(x\theta_0)\right) \xrightarrow{\mathcal{D}} N\left(0,\ K_2 v_{\theta_0}(x\theta_0) f_{\theta_0}(x\theta_0)^{-1}\right). \tag{4.2}$$

Indeed, we can write

$$\begin{aligned}
\widehat{r}_{\widehat{\theta},\widehat{h}}\left(x\widehat{\theta}\right) - r_{\theta_0}(x\theta_0) &= \widehat{r}_{\widehat{\theta},\widehat{h}}\left(x\widehat{\theta}\right) - \widehat{r}_{\theta_0,\widehat{h}}(x\theta_0) + \widehat{r}_{\theta_0,\widehat{h}}(x\theta_0) - r_{\theta_0}(x\theta_0) \\
&= \partial_\theta \widehat{r}_{\theta_0,\widehat{h}}(x\theta_0)\left(\widehat{\theta}-\theta_0\right) + o_P\left(\left|\widehat{\theta}-\theta_0\right|\right) + \widehat{r}_{\theta_0,\widehat{h}}(x\theta_0) - r_{\theta_0}(x\theta_0) \\
&= O_P\left(n^{-1/2}\right) + \widehat{r}_{\theta_0,\widehat{h}}(x\theta_0) - r_{\theta_0}(x\theta_0),
\end{aligned}$$

because $\partial_\theta \widehat{r}_{\theta_0,\widehat{h}}(x\theta_0) \to \partial_\theta r_{\theta_0}(x\theta_0)$, in probability, uniformly over $o_P(1)$ neighborhoods of θ_0 (see Lemma B.3). Thus, the convergence in (4.2) is a consequence of the asymptotic distribution of the Nadaraya-Watson estimator (e.g., Bosq and Lecoutre (1987)).

5 Empirical evidence

In order to illustrate the finite sample properties of our estimator, we conducted a simulation study using a SAS 8.1 program. For optimization we used the NLPNRA routine of SAS/IML software. This routine is based on a Newton-Raphson method. All the estimates reported in this section were obtained with a quartic kernel $K(u) = (15/16)\left(u^2-1\right)^2 I_{[-1,1]}(u)$.

In the first experiment, the data were generated in the following way :

1. $X_i = \left(X_i^{(1)}, X_i^{(2)}, X_i^{(3)}, X_i^{(4)}\right)^T \in \mathbb{R}^4$:

   $$X_i^{(1)} \sim \mathcal{N}(0,1/4)\,,\;\; X_i^{(2)} \sim \mathcal{B}(1,1/2)\,,\;\; X_i^{(3)} \sim \mathcal{N}(0,1/4)\,,\;\; X_i^{(4)} = \left(X_i^{(3)} + Q\right)/2,$$

   with $Q \sim \mathcal{N}(0,1)$, where $X_i^{(1)}$, $X_i^{(2)}$, $X_i^{(3)}$ and Q are independent random variables;

2. $\theta_0 = \left(\theta_0^{(1)}, \theta_0^{(2)}, \theta_0^{(3)}, \theta_0^{(4)}\right) = (1,1,1,1)^T$;

3. the conditional law of Y_i given $X_i = x$ is a negative binomial law of mean $r_0(x\theta_0)$ and variance $r_0(x\theta_0)\left[1 + r_0(x\theta_0)\right]$, where $r_0(t) = \exp\left[(t-3)/2\right]$.

Three types of ψ were used to estimate θ_0:

a) $\psi_{NB}(y,r) = y\,\log r - (y+1)\,\log(1+r)$, corresponding to the true density of Y given X;

b) $\psi_P(y,r) = y\,\log r - r$, corresponding to a Poisson pseudo-likelihood;

c) $\psi_N(y,r) = -r^2 + 2\,r\,y$, yielding the semiparametric least squares estimator considered in Härdle, Hall and Ichimura (1993).

Let $\widehat{\theta}_{NB}$, $\widehat{\theta}_P$ and $\widehat{\theta}_N$ be the $M-$estimators corresponding to ψ_{NB}, ψ_P and ψ_N, respectively. Write $\widehat{\theta} = \left(1, \widehat{\theta}^{(2)}, \widehat{\theta}^{(3)}, \widehat{\theta}^{(4)}\right)^T$ where $\widehat{\theta}$ stands for any of $\widehat{\theta}_{NB}$, $\widehat{\theta}_P$ and $\widehat{\theta}_N$. For each sample size $n \in \{250, 500, 1000\}$ we generated 500 samples $\left(Y_i, X_i^T\right)^T \in \mathbb{R}^5$, $1 \leq i \leq n$. For each sample we computed $\widehat{\theta}_{NB}$, $\widehat{\theta}_P$ and $\widehat{\theta}_N$.

Table 2 contains the mean and the standard deviation of the estimates of the last three components of θ_0. Moreover, we calculate the estimated mean squared error of $\widehat{\theta}$

$$MSE = \frac{1}{500}\sum_{s=1}^{500}\left(\left|\widehat{\theta}_s^{(2)} - \theta_0^{(2)}\right|^2 + \left|\widehat{\theta}_s^{(3)} - \theta_0^{(3)}\right|^2 + \left|\widehat{\theta}_s^{(4)} - \theta_0^{(4)}\right|^2\right),$$

with $\widehat{\theta}$ equal to $\widehat{\theta}_{NB}$, $\widehat{\theta}_P$ and $\widehat{\theta}_N$, respectively.

We remark that $\widehat{\theta}_P$ outperforms $\widehat{\theta}_N$ in terms of bias and variance. Even if the true model we considered is characterized by a significant overdispersion, the semiparametric Poisson pseudo-maximum likelihood estimator behaves almost like the semiparametric maximum likelihood estimator $\widehat{\theta}_{NB}$.

Table 2. The true conditional law of Y given X is negative binomial of mean $r_0(x\theta_0)$ and variance $r_0(x\theta_0)[1+r_0(x\theta_0)]$ with $r_0(t)=\exp[(t-3)/2]$. The true vector θ_0 is $(1,1,1,1)^T$. Let $\widehat{\theta}_{NB}$, $\widehat{\theta}_P$ and $\widehat{\theta}_N$ denote the $M-$ estimators obtained from the true, Poisson and normal log-likelihoods, respectively. The upperscripts indicate the components of the vectors.

n		$\widehat{\theta}_{NB}^{(2)}$	$\widehat{\theta}_{NB}^{(3)}$	$\widehat{\theta}_{NB}^{(4)}$	$\widehat{\theta}_{P}^{(2)}$	$\widehat{\theta}_{P}^{(3)}$	$\widehat{\theta}_{P}^{(4)}$	$\widehat{\theta}_{N}^{(2)}$	$\widehat{\theta}_{N}^{(3)}$	$\widehat{\theta}_{N}^{(4)}$
250	mean	1.144	1.200	1.185	1.156	1.218	1.179	1.297	1.236	1.250
	st. dev.	1.189	1.177	1.203	1.231	1.254	1.277	1.462	1.287	1.338
	MSE		4.331			4.881			5.780	
500	mean	1.052	1.088	1.117	1.096	1.090	1.140	1.115	1.076	1.142
	st. dev.	0.791	0.810	0.796	0.892	0.823	0.840	0.927	0.895	0.907
	MSE		1.937			2.212			2.518	
1000	mean	1.078	1.045	1.077	1.084	1.041	1.072	1.102	1.072	1.124
	st. dev.	0.521	0.524	0.466	0.507	0.527	0.439	0.645	0.651	0.625
	MSE		0.775			0.740			1.260	

The choice of a criterion ψ influences the estimates of θ_0 in two ways. On one hand, the function ψ appears in the asymptotic variance of the $M-$estimator. On the other hand, the semiparametric criterion defined by ψ is also used for choosing the bandwidth. The choice of the bandwidth does not influence the asymptotic variance of the $M-$estimator but it influences its higher order asymptotic properties. In order to distinguish the performance gain due to our new way of choosing the bandwidth, we conducted a second simulation experiment. This time we compute the $M-$estimator and the bandwidth using a sequence of iterations. For a given ψ, two types of iterative procedures are considered. In the first one, which we call it iterative procedure I, we choose h through a $\psi - CV$ function as follows:

I.1 For given $\widehat{\theta}$, find $\widehat{h}$ the maximizer of the $\psi - CV$ function

$$h \to \frac{1}{n}\sum_{i=1}^{n} \psi\left(Y_i, \hat{r}^i_{\widehat{\theta},h}\left(X_i\widehat{\theta}\right)\right) \; ;$$

I.2 For given $\widehat{h}$, an updated estimate $\widehat{\theta}$ is obtained by maximizing

$$\theta \to \frac{1}{n}\sum_{i=1}^{n} \psi\left(Y_i, \hat{r}^i_{\theta,\widehat{h}}\left(X_i\theta\right)\right).$$

For the second type of iterations considered we choose the bandwidth through a classical cross-validation procedure. The following steps defining the iterative procedure II are run until convergence:

II.1 For given $\widehat{\theta}$, find $\widehat{h}$ the minimizer of CV function

$$h \to \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - \hat{r}^i_{\widehat{\theta},h}\left(X_i\widehat{\theta}\right)\right)^2 \; ;$$

II.2 For given $\widehat{h}$, an updated estimate $\widehat{\theta}$ is obtained by maximizing

$$\theta \to \frac{1}{n}\sum_{i=1}^{n}\psi\left(Y_i, \hat{r}^{i}_{\theta,\widehat{h}}\left(X_i\theta\right)\right).$$

Note that repeating Steps I.1 and I.2 until convergence one expects some values $\left(\widehat{\theta},\widehat{h}\right)$ very close to those obtained by joint optimization in θ and h. However, our experience proves that iterating Steps I.1 and I.2 is not only more computational demanding but also leads to more instable results. For the sake of more accurate comparisons, we maintain the iterative procedure I even when we choose the bandwidth through a $\psi-$CV criterion.

For this second simulation experiment we used the same conditional distribution of Y given X and the same one-dimensional regression function $r_0(\cdot)$. For shorter computations, we take only two independent explanatory variables $X^{(1)}$, $X^{(2)} \sim \mathcal{N}(0,1)$ and we fix $\theta_0 = (1,1)^T$. In this case, we only have to calculate the second components $\widehat{\theta}^{(2)}_{NB}$, $\widehat{\theta}^{(2)}_{P}$ and $\widehat{\theta}^{(2)}_{N}$ of the six $M-$estimators we consider (that is, we consider ψ_{NB}, ψ_P and ψ_N in each of the two iterative procedures above). For each $n \in \{100, 200, 400\}$ we draw 500 samples $\left(Y_i, X_i^T\right)^T \in \mathbb{R}^3$, $1 \leq i \leq n$. The results are given in Table 3. Looking at the MSE we notice that the contribution of the bandwidth choice method to the performances of the $M-$estimator is significant. There is a clear improvement of the MSE, mainly because of a smaller variance, when the 'optimal' bandwidth is obtained through a $\psi - CV$ function.

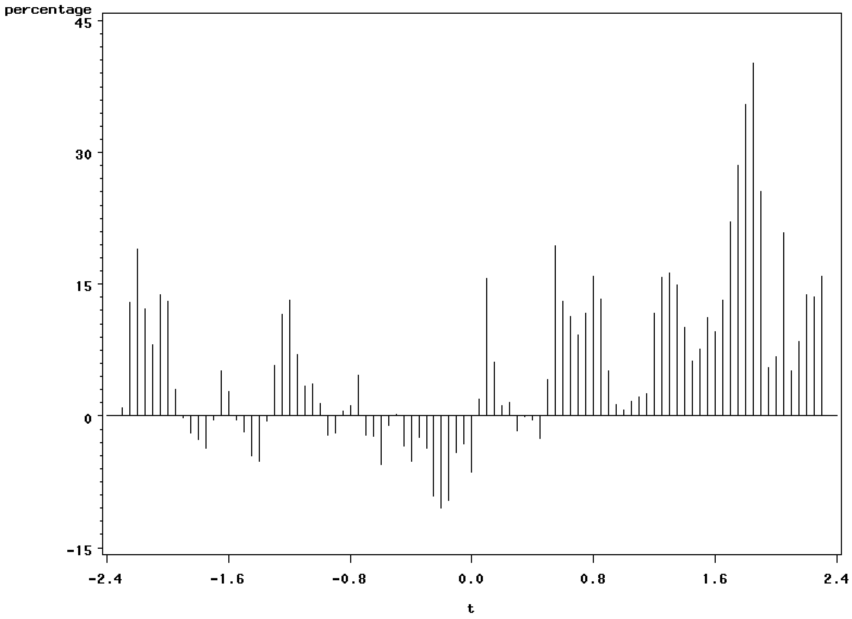
Finally, we analyze the impact of the bandwidth choice on the performances of the nonparametric estimator of the regression function $r_0(\cdot)$ using the results of our second experiment (see also Chiou and Müller (1998), section 5, for a similar analysis). For brevity, only the case $n = 400$ and ψ equal to ψ_P is considered. A grid of points t between -2.3 and 2.3 is fixed. Given the design of our experiment, the probability that $|X\theta_0| \leq 2.3$ is close to 0.9. The Nadaraya-Watson estimators $\hat{r}_{\widehat{\theta},\widehat{h}}(t)$ with t in the grid and $(\widehat{\theta},\widehat{h})$ yielded by each of the iterative procedures I and II were computed.

Table 3. The law of Y given X and the regression $r_0(t)$ are as in Table 2. Moreover, $\theta_0 = (1,1)^T$. The iterative procedures I and II are considered. Each $M-$ estimator is obtained by iterative separate optimization with respect to the second component of θ and with respect to the bandwidth. Two criteria for finding the 'optimal' bandwidth are used, that is $\psi-$CV and usual CV.

n		$\widehat{\theta}^{(2)}_{NB}$ with CV	$\widehat{\theta}^{(2)}_{NB}$ with $\psi-$CV	$\widehat{\theta}^{(2)}_{P}$ with CV	$\widehat{\theta}^{(2)}_{P}$ with $\psi-$CV	$\widehat{\theta}^{(2)}_{N}$ with CV
100	mean	0.993	0.949	1.015	0.958	0.986
	st. dev.	0.722	0.650	0.752	0.663	0.820
	MSE	0.520	0.424	0.564	0.441	0.672
200	mean	1.058	1.001	1.011	0.995	0.991
	st. dev.	0.580	0.497	0.584	0.542	0.635
	MSE	0.339	0.247	0.340	0.293	0.403
400	mean	1.055	1.021	1.051	1.022	1.038
	st. dev.	0.441	0.406	0.459	0.423	0.492
	MSE	0.197	0.165	0.213	0.179	0.243

For any t in the grid, the mean squared error $E[\hat{r}_{\widehat{\theta},\widehat{h}}(t) - r_0(t)]^2$ is estimated by the average of $[\hat{r}_{\widehat{\theta},\widehat{h}}(t) - r_0(t)]^2$ over 500 samples. Let $MSE_I(t)$ and $MSE_{II}(t)$ be the averages corresponding to the iterative procedure I and II, respectively. The curve $t \rightarrow 100 * (MSE_{II}(t)/MSE_I(t) - 1)$ is depicted in Figure 1. It appears that a bandwidth obtained through a $\psi - CV$ function is at least as good as the usual cross-validation bandwidth when used to build the nonparametric estimator of the regression $r_0(\cdot)$.

Figure 1: The difference expressed in percentage between the mean squared errors of the Nadaraya-Watson estimators computed with cross-validation and $\psi-$CV based bandwidths $\widehat{h}$.



6 Conclusions

We introduce a large class of semiparametric $M-$estimators for single-index models and we show their asymptotic normality. The estimates are obtained as maximizers of a criterion

$$\widehat{S}(\theta, h) = \frac{1}{n}\sum_{i=1}^{n} \psi(Y_i, \widehat{r}_{\theta,h}(X_i\theta))\,\tau_n(X_i),$$

where a nonparametric kernel estimator $\widehat{r}_{\theta,h}$ is used to estimate the conditional expectation $r_\theta(\cdot) = E(Y \mid X\theta = \cdot)$. It is well-known that the (first order) asymptotics of $\widehat{\theta} = \arg\min_{\theta\in\Theta} \widehat{S}(\theta, h)$ do not depend on the choice of h, provided that h satisfies some conditions. The decomposition

$$\widehat{S}(\theta, h) = \widetilde{S}(\theta) + T(h) + R(\theta, h)$$

and the order of $R(\theta, h)$ given in (3.3) allows us to derive a large range of values for the smoothing parameter which lead to the same asymptotic law for $\widehat{\theta}$. This range is between

$n^{-1/4}$ and $n^{-1/8}$, if the true regression $r_{\theta_0}(\cdot)$ is twice differentiable. In particular, the optimal rate $n^{-1/5}$ for the bandwidth choice in nonparametric regression is in this range.

In practice, one still has to choose a reliable h. Since the choice of h affects only the higher order asymptotics of an estimator $\widehat{\theta}$, a natural way to deal with the bandwidth choice problem is to find an optimal h for the regression estimation, for example by cross-validation. In the single-index framework this idea was first developed by Härdle, Hall and Ichimura (1993), which, for the particular case of $\psi(y,r) = -(y-r)^2$, proposed to maximize $\widehat{S}(\theta,h)$ jointly in θ and h. This leads to a bandwidth which is asymptotically equivalent to h chosen by cross-validation if θ_0 were known.

In the case of a more general ψ (e.g., quasi-likelihood, maximum likelihood or robust methods), it also seems natural to maximize $\widehat{S}(\theta,h)$ with respect to θ and h. In some sense, this is like considering h an auxiliary parameter for which the $M-$estimation criterion may provide an estimate. We show in this paper that such a choice $\widehat{h}$ for h is optimal for estimating the regression function in the sense that it is equivalent to a bandwidth optimizing a weighted cross-validation criterion. The weights are given by the second derivative of ψ with respect to r (e.g., for the semiparametric quasi-likelihood case, the weights correspond to the inverse of the variance). We prove that $\widehat{h}$ is of order $n^{-1/5}$. The simulation experiments we conducted indicate that $\widehat{h}$ is also preferable for estimating θ_0. Whether our $\widehat{h}$ is optimal for the estimation of θ remains unknown.

The proofs are based on a technical toolbox linking powerful results on $U-$ processes to index regressions. Our technique allows for asymptotic results under weak conditions. In particular, discrete covariates are allowed and no preliminary estimator of θ_0 having a suitable rate of convergence in probability $O_P(n^{-\delta})$, $\delta > 0$, is required.

A Appendix: definitions and assumptions

Assume that Θ is a compact subset of $\mathbb{R}^{d-1}$ with nonvoid interior. Recall that $X\theta$ is a short for $(1,\theta^T)X$.

Assumption 1.1 *The observations* $(Y_1, X_1^T), \ldots, (Y_n, X_n^T)$ *are independent copies of a random vector* $(Y, X^T)^T \in \mathbb{R}^{d+1}$.

Assumption 1.2 *There exists a unique* θ_0 *interior point of* Θ *such that* $E(Y \mid X) = E(Y \mid X\theta_0)$.

Assumption 1.3 *For every* $\theta \in \Theta$, *the random variable* $X\theta$ *admits a density* $f_\theta(\cdot)$ *with respect to the Lebesgue measure on* $\mathbb{R}$.

Assumption 1.4 *There exists* $c_0 > 0$ *and a positive integer* k_0 *such that, for any* $0 < c \leq c_0$ *and* $\theta \in \Theta$, *the set* $\{t : f_\theta(t) = c\}$ *has at most* k_0 *elements.*

The last two assumptions ensure, in particular, that $P(f_{\theta_0}(X\theta_0) = c) = 0$, for any $0 < c \leq c_0$.

Assumption 1.5 $E[\exp(\lambda|X|)] < \infty$, *for some* $\lambda > 0$. *Moreover,* $E(Y^4) < \infty$.

CONDITION L A function $g : \Theta \times \mathbb{R} \to \mathbb{R}$ is said to satisfy *Condition L* if, for any Λ a compact set on the real line, there exists $B > 0$ and $b \in (0, 1]$ such that

$$|g(\theta, t) - g(\theta', t')| \leq B\,|(\theta, t) - (\theta', t')|^{b}, \quad \text{for any} \quad \theta, \theta' \in \Theta, \quad t, t' \in \Lambda.$$

Assumption 1.6 *a) The function $(\theta, t) \mapsto f_\theta(t) \geq 0$, $\theta \in \Theta$, $t \in \mathbb{R}$, satisfies a Lipschitz condition, that is there exists $a \in (0, 1]$ and $C > 0$ such that*

$$|f_\theta(t) - f_{\theta'}(t')| \leq C\ |(\theta, t) - (\theta', t')|^{a} \qquad \textit{for any} \quad \theta, \theta' \in \Theta \qquad \textit{and} \quad t, t' \in \mathbb{R}.$$

b) The function $(\theta, t) \mapsto r_\theta(t)$, $\theta \in \Theta$, $t \in \mathbb{R}$, satisfies Condition L.

c) For any $\theta \in \Theta$, the functions $t \mapsto \gamma_\theta(t)$ and $t \mapsto f_\theta(t)$ are twice differentiable. Let $\gamma''_\theta(t)$ and $f''_\theta(t)$ denote the second order derivatives. The functions $(\theta, t) \mapsto \gamma''_\theta(t)$ and $(\theta, t) \mapsto f''_\theta(t)$, $\theta \in \Theta$, $t \in \mathbb{R}$, satisfy Condition L *with $b = 1$.*

d) For any $\theta \in \Theta$ and any component $X^{(j)}$ of X, the functions $t \mapsto E\left(X^{(j)} \mid X\theta = t\right)$ and $t \mapsto E\left(Y\,X^{(j)} \mid X\theta = t\right)$ are twice differentiable and their second order derivatives satisfy Condition L *with $b = 1$.*

e) For any $t \in \mathbb{R}$, the function $\theta \mapsto r_\theta(t)$ is twice continuously differentiable and, for any $\theta \in \Theta$, the functions $t \mapsto \partial_\theta r_\theta(t)$ and $t \mapsto \partial^2_{\theta\theta} r_\theta(t)$ are continuous. Moreover, the function $(\theta, t) \mapsto \partial_\theta r_\theta(t)$ satisfy Condition L *with $b = 1$.*

Let $v_\theta(t) = var\,(Y|X\theta = t)$ be the conditional variance of Y given $X\theta = t$.

Assumption 1.7 *The function $(\theta, t) \mapsto v_\theta(t)$ satisfies* Condition L.

Consider $\psi : \mathcal{Y} \times R \to \mathbb{R}$, with $\mathcal{Y}$, $R \subset \mathbb{R}$. If $c, \delta > 0$, define $\Lambda = \bigcup_{\theta \in \Theta} \{t : f_\theta(t) \geq c\}$ and

$$D(c, \delta) = \{r : \exists\,(\theta,\, t) \in \Theta \times \Lambda \ \text{such that}\ |r - r_\theta(t)| \leq \delta\}.$$

Assumption 1.8 *If $c > 0$, there exists $\delta > 0$ such that $D(c, \delta)$ is strictly included in R.*

Assumption 1.9 *There exists $F(\cdot)$ such that $\psi(y, r) \leq F(y)$, $\forall r \in R$.*

This assumption allows us to consider $\psi(y, r) \leq 0$, possibly after replacing it with $\psi(y, r) - F(y)$. This condition is fulfilled in all examples we provided above.

Assumption 1.10 *The function $\psi(\cdot, \cdot)$ is twice differentiable in the second argument. For any c and $\delta > 0$ for which $D(c, \delta)$ is strictly included in R, there exists a function $\Psi(\cdot)$ such that*

$$\sup_{r \in D(c,\delta)} \left(|\partial^2_{22}\psi(y, r)| + |\partial_2\psi(y, r)|\right) \leq \Psi(y),$$

$$\sup_{r, r' \in D(c,\delta)} \left|\partial^2_{22}\psi(y, r) - \partial^2_{22}\psi(y, r')\right| \leq \Psi(y)|r - r'|$$

and $E[\Psi(Y)^{4+\varepsilon}]$, for some $\varepsilon > 0$.

Assumption 1.11 *The kernel function $K(\cdot)$ is differentiable, symmetric, positive and compactly supported. Moreover, $K(\cdot)$ and the derivative $K'(\cdot)$ are of bounded variation.*

Assumption 1.12 *The following conditions hold:*

1. *$E\left[\partial_2\psi\left(Y,\ r_{\theta_0}(X\theta_0)\right) \mid X\right] = 0$;*

2. *$E\left[\partial_\theta\partial_2\psi\left(Y,\ r_{\theta_0}(X\theta_0)\right) \mid X\theta_0\right] = 0$.*

Assumption 1.13 *The $(d-1) \times (d-1)$ matrix $W_0 = -E\left[\partial^2_{\theta\theta}\psi\left(Y_i, r_{\theta_0}(X_i\theta_0)\right)\ I_A(X_i)\right]$ is positive definite.*

B Appendix: Technical lemmas

First, let us recall that $\mathcal{F}$, a class of real-valued functions defined on a set $\mathcal{X}$, is Euclidean for the envelope F if there exist positive constants C and V with the following property: if $0 < \varepsilon \leq 1$ and if μ is a measure for which $\int F d\mu < \infty$, then there are functions $f_1, ..., f_k \in \mathcal{F}$ such that $k \leq C\varepsilon^{-V}$ and, for each $f \in \mathcal{F}$, there is an f_i with $\int |f - f_i| \, d\mu \leq \varepsilon \int F d\mu$. Moreover, the constants C and V must not depend on μ. Let us call the functions $f_1, ..., f_k$ approximating functions. Recall also that a class of indicator functions of sets in a class $\mathcal{D}$ is Euclidean (for the envelope $F \equiv 1$) if and only if $\mathcal{D}$ is a Vapnik-Červonenkis (VC) class (cf. Pakes and Pollard (1989)).

Recall that $\mathcal{H}_n$ was defined in (3.1) as the range $\left\{h : \; c_1 n^{-\beta_2} \leq h \leq c_2 n^{-\beta_1}\right\}$, $n \geq 1$, for some fixed $1/8 < \beta_1 < \beta_2 < 1/4$ and $c_1, c_2 > 0$. However, most of the results in the appendices are valid for a larger range of bandwidths. For this reason, consider also the range $H_n = \left[n^{-(1/2-\varepsilon)}, \, n^{-\varepsilon}\right]$, with some small $0 < \varepsilon < 1/2$. Let $\Theta_n = \{\theta : |\theta - \theta_0| \leq d_n\}$, with $\{d_n\}$ some sequence decreasing to zero. We use C to denote a positive constant, not necessarily the same at each occurrence. For simplicity, if $x_1 \in \mathbb{R}^d$ and $x_2 \in \mathbb{R}$, we omit the transpose when writing the vector $\left(x_1^T, x_2\right)^T$. For ease of exposition, we write '$\sup_{a \in A, \, b} F(a, b) = O_P(G(a))$' instead of '$\sup_b F(a, b) = O_P(G(a))$, uniformly in $a \in A$'. A similar notation with $o_P(\cdot)$ is used.

Lemma B.1 *Assume that the kernel K is a symmetric, positive, compactly supported function of bounded variation. Suppose that the map $(\theta, t) \mapsto f_\theta(t) \geq 0$, $\theta \in \Theta$, $t \in \mathbb{R}$, satisfies a Lipschitz condition, that is there exists $a \in (0, 1]$ and $C > 0$ such that*

$$|f_\theta(t) - f_{\theta'}(t')| \leq C \, |(\theta, t) - (\theta', t')|^a \qquad for \quad \theta, \theta' \in \Theta \qquad and \quad t, t' \in \mathbb{R}. \tag{B.1}$$

Then

$$\max_{1 \leq i \leq n} \sup_{\theta, \, x, \, h \in H_n} \left| \widehat{f}^{i}_{\theta,h}(x\theta) - f_\theta(x\theta) \right| = O_P\left(n^{-1/2}h^{-1}\right) + O\left(h^a\right).$$

Proof. Define $f_{\theta,h}(t) = E\left[K_h(t - X\theta)\right]$ and note that

$$\begin{aligned} \sup_{\theta, \, t, \, h \in H_n} |f_{\theta,h}(t) - f_\theta(t)| &= \sup_{\theta, \, t, \, h \in H_n} \left| \int K(u) \left[f_\theta(t + uh) - f_\theta(t)\right] du \right| \\ &\leq \int K(u) \left(\sup_{\theta, \, t, \, h \in H_n} |f_\theta(t + uh) - f_\theta(t)| \right) du \\ &= O\left(h^a\right), \end{aligned}$$

where the last equality is due to (B.1). On the other hand, we have

$$\begin{aligned} \frac{n-1}{n} \left| \widehat{f}^{i}_{\theta,h}(t) - f_{\theta,h}(t) \right| \leq \frac{1}{h} &\left| \frac{1}{n} \sum_{j=1}^{n} K\left(\frac{t - X_j\theta}{h}\right) - E\left[K\left(\frac{t - X\theta}{h}\right)\right] \right| \\ &+ \frac{1}{nh} \left| K\left(\frac{t - X_i\theta}{h}\right) - E\left[K\left(\frac{t - X_i\theta}{h}\right)\right] \right|. \end{aligned}$$

Given the properties of K, the class of functions $\{K\left((t - x\theta)/h\right) : \theta \in \Theta, h \in H_n\}$ is Euclidean for a constant envelope (cf. Lemma 22(ii) of Nolan and Pollard (1987) applied for $\alpha = h^{-1}\theta$ and $\beta = h^{-1}t$). Use, for instance, Corollary 4(ii) of Sherman (1994a), with $k = 1$, and deduce that

$$\max_{1\leq i\leq n} \sup_{\theta,\, x,\, h\in H_n} \left| \widehat{f}^{i}_{\theta,h}\left(x\theta\right) - f_{\theta,h}\left(x\theta\right)\right| = O_P\left(n^{-1/2}h^{-1}\right).$$

■

Now, it becomes more clear why H_n is defined as $\left[n^{-(1/2-\varepsilon)},\, n^{-\varepsilon}\right]$, with $0 < \varepsilon < 1/2$: it is the range for which $\widehat{f}^{i}_{\theta,h}\left(x\theta\right)$ converges to $f_\theta\left(x\theta\right)$ in probability, uniformly in θ, x and h. Note that in the next lemma the definition of Z_n is slightly more general than in section 3 because h belongs to the larger range H_n.

Lemma B.2 *a) If $\delta > 0$, then*

$$\sup_{\theta\in\Theta_n,\, h\in H_n} \left| I_{\left\{x:\, \widehat{f}^{i}_{\theta,h}(x\theta)\geq c\right\}}(X_i) - I_A(X_i)\right| \leq I_{A^\delta}(X_i) + I_{(\delta,\infty)}(Z_n), \qquad 1\leq i\leq n,$$

where $A^\delta = \{x : |f_{\theta_0}\left(x\theta_0\right) - c| \leq \delta\}$ and

$$Z_n = \max_{1\leq i\leq n} \sup_{\theta\in\Theta_n,\, h\in H_n} \left| \widehat{f}^{i}_{\theta,h}\left(X_i\theta\right) - f_{\theta_0}\left(X_i\theta_0\right)\right|.$$

b) Assume that $K\left(\cdot\right)$ and $f_\theta\left(\cdot\right)$ satisfy the assumptions of Lemma B.1 for some $a, C > 0$. Moreover, $E\left[\exp\left(\lambda\left|X\right|\right)\right] < \infty$ for some $\lambda > 0$. Assume $d_n = o\left(1/\ln n\right)$, with d_n from the definition of Θ_n. If $\delta_n \to 0$ such that $\delta_n/n^{-a\varepsilon}$ and $\delta_n\left[d_n \ln n\right]^{-a} \to \infty$, then $I_{(\delta_n,\infty)}(Z_n) = o_P\left(n^{-\alpha}\right)$, $\forall\alpha > 0$.

Proof. a) We have

$$\left| I_{\left\{x:\, \widehat{f}^{i}_{\theta,h}(x\theta)\geq c\right\}}(X_i) - I_A(X_i)\right| \leq I_{\left\{x:\, \widehat{f}^{i}_{\theta,h}(x\theta)\geq c\right\}\setminus A}(X_i) + I_{A\setminus\left\{x:\, \widehat{f}^{i}_{\theta,h}(x\theta)\geq c\right\}}(X_i).$$

For any θ, h and δ, we can write

$$\left\{\widehat{f}^{i}_{\theta,h}\left(X_i\theta\right) \geq c\right\} \setminus A \subset \left\{\, \widehat{f}^{i}_{\theta,h}\left(X_i\theta\right) \geq c,\; f_{\theta_0}\left(X_i\theta_0\right) < c - \delta\right\} \cup \left\{c - \delta \leq f_{\theta_0}\left(X_i\theta_0\right) < c\right\}$$

and

$$A \setminus \left\{\, \widehat{f}^{i}_{\theta,h}\left(X_i\theta\right) \geq c\right\} \subset \left\{\, \widehat{f}^{i}_{\theta,h}\left(X_i\theta\right) < c,\; f_{\theta_0}\left(X_i\theta_0\right) \geq c + \delta\right\} \cup \left\{\, c \leq f_{\theta_0}\left(X_i\theta_0\right) < c + \delta\right\}$$

which proves the inequality.

b) It suffices to prove that $P\left(Z_n > \delta_n\right) \to 0$. Note that, for any x and θ,

$$\left|f_\theta\left(x\theta\right) - f_{\theta'}\left(x\theta'\right)\right| \leq C\left(\left|x\theta - x\theta'\right|^2 + \left|\theta - \theta'\right|^2\right)^{a/2} \leq C\left(1 + |x|\right)^a \left|\theta - \theta'\right|^a.$$

Combine this inequality with the arguments of Lemma B.1 and write

$$\begin{aligned} Z_n &\leq \max_{1\leq i\leq n} \sup_{\theta\in\Theta,\, h\in H_n} \left| \widehat{f}^{i}_{\theta,h}\left(X_i\theta\right) - f_\theta\left(X_i\theta\right)\right| + \max_{1\leq i\leq n} \sup_{\theta\in\Theta_n} \left|f_\theta\left(X_i\theta\right) - f_{\theta_0}\left(X_i\theta_0\right)\right| \\ &\leq \max_{1\leq i\leq n} \sup_{\theta\in\Theta,\, h\in H_n,\, x} \left| \widehat{f}^{i}_{\theta,h}\left(x\theta\right) - f_\theta\left(x\theta\right)\right| + C\left|\theta - \theta_0\right|^a \max_{1\leq i\leq n} \left(1 + |X_i|\right)^a \\ &= O\left(n^{-a\varepsilon}\right) + O_P\left(n^{-\varepsilon}\right) + O\left(d_n^a\right) \max_{1\leq i\leq n} \left(1 + |X_i|\right)^a. \end{aligned}$$

On the other hand, we can write

$$\begin{aligned}P\left(d_n^a \max_{1\leq i\leq n}(1+|X_i|)^a > \delta_n\right) &\leq \sum_{i=1}^n P\left((1+|X_i|)^a > \delta_n/d_n^a\right)\\ &= nP\left[\exp\left(\lambda(1+|X_i|)\right) > \exp\left(\lambda\delta_n^{1/a}/d_n\right)\right]\\ &\leq n\frac{e^\lambda E\left[\exp\left(\lambda\left|X_i\right|\right)\right]}{\exp\left(\lambda\delta_n^{1/a}/d_n\right)}.\end{aligned}$$

Since $\delta_n^{1/a}/d_n \ln n$ and $\delta_n/n^{-a\varepsilon}\to\infty$, deduce that $P\left(Z_n > \delta_n\right)\to 0$. ■

Lemma B.3 *Assume that $E\left[Y^2\right]<\infty$. The kernel function $K(\cdot)$ satisfies the conditions of Lemma B.1. Moreover, $K(\cdot)$ is differentiable with $K'(\cdot)$ of bounded variation. Let $c>0$. Suppose that there exists $a\in(0,1]$ and $C>0$ such that condition (B.1) holds. Assume that, for any $\theta\in\Theta$, the functions $t\mapsto\gamma_\theta(t)$, $t\mapsto f_\theta(t)$, $t\mapsto E\left(X\mid X\theta=t\right)$ and $t\mapsto E\left(Y\,X\mid X\theta=t\right)$ are twice differentiable. Moreover, the second derivatives of these functions satisfy a Lipschitz condition on compacts: if g stands for one of these functions, then for any compact set $D\subset\mathbb{R}$ there exists $b\in(0,1]$, $C>0$, independent of θ, such that, for any $\theta\in\Theta$,*

$$\left|g_\theta''(t_1)-g_\theta''(t_2)\right|\leq C\left|t_1-t_2\right|^b,\qquad t_1,t_2\in D.$$

Consider $d_n=o\left(1/\ln n\right)$, where d_n is the radius of Θ_n, and $H_n=\left[n^{-(1/2-\varepsilon)},\,n^{-\varepsilon}\right]$, with some small $0<\varepsilon<1/2$. Moreover, $E\left[\exp\left(\lambda\left|X\right|\right)\right]<\infty$ for some $\lambda>0$. If ξ stands for f,γ or r, then

$$\max_{1\leq i\leq n}\sup_{\theta\in\Theta_n,\,h\in H_n}\left|\widehat{\xi}_{\theta,h}^i\left(X_i\theta\right)-\xi_\theta\left(X_i\theta\right)\right|I_{\left\{x:\,f_{\theta_0}(x\theta_0)\geq c\right\}}(X_i)=O\left(h^2\right)+O_P\left(n^{-1/2}h^{-1}\right)$$

and

$$\max_{1\leq i\leq n}\sup_{\theta\in\Theta_n,\,h\in H_n}\left|\partial_\theta\widehat{\xi}_{\theta,h}^i\left(X_i\theta\right)-\partial_\theta\xi_\theta\left(X_i\theta\right)\right|I_{\left\{x:\,f_{\theta_0}(x\theta_0)\geq c\right\}}(X_i)=O\left(h^2\right)+O_P\left(n^{-1/2}h^{-2}\right).$$

Proof. Like in Lemma B.2, deduce

$$\left|f_\theta\left(x\theta\right)-f_{\theta'}\left(x\theta'\right)\right|\leq C\left(1+|x|\right)^a\left|\theta-\theta'\right|^a.$$

Next, remark that

$$\left\{x:\,f_{\theta_0}\left(x\theta_0\right)\geq c\right\}\subset\left\{x:\,f_\theta\left(x\theta\right)\geq c/2\right\}\cup\left\{x:C\left(1+|x|\right)^a d_n^a>c/2\right\},\qquad\theta\in\Theta_n.$$

Since $E\left[\exp\left(\lambda\left|X\right|\right)\right]<\infty$, for some $\lambda>0$, and $d_n=o\left(1/\ln n\right)$, we have

$$nP\left(C\left(1+|X|\right)^a d_n^a>c/2\right)\to 0,$$

from which we deduce $I_{\left\{x:\,f_{\theta_0}(x\theta_0)\geq c\right\}}\left(X_i\right)\leq I_{\left\{x:\,f_\theta(x\theta)\geq c/2\right\}}\left(X_i\right)+I_{\left\{x:C(1+|x|)^a d_n^a>c/2\right\}}\left(X_i\right)$, $\theta\in\Theta_n$, with

$$\max_{1\leq i\leq n}I_{\left\{x:C(1+|x|)^a d_n^a>c/2\right\}}\left(X_i\right)=o_P\left(n^{-\alpha}\right),\forall\alpha>0.$$

Thus, we can replace $I_{\left\{x:\, f_{\theta_0}(x\theta_0)\geq c\right\}}\left(X_i\right)$ by $I_{\{x:\, f_\theta(x\theta)\geq c/2\}}\left(X_i\right).$

Next, note that $A\left(\Theta\right) := \bigcup_{\theta\in\Theta}\left\{t : f_\theta\left(t\right)\geq c/2\right\}$, $n\geq 1$ is a bounded set. Indeed, if $A\left(\Theta\right)$ is unbounded, let $\{t_m\}$ and $\{\theta_m\}$ such that $|t_m|\to\infty$ and $f_{\theta_m}\left(t_m\right)\geq c/2$. Let $\theta'\in\Theta$ be a limit point for $\{\theta_m\}$. Use condition (B.1) to deduce that $f_{\theta'}\left(t_m\right)\geq c/4$, if m is sufficiently large, which is impossible because of the uniform continuity of the density $t\mapsto f_{\theta'}\left(t\right)$.

If $f_{\theta,h}\left(t\right) = E\left[K_h\left(t-X\theta\right)\right]$, use a Taylor expansion and the symmetry of K and write

$$f_{\theta,h}\left(t\right)-f_\theta\left(t\right)=\int K(u)\left[f_\theta\left(t+uh\right)-f_\theta\left(t\right)\right]du$$

$$=h^2 f''_\theta\left(t\right)\int u^2K(u)du+h^2\left[f''_\theta\left(\widetilde{t}\right)-f''_\theta\left(t\right)\right]\int u^2K(u)du,$$

with $\widetilde{t}$ between t and $t+uh$. It follows that

$$\sup_{\theta\in\Theta,\, t\in A(\Theta),\, h\in H_n}\left|f_{\theta,h}\left(t\right)-f_\theta\left(t\right)\right|\leq h^2\int u^2K(u)du\left[\sup_{\theta\in\Theta,\, t\in A(\Theta)}\left|f''_\theta\left(t\right)\right|+C\left|uh\right|^b\right].$$

Complete with the arguments in the proof of Lemma B.1 and deduce

$$\max_{1\leq i\leq n}\sup_{\theta\in\Theta_n,\, h\in H_n}\left|\widehat{f}^{i}_{\theta,h}\left(X_i\theta\right)-f_\theta\left(X_i\theta\right)\right|I_{\{x:\, f_\theta(x\theta)\geq c/2\}}(X_i)=O\left(h^2\right)+O_P\left(1/h\sqrt{n}\right).$$

Proceed as for $f_{\theta,h}$ and deduce

$$\sup_{\theta\in\Theta,\, t\in A(\Theta),\, h\in H_n}\left|\gamma_{\theta,h}\left(t\right)-\gamma_\theta\left(t\right)\right|=O\left(h^2\right),$$

with $\gamma_{\theta,h}\left(t\right)=E\left[YK_h\left(t-X\theta\right)\right]$. On the other hand, consider $g_{\theta,t,h}(x,y)=yK((t-x\theta)/h)$ and note that the class of functions $\left\{g_{\theta,t,h} : \theta\in\Theta, t\in A\left(\Theta\right), h\in H_n\right\}$ is Euclidean for an envelope $F(x,y)=C\left|y\right|$ with C such that $K\left(\cdot\right)\leq C$. Then,

$$\begin{aligned}\max_{1\leq i\leq n}\sup_{\theta,\, t\in A(\Theta),\, h\in H_n}\left|\widehat{\gamma}^{i}_{\theta,h}\left(t\right)-\gamma_{\theta,h}\left(t\right)\right|\leq&\frac{n}{\left(n-1\right)h}\left|\frac{1}{n}\sum_{j=1}^{n}g_{\theta,t,h}\left(X_j,Y_j\right)-E\left[g_{\theta,t,h}\left(X,Y\right)\right]\right|\\&+\frac{C}{\left(n-1\right)h}\max_{1\leq i\leq n}\left(\left|Y_i\right|+1\right)\\&=O_P\left(1/h\sqrt{n}\right).\end{aligned}$$

Deduce that

$$\max_{1\leq i\leq n}\sup_{\theta\in\Theta_n,\, h\in H_n}\left|\widehat{\gamma}^{i}_{\theta,h}\left(X_i\theta\right)-\gamma\left(X_i\theta\right)\right|I_{\{x:\, f_\theta(x\theta)\geq c/2\}}(X_i)=O\left(h^2\right)+O_P\left(1/h\sqrt{n}\right).$$

Next, deduce the same result for $\widehat{r}^{i}_{\theta,h}=\widehat{\gamma}^{i}_{\theta,h}\,/\,\widehat{f}^{i}_{\theta,h}$ after writing

$$\widehat{r}^{i}_{\theta,h}-r_\theta=\frac{\widehat{\gamma}^{i}_{\theta,h}}{\widehat{f}^{i}_{\theta,h}}-\frac{\gamma_\theta}{f_\theta}=\frac{1}{\widehat{f}^{i}_{\theta,h}}\left[\widehat{\gamma}^{i}_{\theta,h}-\gamma_\theta\right]-\frac{r_\theta}{\widehat{f}^{i}_{\theta,h}}\left[\widehat{f}^{i}_{\theta,h}-f_\theta\right].$$

The arguments for $\partial_\theta \widehat{\gamma}^i_{\theta,h}$ and $\partial_\theta \widehat{r}^i_{\theta,h}$ are similar and hence omitted. Note only that (see, e.g., Andrews (1995), section 6)

$$\partial_\theta f_\theta (x\theta) = \frac{d}{dt} \left\{ E\left[(x - X) \,|X\theta = t\right] \;f_\theta(t)\right\}\big|_{t=x\theta},$$
$$\partial_\theta \gamma_\theta (x\theta) = \frac{d}{dt} \left\{ E\left[Y\,(x - X) \,|X\theta = t\right] \;f_\theta(t)\right\}\big|_{t=x\theta}.$$

Therefore, at this stage the functions $t \mapsto E\left(X \mid X\theta = t\right)$ and $t \mapsto E\left(Y X \mid X\theta = t\right)$ should satisfy the same conditions as $f_\theta(\cdot)$ and $\gamma_\theta(\cdot)$. ■

REMARK. It is clear from the proofs of Lemmas B.2 b) and B.3 that there is a trade-off between the rate d_n and conditions on the moments of the explanatory variables X. Note that no assumption is required on d_n, the radius of Θ_n, when X is bounded. Indeed, if X lies in a compact, condition (B.1) implies that for any x,

$$|f_\theta(x\theta) - f_{\theta'}(x\theta')| \leq C\,|\theta - \theta'|^a, \qquad \theta, \theta' \in \Theta,$$

with $C > 0$ some constant independent of x. In this case the arguments in the proof of Lemma B.2 b) can be applied for any $\delta_n \to 0$ such that $\delta_n / n^{-a\varepsilon}$ and $\delta_n / d_n^a \to \infty$. Moreover, Lemma B.3 remains true for any $d_n \to 0$.

Lemma B.4 *Assume that Assumption 1.4 holds. Moreover, the functions $t \mapsto f_\theta(t)$, $\theta \in \Theta$, are continuous. Let $A^\delta = \{x : |f_{\theta_0}(x\theta_0) - c| \leq \delta\}$, $c, \delta > 0$. Then:*

a) the class of indicator functions

$$\left\{ (x, \omega) \mapsto I_{\{t:\, f_\theta(t) \geq c\}}(x\theta + \omega h) : \theta \in \Theta, h \in [0,1] \right\},$$

with $(x, \omega) \in \mathbb{R}^d \times [-1, 1]$, is Euclidean for the envelope $F \equiv 1$, provided that $c > 0$ is fixed and sufficiently small. The same remains true for the class $\left\{ x \mapsto I_{\{t:\, f_\theta(t) \geq c\}}(x\theta) : \theta \in \Theta \right\}$, that is in the case where h is fixed equal to zero.

b) the class of indicator functions $\{x \mapsto I_{A^\delta}(x) : 0 < \delta \leq \delta_0\}$ is Euclidean for the envelope $F \equiv 1$, provided that $c, \delta_0 > 0$ are fixed sufficiently small.

Proof. a) The continuity of the functions $t \mapsto f_\theta(t)$, $\theta \in \Theta$, and Assumption 1.4 ensure that any set $\{t : f_\theta(t) \geq c\}$, $\theta \in \Theta$, is a union of at most k_0 intervals, provided that $c > 0$ is sufficiently small. Use Lemma 2.4 (for the space of real-valued linear functions on $\mathbb{R}^{d+1}$) and Lemma 2.5 (ii)-(iii) of Pakes and Pollard (1989) to deduce that the class of all sets of the form $\{(x, \omega) : x\theta + \omega h \in T\}$, with $\theta \in \mathbb{R}^{d-1}$, $h \in [0,1]$ and T a union of at most k_0 intervals on the real line, is a VC class. Thus, the class of sets $\{(x, \omega) :\; f_\theta(x\theta + \omega h) \geq c\}$, $\theta \in \Theta$, $h \in [0,1]$, is a VC class. Similar arguments apply for the second class of functions.

b) Note that any set $\{t : |f_{\theta_0}(t) - c| \leq \delta\}$, $0 < \delta \leq \delta_0$, is a union of at most $2k_0$ intervals, provided that $c, \delta_0 > 0$ are sufficiently small. By the results of Pakes and Pollard (1989) the class of sets A^δ, $0 < \delta \leq \delta_0$, is a VC class. ■

Lemma B.5 *Assume that Assumption 1.4 holds. Moreover, the function $(\theta, t) \mapsto f_\theta(t)$ satisfies condition (B.1). Let $g : \Theta \times \mathbb{R} \to \mathbb{R}$ be a function satisfying* Condition L. *Then the class of functions*

$$\left\{ (x, \omega) \mapsto g\left(\theta, x\theta + \omega h\right) I_{\{t:\, f_\theta(t) \geq c\}}(x\theta + \omega h) : \theta \in \Theta, h \in [0,1] \right\},$$

with $(x,\omega)\in\mathbb{R}^d\times[-1,1]$, *is Euclidean for envelope* $F(x,\omega)=M(1+|x|)^b$, *for some* $M>0$. *The same remains true in the case where* h *is set equal to zero.*

Proof. Denote $(x,\omega)\mapsto I_{\{t:\,f_\theta(t)\geq c\}}(x\theta+\omega h)$ by $I_{\theta,h}(x,\omega)$. By Lemma B.4, the class of functions $I_{\theta,h}(\cdot,\cdot)$, indexed by (θ,h), is Euclidean for the constant envelope equal to one. Take $\varepsilon>0$ and μ such that $\int(1+|x|)^b d\mu<\infty$. In particular, μ is a finite measure. Given ε and μ, let $\left(\widetilde{\theta}_i,\widetilde{h}_i\right)$, $i=1,...,\widetilde{N}$ be the points corresponding to a set of approximating functions for the class of functions $I_{\theta,h}(\cdot,\cdot)$, where $\widetilde{N}\leq C\varepsilon^{-V}$ for some C and V.

Enclose $\Theta\times[0,1]$ in a cube S of side l and partition S in k^d subcubes of side l/k, with k to be determined shortly. For each subcube that intersects $\Theta\times[0,1]$ and for any i, choose arbitrarily a point (θ,h) in the intersection such that $\int|I_{\theta,h}-I_{\widetilde{\theta}_i,\widetilde{h}_i}|d\mu\leq\varepsilon\int d\mu$, when such a point exists. Augment the set of points $\left(\widetilde{\theta}_i,\widetilde{h}_i\right)$ by the points chosen in this way and denote the augmented set by $\{(\theta_j,h_j):j=1,...,N\}$. Note that

$$N\leq C\varepsilon^{-V}\left(1+k^d\right).$$

Each (θ,h) in $\Theta\times[0,1]$ belongs to at least one of the subcubes. By construction, there exists a point (θ_j,h_j) belonging to a same subcube as (θ,h) such that $\int\left|I_{\theta,h}-I_{\theta_j,h_j}\right|d\mu\leq 2\varepsilon\int d\mu$. Moreover, (θ_j,h_j) lies necessarily a distance no greater than $(l/k)\sqrt{d}$ from (θ,h). We can write

$$\begin{aligned}&\left|g(\theta,x\theta+\omega h)I_{\theta,h}(x,\omega)-g(\theta_j,x\theta_j+\omega h_j)I_{\theta_j,h_j}(x,\omega)\right|\\&\leq|g(\theta,x\theta+\omega h)-g(\theta_j,x\theta_j+\omega h_j)|I_{\theta,h}(x,\omega)I_{\theta_j,h_j}(x,\omega)\\&+(|g(\theta,x\theta+\omega h)|+|g(\theta_j,x\theta_j+\omega h_j)|)\left|I_{\theta,h}(x,\omega)-I_{\theta_j,h_j}(x,\omega)\right|\\&=:\Delta_1+\Delta_2.\end{aligned}$$

Since $\Lambda=\bigcup_{\theta\in\Theta}\{t:\,f_\theta(t)\geq c\}$ is a bounded set (see Lemma B.3), there exist $B>0$ and $b\in(0,1]$ (depending only on c) such that

$$\begin{aligned}\Delta_1&\leq B\left(|\theta-\theta_j|^2+|(\theta,h)-(\theta_j,h_j)|^2|(x,\omega)|^2\right)^{b/2}\\&\leq B|(\theta,h)-(\theta_j,h_j)|^b\left[2+|x|^2\right]^{b/2}\\&\leq C_1k^{-b}(1+|x|)^b,\end{aligned}$$

for some $C_1>0$. Deduce that $\int\Delta_1 d\mu\leq C_1\varepsilon\int(1+|x|)^b d\mu$, provided that $k^{-b}\leq\varepsilon$. On the other hand, there exists a constant $C_2>0$ such that

$$\sup_{\theta\in\Theta,\,h\in[0,1]}|g(\theta,x\theta+\omega h)|I_{\theta,h}(x,\omega)\leq C_2,\qquad(x,\omega)\in\mathbb{R}^d\times[-1,1].$$

Deduce that there exist some constant M (depending only on c) such that

$$\Delta_1+\Delta_2\leq M\varepsilon\int(1+|x|)^b d\mu,$$

if k is the smallest integer greater than $\varepsilon^{-1/b}$. This choice of k guarantees that N grows at a rate that characterize a Euclidean class. ■

The proof of the following lemma is an obvious adaptation of the proof of Lemma B.5 and therefore it will be omitted. See also Lemma 2.13 of Pakes and Pollard (1989).

Lemma B.6 *Suppose that Assumption 1.4 holds. Moreover, the function $(\theta, t) \mapsto f_\theta(t)$ satisfies condition (B.1). Consider a function $(\theta, t, y) \mapsto g(\theta, t, y)$, with $\theta \in \Theta$, $t \in \mathbb{R}$ and $y \in \mathcal{Y} \subset \mathbb{R}$, such that for any compact $\Lambda \subset \mathbb{R}$ there exist a real-valued function $B(\cdot)$ and $b \in (0, 1]$ for which*

$$|g(\theta, t, y) - g(\theta', t', y)| \leq B(y)|(\theta, t) - (\theta', t')|^b, \qquad \theta, \theta' \in \Theta,\ t, t' \in \Lambda,\ y \in \mathcal{Y}.$$

Then the family of functions

$$\left\{(x, \omega, y) \mapsto g(\theta, x\theta + \omega h, y) I_{\{t\colon f_\theta(t) \geq c\}}(x\theta + \omega h) : \theta \in \Theta, h \in [0, 1]\right\},$$

with $(x, \omega, y) \in \mathbb{R}^d \times [-1, 1] \times \mathcal{Y}$, is Euclidean for the envelope

$$|g(\theta, x\theta + \omega h, y)| I_{\{t\colon f_\theta(t) \geq c\}}(x\theta + \omega h) + MB(y)(1 + |x|)^b,$$

for some arbitrary θ, h in $\Theta \times [0, 1]$ and some function $B(\cdot)$ and $M > 0$.

Lemma B.7 *Suppose that the map $(\theta, t) \mapsto f_\theta(t) \geq 0$, $\theta \in \Theta$, $t \in \mathbb{R}$, satisfies the Lipschitz condition (B.1) for some $C > 0$ and $a \in (0, 1]$. Then the family*

$$\left\{(x, \omega) \mapsto f_\theta^{-1}(x\,\theta + \omega h)\, I_{\{t\colon f_\theta(t) \geq c\}}(x\,\theta + \omega h) : \theta \in \Theta, h \in [0, 1]\right\},$$

$c > 0$, is Euclidean for the envelope $C'(1 + |x|)^a$, where C' is some positive constant. The same remains true when h is fixed equal to zero.

Proof. First, note that

$$f_\theta^{-1}(t)\; I_{\{t'\colon f_\theta(t') \geq c\}}(t) = [\max(f_\theta(t), c)]^{-1}\; I_{\{t'\colon f_\theta(t') \geq c\}}(t).$$

Next, by little algebra deduce that

$$|1/\max(f_\theta(t), c) - 1/\max(f_{\theta'}(t'), c)| \leq C_1 |(\theta, t) - (\theta', t')|^a,$$

for some $C_1 > 0$. Finally, apply Lemma B.5 above for $g(\theta, t) = [\max(f_\theta(t), c)]^{-1}$. ∎

C Appendix: The expansion of $T(h)$

Denote $(n)_k = n(n-1)...(n-k+1)$ and recall that H_n is the range $\left[n^{-(1/2-\varepsilon)}, n^{-\varepsilon}\right]$, $0 < \varepsilon < 1/2$. Let Θ_n be defined as in Lemmas B.2 and B.3. When X is bounded, see the remark following Lemma B.3. The following lemma is a refined version of a standard result for cross-validation in nonparametric regression (e.g., Härdle and Marron (1985)).

Lemma C.1 *Let $Z_1, Z_2, ...$ be independent copies of a random vector $Z = \left(Y, X^T\right)^T \in \mathbb{R}^{d+1}$. Define $r_\theta\left(t\right) = E\left(Y|X\theta = t\right)$ and $v_\theta\left(t\right) = var\left(Y|X\theta = t\right)$. Suppose that Assumption 1.4 holds, fix some small $c > 0$ and let $\Lambda = \bigcup_{\theta \in \Theta}\{t : f_\theta(t) \geq c\}$. Moreover, suppose that:*

1. *$E(Y^4) < \infty$ and $E\left[\exp\left(\lambda\left|X\right|\right)\right] < \infty$ for some $\lambda > 0$.*
2. *for any $\theta \in \Theta$, the random variable $X\theta$ admits a density f_θ with respect to the Lebesgue measure on $\mathbb{R}$.*
3. *$(\theta, t) \mapsto f_\theta\left(t\right)$, $\theta \in \Theta$, $t \in \mathbb{R}$, satisfies the Lipschitz condition (B.1) for some $C > 0$ and $a \in (0, 1]$.*
4. *$(\theta, t) \mapsto r_\theta\left(t\right)$, $\theta \in \Theta$, $t \in \mathbb{R}$, satisfies* Condition L.
5. *for any $\theta \in \Theta$, the functions $t \mapsto \gamma_\theta\left(t\right) = r_\theta\left(t\right) f_\theta\left(t\right)$ and $t \mapsto f_\theta\left(t\right)$, $t \in \mathbb{R}$, are twice continuously differentiable.*
6. *$(\theta, t) \mapsto f''_\theta\left(t\right)$ and $(\theta, t) \mapsto \gamma''_\theta\left(t\right)$, $\theta \in \Theta$, $t \in \mathbb{R}$, satisfy* Condition L *with $b = 1$.*
7. *$(\theta, t) \mapsto v_\theta\left(t\right)$, $\theta \in \Theta$, $t \in \mathbb{R}$, satisfies* Condition L.

Let $(y, t) \mapsto w_\theta(y, t)$, $\theta \in \Theta$, be a family of functions for which there exist a real-valued function $B(\cdot)$ with $E[B(Y)^{4+\varepsilon}] < \infty$, for some $\varepsilon > 0$, and $b' \in (0, 1]$ such that, for each y

$$|w_\theta(y, t) - w_{\theta'}(y, t')| \leq B(y)|(\theta, t) - (\theta', t')|^{b'}, \qquad \theta, \theta' \in \Theta,\ t, t' \in \Lambda.$$

Moreover, there exist θ and $\widetilde{B}(\cdot)$ such that $sup_{t \in \Lambda}|w_\theta(\cdot, t)| \leq \widetilde{B}(\cdot)$ and $E[\widetilde{B}(Y)^{4+\varepsilon}] < \infty$.

Define

$$U\left(\theta, h\right) = \frac{1}{n}\sum_{i=1}^{n} w_\theta\left(Y_i, X_i\theta\right)\left[\hat{r}^i_{\theta,h}\left(X_i\theta\right) - r_\theta\left(X_i\theta\right)\right]^2 I_{\{x:\, f_\theta(x\theta) \geq c\}}\left(X_i\right), \qquad \theta \in \Theta, h \in H_n,$$

where $\hat{r}^i_{\theta,h}\left(\cdot\right)$ denotes the leave-one-out kernel estimator of $r_\theta\left(\cdot\right)$; the kernel is a continuous probability density function K with the support in $[-1, 1]$. Moreover, K is of bounded variation and symmetric. Then,

$$U\left(\theta, h\right) = -h^4 C_1\left(\theta\right) - \frac{1}{nh} C_2\left(\theta\right) + \rho\left(\theta, h\right),$$

where

$$C_1\left(\theta\right) = \frac{K_1^2}{4}\ E\left\{- w_\theta\left(Y, X\theta\right)\ \left[r''_\theta\left(X_i\theta\right) + \frac{2\ r'_\theta\left(X\theta\right)\ f'_\theta\left(X\theta\right)}{f_\theta\left(X\theta\right)}\right]^2 I_{\{x:\, f_\theta(x\theta) \geq c\}}\left(X\right)\right\},$$

$$C_2\left(\theta\right) = K_2\ E\left\{- \frac{w_\theta\left(Y, X\theta\right)}{f_\theta\left(X\theta\right)}\ v_\theta\left(X\theta\right)\ I_{\{x:\, f_\theta(x\theta) \geq c\}}\left(X\right)\right\}, \tag{C.1}$$

with $K_1 = \int u^2 K\left(u\right) du$, $K_2 = \int K^2\left(u\right) du$ and the reminder $\rho\left(\theta, h\right)$ satisfies

$$\sup_{\theta \in \Theta,\, h \in H_n} \rho\left(\theta, h\right) = o_P\left(h^4 + (nh)^{-1}\right).$$

Proof. For brevity, assume that, for any θ, the set $\{t : f_\theta(t) \geq c\}$ is an interval, necessarily bounded. The extension to the case where $\{t : f_\theta(t) \geq c\}$ is the union of at most k_0 bounded intervals is straight.

First, simplify the notation: when there is no possible confusion, omit the arguments $X_i\theta$ and $x_i\theta$ or replace them by V_i and v_i, respectively. Write $\widehat{r}^i = \widehat{\gamma}^i / \widehat{f}^{\,i}$ and $r = \gamma \,/\, f$ instead of $\hat{r}^i_{\theta,h} = \widehat{\gamma}^i_{\theta,h} / \widehat{f}^i_{\theta,h}$ and $r_\theta = \gamma_\theta / f_\theta$, respectively. Similarly, w is a short for w_θ. Moreover, write I_i instead of $I_{\{x:\, f_\theta(x\theta) \geq c\}}(X_i)$.

By Taylor expansion, $1/\widehat{f}^{\,i\,2} = 1/f^2 - 2[(\overline{f}^i)^{-3}(\widehat{f}^{\,i} - f)]$, with $\overline{f}^i$ between $\widehat{f}^{\,i}$ and f. Thus

$$U(\theta, h) = \frac{1}{n}\sum_{i=1}^{n} a\left(\widehat{\gamma}^{\,i} - r\widehat{f}^{\,i}\right)^2 I_i - \frac{1}{n}\sum_{i=1}^{n} a\ b_n\ \left(\widehat{\gamma}^{\,i} - r\widehat{f}^{\,i}\right)^2 I_i,$$

with $a = wf^{-2}$ and $b_n = 2[(\overline{f}^i)^{-3}(\widehat{f}^{\,i} - f)]$. Clearly, $a\, b_n I_i = o_P(1)$ (see also Lemma B.1). Using the definition of $\widehat{\gamma}^{\,i}$ and $\widehat{f}^{\,i}$,

$$U(\theta, h) = \frac{n-2}{n-1}\, U_1(\theta, h) + \frac{1}{n-1}\, U_2(\theta, h) + \{\textit{terms of smaller order}\}$$

with

$$U_1(\theta, h) = (n)_3^{-1} \sum_{i \neq j \neq l} f_1(Z_i, Z_j, Z_l; \theta, h), \qquad U_2(\theta, h) = (n)_2^{-1} \sum_{i \neq j} f_2(Z_i, Z_j; \theta, h),$$

$$f_1(z_i, z_j, z_l; \theta, h) = a(v_i)\,[y_j - r_\theta(v_i)]\ \ [y_l - r_\theta(v_i)]\ \ K_h(v_i - v_j)\, K_h(v_i - v_l)\, I_i$$

and

$$f_2(z_i, z_j; \theta, h) = a(v_i)\ \ [y_j - r_\theta(v_i)]^2\, K_h^2(v_i - v_j)\ I_i\,. \tag{C.2}$$

The order of $U_1(\theta, h)$. For any $i \neq j$ denote by E_i and E_{ij} the conditional expectation operators $E(\cdot\,|Z_i)$ and $E(\cdot\,|Z_i, Z_j)$, respectively. By the usual decomposition of a $U-$statistics in *degenerate* $U-$statistics [see, e.g., Serfling(1980)], write

$$U_1(\theta, h) = U_n^3 f_{1,3}(\cdot, \cdot, \cdot; \theta, h) + U_n^2 f_{1,2}(\cdot, \cdot; \theta, h) + P_n^1 f_{1,1}(\cdot; \theta, h) + E(f_1),$$

where

$$U_n^3 f_{1,3} = (n)_3^{-1} \sum_{i \neq j \neq l} \{f_1 - [E_{ij} + E_{il} + E_{jl}](f_1) + [E_i + E_j + E_l](f_1) - E(f_1)\}$$

$$U_n^2 f_{1,2} = \frac{1}{(n)_2}[\sum_{i \neq j} E_{ij} + \sum_{i \neq l} E_{il} + \sum_{j \neq l} E_{jl}](f_1) - \frac{2}{n}[\sum_i E_i + \sum_j E_j + \sum_l E_l](f_1) + 3E(f_1)$$

$$P_n^1 f_{1,1} = n^{-1}[\sum_i E_i + \sum_j E_j + \sum_l E_l](f_1) - 3E(f_1).$$

Using the results of Sherman (1994a) on the rates of uniform convergence for degenerate $U-$processes indexed by classes which are Euclidean for a squared integrable envelope, we show that $E(f_1)$ is the dominant term in the decomposition of $U_1(\theta, h)$. To prove the Euclidean property we use results from Nolan and Pollard (1987) (hereafter NP87), Pakes and Pollard (1989) (abbreviated by PP89), Sherman (1994a) and Lemmas B.4 to B.7

above. Let us simply recall that pointwise addition or pointwise product of two Euclidean classes as well as integration with respect to one of the arguments of the functions in an Euclidean class preserve the Euclidean property.

First, note that the family

$$\{(z_i, z_j, z_l) \mapsto K\left[(x_i\theta - x_j\theta)/h\right] K\left[(x_j\theta - x_l\theta)/h\right] : \theta \in \Theta, h \in H_n\}$$

is Euclidean for a constant envelope (apply Lemma 22(ii) of NP87). Next, use Lemmas B.4, B.5 and B.7 above and deduce that the family $\{h^2 f_{1,3}(\cdot,\cdot,\cdot;\theta,h)\}$ is Euclidean for a squared integrable envelope. Apply Corollary 4 of Sherman (1994a) and deduce

$$U_n^3 f_{1,3}(\cdot,\cdot,\cdot;\theta,h) = h^{-2} O_P(n^{-3/2}) = O_P(h^{-2} n^{-3/2}), \tag{C.3}$$

uniformly in $\theta \in \Theta$ and $h \in H_n$.

Next, we look for the order of $U_n^2 f_{1,2}$. Remark that

$$\begin{aligned} E_{ij}(f_1) &= a(V_i)\ \left[Y_j - r(V_i)\right]\ K_h(V_i - V_j)\ E_{ij}\left\{\left[Y_l - r(V_i)\right]\ K_h(V_i - V_l)\right\} I_i \\ &= a(V_i)\ \left[Y_j - r(V_i)\right]\ K_h(V_i - V_j)\, E\left\{\left[r(V_l) - r(V_i)\right]\ K_h(V_i - V_l)\,|V_i\right\} I_i. \end{aligned}$$

If v is fixed, the symmetry of K, a simple change of variables and Taylor expansion yield

$$\begin{aligned} g_1(v;\theta,h) &= E\left\{\left[r_\theta(V_l) - r_\theta(v)\right]\ K_h(v - V_l)\right\} \\ &= \int \gamma_\theta(u) K_h(v-u)\, du - r_\theta(v) \int f_\theta(u) K_h(v-u)\, du \\ &= \int \gamma_\theta(v - \omega h)\, K(\omega) d\omega - r_\theta(v) \int f_\theta(v - \omega h)\, K(\omega) d\omega \\ &= \left(h^2/2\right)\left[\gamma_\theta''(v) - r_\theta(v) f_\theta''(v)\right] \int \omega^2 K(\omega)\, d\omega \qquad\qquad \text{(C.4)} \\ &\quad + \int s_\gamma(v,\omega;\theta,h)\, K(\omega)\, d\omega - r_\theta(v) \int s_f(v,\omega;\theta,h)\, K(\omega)\, d\omega, \end{aligned}$$

where s_γ and s_f are the reminders under integral form of the Taylor expansions of $\gamma_\theta(\cdot)$ and $f_\theta(\cdot)$, respectively, that is,

$$s_L(v,\omega;\theta,h) = \int_v^{v-\omega h} (v - \omega h - s)\ \left[L_\theta''(s) - L_\theta''(v)\right]\, ds, \tag{C.5}$$

with $L = \gamma$ or f. By Condition 6,

$$\begin{aligned} \left|\frac{s_L(v,\omega;\theta,h)}{h^2} - \frac{s_L(v,\omega;\theta,h')}{h'^2}\right| &\leq \left|\int_v^{v-\omega h} \left[\frac{v - \omega h - s}{h^2} - \frac{v - \omega h' - s}{h'^2}\right]\right. \\ &\qquad\qquad \times \left[L_\theta''(s) - L_\theta''(v)\right]\, ds| \\ &\quad + \left|\int_{v-\omega h}^{v-\omega h'} \frac{v - \omega h' - s}{h'^2}\left[L_\theta''(s) - L_\theta''(v)\right]\, ds\right| \\ &\leq 2C(h'-h)|\omega|^3 + C(h'-h)|\omega|^3/2, \qquad \text{(C.6)} \end{aligned}$$

for some $C > 0$, provided that $0 < h \leq h'$. On the other hand, using again condition 6,

$$\begin{aligned} \left|\frac{s_L(v,\omega;\theta,h')}{h'^2} - \frac{s_L(v,\omega;\theta',h')}{h'^2}\right| &\leq \int_v^{v-\omega h'} \frac{|v - \omega h' - s|}{h'^2}\left|L_\theta''(s) - L_{\theta'}''(s)\right|\, ds \\ &\quad + \left|L_\theta''(v) - L_{\theta'}''(v)\right| \int_v^{v-\omega h'} \frac{|v - \omega h' - s|}{h'^2}\, ds \\ &\leq 2C\omega^2|\theta - \theta'|. \qquad \text{(C.7)} \end{aligned}$$

The last two displays show that the families of functions

$$(x,\omega)\mapsto h^{-2}s_L(x\,\theta,\omega;\theta,h)I_{\{x'\colon f_\theta(x'\theta)\geq c\}}\,(x) \tag{C.8}$$

(with $L=\gamma$ or f) indexed by (θ,h), satisfy the Lipschitz condition of Lemma 2.13 of PP89. Deduce that these families are Euclidean for a constant envelope. The Euclidean property is preserved for the last two integrals in (C.4). Finally, apply Lemma B.5 and deduce that the family of functions

$$x\mapsto\left[\gamma''_\theta(x\,\theta)-r_\theta(x\theta)f''_\theta(x\,\theta)\right]I_{\{x'\colon f_\theta(x'\theta)\geq c\}}\,(x)$$

is Euclidean for an envelope $M(1+|x|)^{1+b}$, for some $M>0$ and $b\in(0,1]$. Now, we have all the ingredients necessary to conclude that the family of functions

$$x\mapsto h^{-2}g_1(x\,\theta\,;\theta,h)I_{\{x'\colon f_\theta(x'\theta)\geq c\}}\,(x)\,,$$

indexed by (θ,h) is Euclidean for the envelope $C(1+|x|)^{1+b}$, for some $C>0$ and $b\in(0,1]$. Consequently, $\{\,h^{-1}\,E\,[\,f_1\,(z_i,z_j,\cdot\,;\theta,h)]\}$ is Euclidean for a squared integrable envelope. By similar arguments deduce that $\{\,h^{-1}\,E\,[\,f_1\,(z_i,\cdot\,,z_l;\theta,h)]\}$ is Euclidean for the corresponding squared integrable envelope.

The last term of $U_n^2 f_{1,2}\,(\cdot,\cdot;\theta,h)$ to be studied is $E_{jl}\,(f_1)$. By a change of variable,

$$\begin{aligned}E\left[f_1|\,(Y_j,V_j)=(y_j,v_j)\,,(Y_l,V_l)=(y_l,v_l)\right]&\\ =h^{-1}\int a_\theta(v_j+\omega h)\;\left[y_j-r_\theta\left(v_j+\omega h\right)\right]\;\left[y_l-r_\theta\left(v_j+\omega h\right)\right]&\\ \times K\left(\omega\right)\;K\left[\left(v_j-v_l\right)/h+\omega\right]\;I_{\{t\colon f_\theta(t)\geq c\}}\left(v_j+\omega h\right)\;f_\theta(v_j+\omega h)\;d\omega.&\end{aligned}$$

Apply Lemmas B.4 to B.7 and deduce that the family of functions

$$\begin{aligned}(\omega,z_j,z_l)\mapsto a_\theta(x_j\,\theta+\omega h)\;\left[y_j-r_\theta\left(x_j\,\theta+\omega h\right)\right]\;\left[y_l-r_\theta\left(x_j\,\theta+\omega h\right)\right]\\ \times K\left[\left(x_j\,\theta-x_l\,\theta\right)/h+\omega\right]\;I_{\{t\colon f_\theta(t)\geq c\}}\left(x_j\,\theta+\omega h\right)\;f_\theta(x_j\,\theta+\omega h),\end{aligned}$$

indexed by (θ,h), is Euclidean for a squared integrable envelope. Consider the last integral above as an expectation with respect to the probability defined by $K\left(\omega\right)d\omega$ and deduce that the class $\{\,h\,E\,[\,f_1\,(\cdot,z_j,z_l\,;\theta,h)]\}$ indexed by (θ,h) is Euclidean for a squared integrable envelope.

The previous findings indicate that $\{\,h\,f_{1,2}\,(\cdot,\cdot\,;\theta,h)\,;\,\theta\in\Theta,h\in H_n\}$ is Euclidean for a squared integrable envelope. Moreover, for any θ and $i_1\neq i_2$

$$\sup_{\theta\in\Theta,\,h\in H_n} h\,E\,\left(|f_{1,2}\left(z_{i_1},z_{i_2};\theta,h\right)|\right)\to 0. \tag{C.9}$$

Since

$$\sup_{\theta\in\Theta,\,h\in H_n}\left|a_\theta\left(X_i\theta\right)\left[Y_j-r_\theta\left(X_i\theta\right)\right]\;\left[Y_l-r_\theta\left(X_i\theta\right)\right]\right|I_{\{x\colon f_\theta(x\theta)\geq c\}}\left(X_i\right)$$

is integrable, to prove property (C.9) it suffices to show that

$$\sup_{\theta\in\Theta,\,h\in H_n} h\,E\,\left[K_h\left(X_i\theta-X_j\theta\right)K_h\left(X_j\theta-X_l\theta\right)\right]\to 0,\qquad i\neq j\neq l.$$

For this note that, for any θ, $P\left(X_j\theta - X_l\theta = 0\right) = 0$ and recall that the support of K is bounded. By the same arguments as used in the proof of Corollary 8(ii) of Sherman (1994a), deduce that uniformly over $\Theta \times H_n$,

$$U_n^2 f_{1,2}\left(\cdot,\cdot;\theta,h\right) = o_P(h^{-1}n^{-1}). \tag{C.10}$$

For the order of $P_n^1 f_{1,1}\left(\cdot;\theta,h\right)$ let us write

$$\begin{aligned} E_i\left(f_1\right) = a\left(V_i\right)\ E_i\{\ E\left[\left(Y_j - r\left(V_i\right)\right)\ \mid Z_i, V_j\right]\ E\left[\left(Y_l - r\left(V_i\right)\right)\ \mid Z_i, V_l\right] \\ \times K_h\left(V_i - V_j\right)\ K_h\left(V_i - V_l\right)\} \, I_i \\ = a\left(V_i\right)\ \left[g_1(V_i;\theta,h)\right]^2 I_i, \end{aligned} \tag{C.11}$$

with $g_1(\cdot;\theta,h)$ defined as in (C.4). Deduce that $\{h^{-4}\, E\left[\, f_1\left(z_i,\cdot,\cdot;h\right)\right]\}$ is Euclidean for a constant envelope. On the other hand, by simple algebra we obtain

$$\begin{aligned} E_j\left(f_1\right) &= E_j\left\{E_{ij}\left\{a\left(V_i\right)\ \left[Y_j - r\left(V_i\right)\right]\ \left[Y_l - r\left(V_i\right)\right]\ K_h\left(V_i - V_j\right)\ K_h\left(V_i - V_l\right)\ I_i\right\}\right\} \\ &= E\left\{a\left(V_i\right)\left[Y_j - r\left(V_i\right)\right]\ g_1(V_i\,;\theta,h)\ K_h\left(V_i - V_j\right) I_i \mid Y_j, V_j\right\}. \end{aligned}$$

Moreover, by a change of variables

$$\begin{aligned} & E\left\{a\left(V_i\right)\ \left[Y_j - r\left(V_i\right)\right]\ g_1(V_i;\theta,h)\ K_h\left(V_i - V_j\right) I_i \mid Y_j = y, V_j = v\right\} \\ & = \int g_1(t;\theta,h)\ a(t)\left[y - r_\theta\left(t\right)\right] K_h\left(t - v\right)\ f_\theta(t)\ I_{\{t' \colon f_\theta(t') \geq c\}}\left(t\right) dt \\ = & \int g_1(uh+v;\theta,h) a(uh+v)\left[y - r_\theta\left(uh+v\right)\right] K\left(u\right) f_\theta(uh+v) I_{\{t' \colon f_\theta(t') \geq c\}}\left(uh+v\right) du. \end{aligned}$$

In view of (C.4), write

$$\begin{aligned} g_1(uh+v;\theta,h) = h^2\left(K_1/2\right)\left[\gamma_\theta'' - r_\theta f_\theta''\right](uh+v) \\ + \int s_\gamma(uh+v,\omega;\theta,h)\, K(\omega)\, d\omega \\ - r_\theta(uh+v)\int s_f(uh+v,\omega;\theta,h)\, K(\omega)\, d\omega, \end{aligned}$$

with s_γ and s_f defined as in (C.5). By the same arguments as used for the families written in (C.8), the families

$$(x,u,\omega) \mapsto h^{-2} s_L(uh + x\,\theta, \omega;\theta,h) I_{\{x' \colon f_\theta(x'\theta) \geq c\}}\left(x\right),$$

(with $L = \gamma$ or f) indexed by $\left(\theta,h\right)$, are Euclidean for a constant envelope. Next, after integrating out ω, deduce that

$$(x,u) \mapsto h^{-2} g_1(uh + x\,\theta;\theta,h) I_{\{x' \colon f_\theta(x'\theta) \geq c\}}\left(x\right),$$

is Euclidean for a squared integrable envelope. Finally, after integrating out u, deduce that the family $\{\, h^{-2}\, E\left[\, f_1\left(\cdot, z_j, \cdot\,;\theta,h\right)\right]\}$ is Euclidean for a squared integrable envelope. Similar arguments apply for $E_l\left(f_1\right)$. Use Corollary 4 of Sherman(1994a) to deduce

$$P_n^1 f_{1,1}\left(\cdot;\theta,h\right) = O_P\left(h^2 n^{-1/2}\right), \tag{C.12}$$

uniformly in $(\theta, h) \in \Theta \times H_n$. In view of (C.4), (C.6), (C.7) and (C.11), deduce

$$E_i(f_1) = h^4 a(V_i) \frac{K_1^2}{4} \left[(r_\theta f_\theta)''(V_i) - r_\theta(V_i)\ f_\theta''(V_i) \right]^2 I_{\{x:\, f_\theta(x\theta) \geq c\}}(X_i) + h^4 R_1(V_i; \theta, h),$$

with $|R_1(v; \theta, h)| \to 0$ as $h \to 0$, uniformly in v and θ. Thus, uniformly in $(\theta, h) \in \Theta \times H_n$,

$$E(f_1) = E[E_i(f_1)] = -h^4 C_1(\theta) + o\left(h^4\right).$$

The order of $U_2(\theta, h)$. Write $U_2(\theta, h) = U_n^2 f_{2,2} + P_n^1 f_{2,1} + E(f_2)$, with

$$U_n^2 f_{2,2}(\cdot, \cdot; \theta, h) = (n)_2^{-1} \sum_{i \neq j} \left[f_2(Z_i, Z_j) - E_i(f_2) - E_j(f_2) + E(f_2) \right],$$

$$P_n^1 f_{2,1}(\cdot; \theta, h) = n^{-1} [\sum_i E_i + \sum_j E_j](f_2) - 2E(f_2)$$

and f_2 is defined in (C.2). The order of $U_2(\theta, h)$ is given by $E(f_2)$. Indeed, use the same arguments as for $U_n^3 f_{1,3}$ and deduce that uniformly over $\Theta \times H_n$,

$$U_n^2 f_{2,2}(\cdot, \cdot; \theta, h) = h^{-2} O_P\left(n^{-1}\right) = O_P\left(h^{-2} n^{-1}\right). \tag{C.13}$$

For the order of $P_n^1 f_{2,1}(\cdot, \cdot; \theta, h)$ write

$$\begin{aligned} E_i(f_2) = a(V_i)\ E_i \left\{ E\left\{ [Y_j - r_\theta(V_j) + r_\theta(V_j) - r_\theta(V_i)]^2 \mid V_i, V_j \right\}\ K_h^2(V_i - V_j) \right\} I_i \\ = a(V_i)\ E\left[v_\theta(V_j)\ K_h^2(V_i - V_j) \, |V_i \right] I_i \\ + a(V_i)\ E\left\{ [r_\theta(V_j) - r_\theta(V_i)]^2\ K_h^2(V_i - V_j) \, |V_i \right\} I_i. \end{aligned}$$

Use again a change of variables and arguments as used for $E_{jl}(f_1)$ and deduce that the family $\{ h\, E[\, f_2(z_i, \cdot\,; \theta, h)] \}$ is Euclidean for a squared integrable envelope (use also condition 7). Similar arguments apply to the family $\{ h\, E[\, f_2(\cdot, z_j\,; \theta, h)] \}$. Corollary 4 of Sherman (1994a) yields $P_n^1 f_{2,1}(\cdot; \theta, h) = O_P\left(h^{-1} n^{-1/2}\right)$. Finally, by simple algebra

$$E_i(f_2) = h^{-1} K_2\ a(V_i)\, v_\theta(V_i)\ f_\theta(V_i) + R_2(V_i; \theta, h),$$

with $|R_2(v; \theta, h)| \to 0$ as $h \to 0$, uniformly in v and θ. Consequently,

$$E(f_2) = E[E_i(f_2)] = -h^{-1} C_2(\theta) + o\left(h^{-1}\right),$$

uniformly in $(\theta, h) \in \Theta \times H_n$. Deduce that

$$(n-1)^{-1} U_2(\theta, h) = -n^{-1} h^{-1} C_2(\theta) + o\left(n^{-1} h^{-1}\right),$$

uniformly over $\Theta \times H_n$. Now the proof is complete. ■

The following lemma indicates, in particular, that the order of $T(h)$ is given by $T_2(h)$. For this we have to constrain h to the range $\mathcal{H}_n$ defined in (3.1), that is $\mathcal{H}_n = \left\{ h :\ c_1 n^{-\beta_2} \leq h \leq c_2 n^{-\beta_1} \right\}$, for some fixed $1/8 < \beta_1 < \beta_2 < 1/4$ and $c_1, c_2 > 0$.

Lemma C.2 *Let $Z_1, Z_2, ...$ be independent copies of a random vector $Z = \left(Y, X^T\right)^T \in \mathbb{R}^{d+1}$. Assume that conditions 1 to 6 of Lemma C.1 hold. Moreover, consider a kernel K as in Lemma C.1. Let*

$$\widetilde{T}(\theta, h) = \frac{1}{n} \sum_{i=1}^{n} \ \alpha\left(Z_i\right) \left[\hat{r}_{\theta,h}^{i}\left(X_i\theta\right) - r_\theta\left(X_i\theta\right)\right] \ I_{\left\{x\colon f_{\theta_0}(x\theta_0)\geq c\right\}}\left(X_i\right),$$

with $E\left[\alpha\left(Z\right) \mid X\right] = 0$ and $E\left[\left|\alpha\left(Z\right)\right|^{4+\varepsilon} I_{\left\{x\colon f_{\theta_0}(x\theta_0)\geq c\right\}}\left(X\right)\right] < \infty$, for some $\varepsilon > 0$. Then,

$$\widetilde{T}(\theta, h) = o_P\left(h^4\right) + o_P\left(n^{-1}h^{-1}\right),$$

uniformly in $h \in \mathcal{H}_n$ and in $\theta \in \Theta_n$.

Proof. As in the proof of Lemma B.3, we can replace $I_{\left\{x\colon f_{\theta_0}(x\theta_0)\geq c\right\}}\left(X_i\right)$ by $I_i = I_{\{x\colon f_\theta(x\theta)\geq c/2\}}\left(X_i\right)$. We use the same simplified notation as in the proof of Lemma C.1. Moreover, $\alpha_i = \alpha\left(Z_i\right)$. By Taylor expansion

$$\frac{1}{\widehat{f}^{\,i}} = \frac{1}{f} - \frac{1}{f^2}\left(\widehat{f}^{\,i} - f\right) + O_P\left(\left|\widehat{f}^{\,i} - f\right|^2\right) = \frac{2}{f} - \frac{\widehat{f}^{\,i}}{f^2} + O_P\left(\left|\widehat{f}^{\,i} - f\right|^2\right).$$

Therefore, we can write

$$\widetilde{T}(\theta, h) = \frac{1}{n}\sum_{i=1}^{n}\alpha_i\left(\widehat{f}^{\,i}\right)^{-1}\left(\widehat{\gamma}^{\,i} - r\widehat{f}^{\,i}\right)I_i$$
$$= \frac{2}{n}\sum_{i=1}^{n}\alpha_i f^{-1}\left(\widehat{\gamma}^{\,i} - r\widehat{f}^{\,i}\right)I_i - \frac{1}{n}\sum_{i=1}^{n}\alpha_i f^{-2}\left(\widehat{\gamma}^{\,i} - r\widehat{f}^{\,i}\right)\widehat{f}^{\,i} + \widetilde{R}(h)$$
$$=: \widetilde{T}_1(\theta, h) + \widetilde{T}_2(\theta, h) + \widetilde{R}(h).$$

It is easy to check that the reminder $\widetilde{R}(h)$ has the order $o_P\left(h^4 + n^{-1}h^{-1}\right)$ for $h \in \mathcal{H}_n$ (apply Lemma B.3). More precisely,

$$\widetilde{R}(h) = \left[O\left(h^2\right) + O_P\left(h^{-1}n^{-1/2}\right)\right]^3$$
$$= O\left(h^6\right) + O_P\left(h^3n^{-1/2}\right) + O_P\left(n^{-1}\right) + O_P\left(h^{-3}n^{-3/2}\right).$$

Next,

$$\widetilde{T}_1(\theta, h) = (n)_2^{-1}\sum_{i\neq j}\frac{2\alpha_i}{f\left(V_i\right)}\left(Y_j - r\left(V_i\right)\right)K_h\left(V_i - V_j\right)I_i =: U_n^2 g\left(\cdot,\cdot;\theta,h\right).$$

Since $E\left[\alpha\left(Z_i\right) \mid X_i\right] = 0$, the conditional expectation of g given Z_j, denoted $E_j\left(g\right)$, and $E\left(g\right)$ are null. Hence, the second order $U-$statistics $U_n^2 g$ can be decomposed in degenerated $U-$statistics like

$$U_n^2 g\left(\cdot,\cdot;\theta,h\right) = U_n^2 g_2\left(\cdot,\cdot;\theta,h\right) + P_n^1 g_1\left(\cdot,\cdot;\theta,h\right)$$
$$= (n)_2^{-1}\sum_{i\neq j}\left[g\left(Z_i, Z_j;\theta,h\right) - E_i\left(g\left(Z_i, Z_j;\theta,h\right)\right)\right]$$
$$+n^{-1}\sum_{i=1}^{n}E_i\left(g\left(Z_i, Z_j;\theta,h\right)\right).$$

By similar arguments as in Lemma C.1 deduce that the class $\{hg_2(\cdot,\cdot;\theta,h) : h\in\mathcal{H}_n\}$ is Euclidean for a squared integrable envelope. Moreover, if $i\neq j$,

$$\sup_{\theta\in\Theta_n,\, h\in\mathcal{H}_n} h\,E\left(\left|g_2\left(z_i,z_j;\theta,h\right)\right|\right)\to 0,$$

(see the arguments following equation (C.9)). Corollary 8(ii) of Sherman (1994a) implies $U_n^2 g_2(\cdot,\cdot;\theta,h) = o_P(h^{-1}n^{-1})$. On the other hand,

$$E_i\left(g\right) = \alpha_i f\left(V_i\right)^{-1} E\left[\left(r\left(V_j\right) - r\left(V_i\right)\right) K_h\left(V_i - V_j\right) \mid V_i\right] I_i$$

and thus the arguments used in Lemma C.1 for the function g_1 in equation (C.4) apply again. Deduce that the class $\{\, h^{-2}\, E[\, g(z_i,\cdot\,;\theta,h)] : h\in\mathcal{H}_n,\ \theta\in\Theta_n\}$ is Euclidean for a squared integrable envelope and $P_n^1 g_1(\cdot,\cdot;\theta,h) = O_P(h^2n^{-1/2})$. Thus,

$$\widetilde{T}_1(\theta,h) = o_P\left(n^{-1}h^{-1}\right) + O_P\left(h^2 n^{-1/2}\right)$$

uniformly in $h\in\mathcal{H}_n$, $\theta\in\Theta_n$.

On the other hand,

$$\widetilde{T}_2(\theta,h) = \frac{1}{n}\sum_{i=1}^{n}\frac{\alpha_i}{f^2}(\widehat{\gamma}^{\,i} - r\widehat{f}^{\,i})\widehat{f}^{\,i} I_i$$

$$= \frac{1}{(n)_3}\sum_{i\neq j\neq l}\frac{\alpha_i}{f^2}(Y_j - r(V_i))K_h(V_i - V_j)K_h(V_i - V_l)I_i$$

$$+\frac{1}{(n-2)}\frac{1}{(n)_2}\sum_{i\neq j}\frac{\alpha_i}{f^2}(Y_j - r(V_i))K_h^2(V_i - V_j)I_i.$$

It remains to study the orders of the two U-statistics on the right-hand side of the last display using arguments as in Lemma C.1. For the first one, we obtain the order

$$O_P(h^{-2}n^{-3/2}) + o_P(h^{-1}n^{-1}) + O_P(h^2n^{-1/2})$$

(see also the arguments used to obtain the orders in (C.3), (C.10) and (C.12)), while for the second $U-$statistics to be analyzed the order is $n^{-1}\left[O_P(h^{-2}n^{-1}) + O_P(h^{-1}n^{-1/2})\right]$. Consequently,

$$\widetilde{T}_2(\theta,h) = O_P\left(h^{-2}n^{-3/2} + h^2n^{-1/2} + h^{-2}n^{-2} + h^{-1}\,n^{-3/2}\right) + o_P\left(n^{-1}h^{-1}\right),$$

uniformly in $h\in\mathcal{H}_n$, $\theta\in\Theta_n$. ■

Corollary 3.1 *Under the assumptions of Appendix A,*

$$T(h) = -C_1h^4 - C_2/nh + o_P(h^4 + 1/nh)$$

uniformly in $h\in\mathcal{H}_n$*, with* C_1, C_2 *defined in equation (4.1).*

Proof. By Taylor expansion

$$T\left(h\right) = T_0 + T_1(h) + T_2(h) + T_3(h),$$

with $T_0 = n^{-1}\sum_i \psi\left(Y_i, r_{\theta_0}\left(X_i\theta_0\right)\right) I_A\left(X_i\right)$, $T_1(h)$ and $T_2(h)$ defined in section 3.2, and $T_3(h)$ the reminder. Use Assumption 1.12-1) and Lemma C.2 to deduce that $T_1(h) = o_P(h^4 + 1/nh)$. Moreover, by Lemma C.1, $T_2(h) = -C_1h^4 - C_2/nh + o_P(h^4 + 1/nh)$. Finally, note that the order of $T_3(h)$ is given by the cubic terms $\left|\hat{r}^i_{\theta,h}\left(X_i\theta\right) - r_\theta\left(X_i\theta\right)\right|^3$. Use Lemma B.3 to deduce that $T_3(h) = o_P(h^4 + 1/nh)$. ■

D Appendix: The expansion of $R(\theta, h)$

Proposition D.1 *Let $A = \{x : f_{\theta_0}(x\theta_0) \geq c\} \subset \mathbb{R}^d$, for some $c > 0$. Under the Assumptions of Theorem 4.1*

$$R(\theta, h) = \frac{1}{n}\sum_{i=1}^{n}\left[\psi\left(Y_i, \hat{r}^i_{\theta,h}(X_i\theta)\right) - \psi\left(Y_i, r_\theta(X_i\theta)\right)\right]\ I_A(X_i)$$
$$-\frac{1}{n}\sum_{i=1}^{n}\left[\psi\left(Y_i, \hat{r}^i_{\theta_0,h}(X_i\theta_0)\right) - \psi\left(Y_i, r_{\theta_0}(X_i\theta_0)\right)\right]\ I_A(X_i)$$
$$= \left[O_P\left(h^4\right) + O_P\left(\frac{1}{nh^2}\right) + O_P\left(\frac{h^2}{\sqrt{n}}\right) + O_P\left(\frac{1}{n\sqrt{n}h^4}\right)\right] \times O_P\left(|\theta - \theta_0|\right)$$
$$+\left[O\left(h^2\right) + O_P\left(\frac{1}{\sqrt{n}h^2}\right)\right] \times O_P\left(|\theta - \theta_0|^2\right)$$

when $n \to \infty$, uniformly in $h \in \left[n^{-(1/2-\varepsilon)}, n^{-\varepsilon}\right]$, with $0 < \varepsilon < 1/2$, and uniformly in $\theta \in \Theta_n$. Moreover,

$$R(\theta, h) = o_P\left(n^{-1/2}\ |\theta - \theta_0|\right) + o_P\left(|\theta - \theta_0|^2\right)$$

uniformly in $h \in \mathcal{H}_n$ defined in (3.1) and uniformly in $\theta \in \Theta_n$.

Proof. We use again the same simplified notation as in the proof of Lemma C.1 whenever is possible. Write

$$R(\theta, h) = R_1(\theta, h) - R_1(\theta_0, h) = \partial_1 R_1\left(\overline{\theta}, h\right)\ (\theta - \theta_0),$$

with $\overline{\theta}$ between θ and θ_0, where

$$R_1(\theta, h) = \frac{1}{n}\sum_{i=1}^{n}\left[\psi\left(Y_i, \hat{r}^i_{\theta,h}\right) - \psi\left(Y_i, r_\theta\right)\right]\ I_A(X_i),$$

so that

$$\partial_1 R_1(\theta, h) = \frac{1}{n}\sum_{i=1}^{n}\left[\partial_2\psi\left(Y_i, \hat{r}^i_{\theta,h}\right)\ \partial_\theta \hat{r}^i_{\theta,h} - \partial_2\psi\left(Y_i, r_\theta\right)\ \partial_\theta r_\theta\right]\ I_A(X_i).$$

Recall that whenever is necessary, modulo arbitrarily small terms, $I_A(X_i)$ may be replaced by $I_{\{x:\, f_\theta(x\theta) \geq c/2\}}(X_i)$ (see the proof of Lemma B.3). Let I_i denote any of these indicator functions. We can write

$$\partial_1 R_1(\theta, h) = \frac{1}{n}\sum_{i=1}^{n}\left[\partial_2\psi\left(Y_i, \hat{r}^i\right) - \partial_2\psi\left(Y_i, r\right)\right]\ \left(\partial_\theta \hat{r}^i - \partial_\theta r\right)\ I_i$$
$$+\frac{1}{n}\sum_{i=1}^{n}\partial_2\psi\left(Y_i, r\right)\ \left(\partial_\theta \hat{r}^i - \partial_\theta r\right)\ I_i$$
$$+\frac{1}{n}\sum_{i=1}^{n}\partial_\theta r\ \left[\partial_2\psi\left(Y_i, \hat{r}^i\right) - \partial_2\psi\left(Y_i, r\right)\right]\ I_i$$
$$=: R_{11}(\theta, h) + R_{12}(\theta, h) + R_{13}(\theta, h).$$

The order of $R_{11}(\theta,h)$. By the mean value theorem, $R_{11}(\theta,h)$ can be written as

$$R_{11}(\theta,h) = \frac{1}{n}\sum_{i=1}^{n} \partial_{22}^2\psi\left(Y_i,\overline{r}^i\right)\ \left(\hat{r}^i - r\right)\ \left(\partial_\theta\hat{r}^i - \partial_\theta r\right)\ I_i$$

$$= \frac{1}{n}\sum_{i=1}^{n} \partial_{22}^2\psi\left(Y_i,r\right)\ \left(\hat{r}^i - r\right)\ \left(\partial_\theta\hat{r}^i - \partial_\theta r\right)\ I_i$$

$$+\frac{1}{n}\sum_{i=1}^{n}\left[\partial_{22}^2\psi\left(Y_i,\overline{r}^i\right) - \partial_{22}^2\psi\left(Y_i,r\right)\right]\ \left(\hat{r}^i - r\right)\ \left(\partial_\theta\hat{r}^i - \partial_\theta r\right)\ I_i,$$

with $\overline{r}^i$ between $\hat{r}^i$ and r. The first term can be handled as in Lemma C.1. It is of order $O\left(h^4\right) + O_P\left(n^{-1}h^{-2}\right)$. Note that the bias term is of the same order h^4, since the bias in estimating the regression function or its derivative, under the assumptions we made, is the same, namely h^2. Only the variance term changes, due to the extra h^{-1} coming from the derivation of $K_h(\cdot)$. The second term in the decomposition of R_{11} is similar to the reminder term $\widetilde{R}(h)$ in the decomposition of $\widetilde{T}(h)$ in Lemma C.2. We then obtain

$$R_{11}(\theta,h) = O\left(h^4\right) + O_P\left(n^{-1}h^{-2}\right)$$
$$+\left[O\left(h^2\right) + O_P\left(n^{-1/2}h^{-1}\right)\right]^2 \times \left[O\left(h^2\right) + O_P\left(n^{-1/2}h^{-2}\right)\right]$$
$$= O\left(h^4\right) + O_P\left(n^{-1}h^{-2} + n^{-1/2}h^2 + n^{-3/2}h^{-4}\right),$$

uniformly in $h \in H_n$ and $\theta \in \Theta_n$.

The order of $R_{12}(\theta,h)$. We can write

$$R_{12}(\theta,h) = \frac{1}{n}\sum_{i=1}^{n}\partial_2\psi\left(Y_i, r_{\theta_0}\left(X_i\theta_0\right)\right)\ \left[\partial_\theta\widehat{r}^i_{\theta,h}\left(X_i\theta\right) - \partial_\theta r_\theta\left(X_i\theta\right)\right]\ I_i$$
$$+\frac{1}{n}\sum_{i=1}^{n}\left[\partial_2\psi\left(Y_i, r_\theta\left(X_i\theta\right)\right) - \partial_2\psi\left(Y_i, r_{\theta_0}\left(X_i\theta_0\right)\right)\right]$$
$$\times\ \left[\partial_\theta\widehat{r}^i_{\theta,h}\left(X_i\theta\right) - \partial_\theta r_\theta\left(X_i\theta\right)\right]\ I_i$$
$$=: R_{121}(\theta,h) + R_{122}(\theta,h).$$

Use Assumption 1.12-1) and mimic the arguments in Lemma C.2 in order to show that

$$R_{121} = o_P\left(n^{-1}h^{-2}\right) + O_P\left(h^2n^{-1/2} + n^{-3/2}h^{-4}\right) + O\left(h^6\right).$$

In fact, R_{121}, which at a first look has the same order as $h^{-1}\widetilde{T}_1(\theta,h)$ (we write this $R_{121} \approx h^{-1}\widetilde{T}(\theta,h)$), can be written as

$$R_{121} = o_P\left(\frac{1}{nh^2}\right) + O_P\left(\frac{h^2}{\sqrt{n}}\right) \qquad \left\{\approx \frac{1}{h}\widetilde{T}_1(\theta,h)\right\}$$
$$+O_P\left(\frac{1}{h^3n\sqrt{n}}\right) + o_P\left(\frac{1}{nh^2}\right) + O_P\left(\frac{h^2}{\sqrt{n}}\right) \qquad \left\{\approx \frac{1}{h}\widetilde{T}_{21}(\theta,h)\right\}$$
$$+O_P\left(\frac{1}{h^3n^2}\right) + O_P\left(\frac{1}{h^2n\sqrt{n}}\right) \qquad \left\{\approx \frac{1}{h}\widetilde{T}_{22}(\theta,h)\right\}$$
$$+\left[O\left(h^2\right) + O_P\left(\frac{1}{h\sqrt{n}}\right)\right]^2\left[O\left(h^2\right) + O_P\left(\frac{1}{h^2\sqrt{n}}\right)\right]. \qquad \left\{\approx \frac{1}{h}\widetilde{R}(h) \approx R_{112}\right\}$$

$\widetilde{T}_{21}(\theta,h)$ and $\widetilde{T}_{22}(\theta,h)$ denote the first and the second sum in the expression of $\widetilde{T}_2(\theta,h)$ appearing in the proof of Lemma C.2. Similarly, R_{112} stands for the second sum appearing in the expression of R_{11}. For example, the first term in this decomposition is, crudely speaking, $O_P\left(h^{-1}\widetilde{T}_1(\theta,h)\right)$, which would mean to be of order $o_P\left(n^{-1}h^{-2}\right)+O_P\left(n^{-1/2}h\right)$. But the $O_P\left(n^{-1/2}h^2\right)$ term in $\widetilde{T}_1(\theta,h)$ is a degenerated $U-$statistics of order 1 with h^2 coming from the bias of a kernel estimator. The corresponding term in R_{121} has thus the same order $O_P\left(n^{-1/2}h^2\right)$, and not only $O_P\left(n^{-1/2}h\right)$, since it involves a bias term for a derivative kernel estimator which is still of order h^2.

Next, use Assumptions 1.6, 1.10 and Lemma B.3 to obtain that

$$R_{122}=\left[O\left(h^2\right)+O_P\left(n^{-1/2}h^{-2}\right)\right]\,O_P\left(|\theta-\theta_0|\right).$$

The order of $R_{13}(\theta,h)$. Write $R_{13}(\theta,h)=R_{13}(\theta_0,h)+\left[R_{13}(\theta,h)-R_{13}(\theta_0,h)\right]$. We show that $R_{13}(\theta_0,h)$ has the same order as $\widetilde{T}(\theta_0,h)$ in Lemma C.2 plus a $o_P\left(|\theta-\theta_0|\right)$ term, while $R_{13}(\theta,h)-R_{13}(\theta_0,h)=o_P\left(|\theta-\theta_0|\right)$. First,

$$\begin{aligned}
R_{13}(\theta_0,h) &= \frac{1}{n}\sum_{i=1}^{n}\partial_\theta r_{\theta_0}(X_i\theta_0)\,\left[\partial_2\psi\left(Y_i,\hat{r}^i_{\theta_0,h}(X_i\theta_0)\right)-\partial_2\psi\left(Y_i,r_{\theta_0}(X_i\theta_0)\right)\right]\,I_i\\
&= \frac{1}{n}\sum_{i=1}^{n}\partial_\theta r_{\theta_0}(X_i\theta_0)\,\partial^2_{22}\psi\left(Y_i,r_{\theta_0}(X_i\theta_0)\right)\left[\hat{r}^i_{\theta_0,h}(X_i\theta_0)-r_{\theta_0}(X_i\theta_0)\right]\,I_i\\
&\quad+\frac{1}{n}\sum_{i=1}^{n}\partial_\theta r_{\theta_0}(X_i\theta_0)\,\left[\partial^2_{22}\psi\left(Y_i,\overline{r}_{\theta_0}(X_i\theta_0)\right)-\partial^2_{22}\psi\left(Y_i,r_{\theta_0}(X_i\theta_0)\right)\right]\\
&\qquad\times\,\left[\hat{r}^i_{\theta_0,h}(X_i\theta_0)-r_{\theta_0}(X_i\theta_0)\right]\,I_i\\
&=:R_{131}(\theta_0,h)+R_{132}(\theta_0,h),
\end{aligned}$$

where $\overline{r}_{\theta_0}(X_i\theta_0)$ is between $\hat{r}^i_{\theta_0,h}(X_i\theta_0)$ and $r_{\theta_0}(X_i\theta_0)$. Use Assumption 1.12-2) and argue as in Lemma C.2 to show that

$$R_{131}(\theta_0,h)=o_P\left(n^{-1}h^{-1}\right)+O_P\left(h^2n^{-1/2}+n^{-3/2}h^{-3}\right)+O\left(h^6\right).$$

Then, use Assumptions 1.6, 1.10 and Lemma B.3 to obtain that

$$R_{132}(\theta_0,h)=\left[O\left(h^2\right)+O_P\left(h^{-1}n^{-1/2}\right)\right]\,O_P\left(|\theta-\theta_0|\right).$$

It remains to study the order of

$$\begin{aligned}
R_{13}(\theta,h)-R_{13}(\theta_0,h) &= \frac{1}{n}\sum_{i=1}^{n}\partial_\theta r_{\theta_0}(X_i\theta_0)\,I_i\\
&\qquad\times\left\{\left[\partial_2\psi\left(Y_i,\hat{r}^i_{\theta,h}(X_i\theta)\right)-\partial_2\psi\left(Y_i,r_\theta(X_i\theta)\right)\right]\right.\\
&\qquad\quad\left.-\left[\partial_2\psi\left(Y_i,\hat{r}^i_{\theta_0,h}(X_i\theta_0)\right)-\partial_2\psi\left(Y_i,r_{\theta_0}(X_i\theta_0)\right)\right]\right\}\\
&\quad+\frac{1}{n}\sum_{i=1}^{n}\left[\partial_\theta r_\theta(X_i\theta)-\partial_\theta r_{\theta_0}(X_i\theta_0)\right]\\
&\qquad\times\,\left[\partial_2\psi\left(Y_i,\hat{r}^i_{\theta,h}(X_i\theta)\right)-\partial_2\psi\left(Y_i,r_\theta(X_i\theta)\right)\right]\,I_i
\end{aligned}$$

This can be done by applying the mean value theorem and using Assumption 1.6 and 1.10 and Lemma B.3, in order to obtain

$$R_{13}(\theta, h) - R_{13}(\theta_0, h) = O_P(|\theta - \theta_0|) \times \left[O\left(h^2\right) + O_P\left(n^{-1/2}h^{-2}\right)\right].$$

The proof of the first identity for $R(\theta, h)$ is complete. Now, the order of $R(\theta, h)$ when $h \in \mathcal{H}_n$ is obvious. ■

E Appendix: Preliminary estimate for θ_0

A preliminary estimator of θ_0 can be easily obtained using a fixed trimming (see also Härdle, Hall and Ichimura (1993)). Fix some small $c > 0$ and let B be a subset of $\mathbb{R}^d$ such that $f_\theta(x\theta) \geq c > 0$, $x \in B, \theta \in \Theta$. Define

$$\theta_n = \underset{\theta \in \Theta}{\arg\max} \frac{1}{n}\sum_{i=1}^{n} \psi\left(Y_i, \hat{r}^i_{\theta,h}(X_i\theta)\right) I_B(X_i), \tag{E.1}$$

with $h \in H_n = \left[n^{-(1/2-\varepsilon)},\, n^{-\varepsilon}\right]$ for some small $0 < \varepsilon < 1/2$. To ensure consistency for θ_n, we have to check that

$$\theta_0 = \underset{\theta \in \Theta}{\arg\max}\, E\left[\psi(Y, r_\theta(X\theta))\; I_B(X)\right], \tag{E.2}$$

and θ_0 is unique with this property. In all examples we have in mind, the SIM condition that specifies θ_0 as the unique vector in Θ satisfying $E[Y \mid X] = E[Y \mid X\theta_0]$ implies that, for any x,

$$\theta_0 = \underset{\theta \in \Theta}{\arg\max}\, E\left[\psi(Y, r_\theta(x\theta))\right]$$

and θ_0 is the unique maximizer. This is a version of the so-called conditional Fisher consistency assumption (e.g., Kunsch, Stefanski and Carroll (1989)) which, in particular, implies the identification condition (E.2). The other ingredient for proving consistency is the convergence in probability of the objective function. Let $H_n = \left[n^{-(1/2-\varepsilon)},\, n^{-\varepsilon}\right]$, with some small $0 < \varepsilon < 1/2$. We prove that,

$$\frac{1}{n}\sum_{i=1}^{n} \psi\left(Y_i, \hat{r}^i_{\theta,h}(X_i\theta)\right) I_B(X_i) \to E[\psi(Y, r_\theta(X\theta))\, I_B(X)], \tag{E.3}$$

uniformly in $\theta \in \Theta$ and $h \in H_n$.

Proposition E.1 *(A consistent preliminary estimator) Assume that $E[Y^2] < \infty$ and condition (E.2) holds. Let the kernel $K(\cdot)$ be as in Lemma B.1. Moreover, $(\theta, t) \mapsto f_\theta(t)$ and $(\theta, t) \mapsto \gamma_\theta(t)$ satisfy* Condition L. *Consider $\psi : \mathcal{Y} \times R \to \mathbb{R}$, with $\mathcal{Y}$, $R \subset \mathbb{R}$ such that*

i) $\psi(\cdot, \cdot)$ is twice differentiable in its second argument.

ii) there exists $\delta > 0$ such that the set

$$D_{B,\delta} = \{r : \exists\, (\theta, x) \in \Theta \times B \; such\ that\; |r - r_\theta(x\theta)| < \delta\}$$

is strictly included in R.

iii)

$$\sup_{\theta\in D_{B,\delta}} \left(\left|\psi\left(y,r\right)\right| + \left|\partial_2\psi\left(y,r\right)\right|\right) \leq \Psi\left(y\right)$$

for some $\delta > 0$ *and some squared integrable function* $\Psi\left(\cdot\right).$

Then $\theta_n \to \theta_0$, *in probability. If, in addition, for any* $t \in \{x\theta : x \in B, \theta \in \Theta\}$, *the function* $\theta \mapsto r_\theta\left(t\right)$ *is twice continuously differentiable and the* $(d-1)\times(d-1)$ *matrix* $W_0 = -E\left[\partial^2_{\theta\theta}\psi\left(Y_i, r_{\theta_0}\left(X_i\theta_0\right)\right)\ I_B\left(X_i\right)\right]$ *is positive definite, then* $\theta_n - \theta_0 = O_P\left(n^{-a\varepsilon/2}\right).$

Proof. Let us write

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \psi\left(Y_i, \hat r^i_{\theta,h}\left(X_i\theta\right)\right)\ I_B(X_i) - E[\psi\left(Y, r_\theta\left(X\theta\right)\right) I_B(X)] \\
= &\ \frac{1}{n}\sum_{i=1}^n \left[\psi\left(Y_i, \hat r^i_{\theta,h}\left(X_i\theta\right)\right) I_B\left(X_i\right)\ - \psi\left(Y_i, r_\theta\left(X_i\theta\right)\right) I_B\left(X_i\right)\right] \\
&+\frac{1}{n}\sum_{i=1}^n \psi\left(Y_i, r_\theta\left(X_i\theta\right)\right) I_B\left(X_i\right) - E\left(\psi\left(Y, r_\theta\left(X\theta\right)\right) I_B\left(X\right)\right) \\
= &\ : \widehat{S}_1\left(\theta,h\right) + \widehat{S}_2\left(\theta\right).
\end{aligned}$$

By Taylor expansion, $\left|\widehat{S}_1\left(\theta,h\right)\right|$ is bounded by

$$\left[\max_{1\leq i\leq n}\sup_{\theta,\,h\in H_n}\left|\hat r^i_{\theta,h}\left(X_i\theta\right) - r_\theta\left(X_i\theta\right)\right| I_B\left(X_i\right)\right]\frac{1}{n}\sum_{i=1}^n \left|\partial_2\psi\left(Y_i,\overline{r}_i\right)\right| I_B\left(X_i\right),$$

with $\overline{r}_i$ somewhere between $\hat r^i_{\theta,h}\left(X_i\theta\right)$ and $r_\theta\left(X_i\theta\right)$. Given the assumptions, the sum in the last display is bounded in probability. Next, in view of the proofs of Lemmas B.1 and B.3, deduce that

$$\max_{1\leq i\leq n}\sup_{\theta,\,h\in H_n}\left|\hat r^i_{\theta,h}\left(X_i\theta\right) - r_\theta\left(X_i\theta\right)\right| I_B\left(X_i\right) = O_P\left(n^{-a\varepsilon}\right).$$

It follows that $\sup_{\theta\in\Theta,\,h\in H_n}\widehat{S}_1\left(\theta,h\right) = O_P\left(n^{-a\varepsilon}\right).$

For the uniform convergence of $\widehat{S}_2\left(\theta\right)$, use a uniform law of large numbers (e.g., Pakes and Pollard (1989)). The family of functions $\{(x,y)\mapsto\psi\left(y, r_\theta\left(x\theta\right)\right) I_B\left(x\right),\ \theta\in\Theta\}$ admit the integrable envelope Ψ. Moreover, this family is Euclidean for this envelope (see Lemma B.6). Deduce that $\sup_\theta\left|\widehat{S}_2\left(\theta\right)\right| = o_P\left(1\right)$. The uniform convergence of $\widehat{S}_1\left(\theta,h\right) + \widehat{S}_2\left(\theta\right)$ ensures $\theta_n - \theta_0 = o_P\left(1\right).$

For the second part, use a Taylor expansion and deduce that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \psi\left(Y_i, \hat r^i_{\theta,h}\left(X_i\theta\right)\right)\ I_B(X_i) &= \widehat{S}_1\left(\theta,h\right) + \frac{1}{n}\sum_{i=1}^n \psi\left(Y_i, r_\theta\left(X_i\theta\right)\right) I_B\left(X_i\right) \\
&= O_P\left(n^{-a\varepsilon}\right) + \frac{1}{n}\sum_{i=1}^n \psi\left(Y_i, r_{\theta_0}\left(X_i\theta_0\right)\right) I_B\left(X_i\right) \\
&\quad + \frac{1}{\sqrt{n}}\left(\theta-\theta_0\right)^T V_n - \frac{1}{2}\left(\theta-\theta_0\right)^T\ W_n\ \left(\theta-\theta_0\right) + o_P\left(\left|\theta-\theta_0\right|^2\right),
\end{aligned}$$

uniformly over $o_P(1)$ neighborhoods of θ_0, where

$$V_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \partial_\theta \psi \left(Y_i, r_{\theta_0}(X_i\theta_0)\right) I_B(X_i), \qquad W_n = -\frac{1}{n} \sum_{i=1}^{n} \partial_{\theta\theta}^2 \psi \left(Y_i, r_{\theta_0}(X_i\theta_0)\right) I_B(X_i).$$

Use Theorem 1 of Sherman (1994b) and deduce $\theta_n - \theta_0 = O_P\left(n^{-a\varepsilon/2}\right)$. ■

REFERENCES

Andrews, D.W.K. (1995). Nonparametric kernel estimation for semiparametric models. *Econometric Theory*, **11**, 560-596.

Bonneu, M. and Gba, M. (1998). Estimation semi-paramétrique de quasi-score. *Bull. Belg. Math. Soc.*, **5**, 693-712.

Bosq, D. and Lecoutre, J-P. (1987). *Théorie de l'estimation fonctionnelle.* Economica, Paris.

Carroll, R.J., Fan, J., Gijbels, I. and Wand, M.P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.*, **92**, 477-489.

Chen, H. (1995). Asymptotically efficient estimation in semiparametric generalized linear models. *Ann. Statist.*, **23**, 1102-1129.

Chiou, J.-M. and Müller, H.-G. (1998). Quasi-likelihood regression with unknown link and variance functions. *J. Amer. Statist. Assoc.*, **93**, 1376-1387.

Chiou, J.-M. and Müller, H.-G. (1999). Nonparametric quasi-likelihood. *Ann. Statist.*, **27**, 36-64.

Clark, R.M. (1975). A calibration curve for radio carbon dates. *Antiquity*, **49**, 251-266.

Delecroix, M. and Hristache, M. (1999). M-estimateurs semi-paramétriques dans les modèles à direction révélatrice unique. *Bull. Belg. Math. Soc.*, **6**, 161-185.

Delecroix, M. and Hristache, M., Patilea, V. (2003). On semiparametric M-estimation in single-index regression. Working Paper 2003-19, Université d'Orléans (http://www.univ-orleans.fr/DEG/LEO).

Fraiman, R., Yohai, V.J. and Zamar, R.H. (2001). Optimal robust M-estimates of location. *Ann. Statist.*, **29**, 194-223.

Gouriéroux, C. Monfort, A. and Trognon, A. (1984). Pseudo maximum likelihood methods: theory. *Econometrica*, **52**, 681-700.

Härdle, W., Hall, P and Ichimura, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.*, **21**, 157-178.

Härdle, W. and Marron, J.S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.*, **13**, 1465-1481.

Härdle, W. and Stoker, T.M. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.*, **84**, 986-995.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). *Robust Statistics. The Approach Based on Influence Functions.* John Wiley & Sons, New York.

Hristache, M., Juditsky, A., Polzehl, J. and Spokoiny, V. (2001). Structure adaptive approach for dimension reduction. *Ann. Statist.*, **29**, 1537-1566.

Hristache, M., Juditsky, A. and Spokoiny, V. (2001). Direct estimation of the index coefficient in a single-index model. *Ann. Statist.*, **29**, 595-623.

Huh, P.J. and Park, B.U. (2002). Likelihood-based local polynomial fitting for single-index models. *J. Multivariate Anal.*, **80,** 302-321.

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

Ichimura, H and Lee, L.-F. (1991). Semiparametric least squares estimation of multiple index models: single equation estimation. In *Nonparametric and Semiparametric Methods in Statistics and Econometrics,* W. A. Barnett, J. Powell and G. Tauchen, eds., Cambridge University Press, Ch. 1.

Klein, R.W. and Spady, R.H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica*, **61**, 387-421.

Kunsch, H.R., Stefanski, L.A. and Carroll, R.J. (1989). Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. *J. Amer. Statist. Assoc.,* **84**, 460-466.

McCullagh, P and Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman and Hall, London.

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

Nolan, D and Pollard, D. (1987). $U-$processes : rates of convergence. *Ann. Statist.*, **15**, 780-799.

Pakes, A. and Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica*, **57**, 1027-1057.

Picone, G.A. and Butler, J.S. (2000). Semiparametric estimation of multiple equation models. *Econometric Theory*, **16**, 551-575.

Powell, J.L., Stock, J.M. and Stoker, T.M. (1989). Semiparametric estimation of index coefficients. *Econometrica*, **57**, 1403-1430.

Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics.* Wiley, New-York.

Sherman, R.P. (1994a). Maximal inequalities for degenerate $U-$processes with applications to optimization estimators. *Ann. Statist.*, **22**, 439-459.

Sherman, R.P. (1994b). U-processes in the analysis of a generalized semiparametric regression estimator. *Econometric Theory*, **10**, 372-395.

Stone, C.J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, **10**, 1040-1053.

Vajda, I. (1989). *Theory of Statistical Inference and Information.* Kluwert, Dordrecht, Boston.

Weisberg, S. and Welsh, A.H. (1994). Adapting for the missing link. *Ann. Statist.*, **22**, 1674-1700.

Xia, Y. and Härdle, W. (2002). Semi-parametric estimation of generalized partially linear single-index models. Discussion paper n°2002-56, SFB373.

Xia, Y. and Li, W.K. (1999). On single-index coefficient regression models. *J. Amer. Statist. Assoc.*, **94**, 1275-1285.

Xia, Y., Tong, H. and Li, W.K. (1999). On extended partially linear single-index models. *Biometrika*, **86**, 831-842.

Xia, Y. , Tong, H. and Li, W.K. (2002). Single-index volatility models and estimation. *Statist. Sinica*, **12**, 785-799.