Semi-parametric estimation in a single-index model with endogenous variables

Melanie Birke *
Institute of Mathematics
Universität Bayreuth

Sébastien Van Bellegem
CORE
Université catholique de Louvain

Ingrid Van Keilegom †
Institute of Statistics, Biostatistics and Actuarial Sciences
Université catholique de Louvain

June 4, 2016

Abstract

We consider a semiparametric single-index model, and suppose that endogeneity is present in the explanatory variables. The presence of an instrument is assumed that is non-correlated with the error term. We propose an estimator of the parametric component of the model, which is the solution of an ill-posed inverse problem. The estimator is shown to be asymptotically normal under certain regularity conditions. A simulation study is conducted to illustrate the finite sample performance of the proposed estimator.

Key Words: Endogeneity; ill-posed inverse problem; instrumental variable; semiparametric regression; single-index model; Tikhonov regularization.

Running Title: Single-index models under endogeneity

*Part of this research has been carried out during a stay at the Université catholique de Louvain, financed by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement No. 203650. A further part has been financed by SFB 823, Project C4: "Reconstruction of time variable distributions in statistical inverse problems" of the German Research Foundation.

†Research supported by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement No. 203650, by IAP research network grant nr. P7/06 of the Belgian government (Belgian Science Policy), and by the contract "Projet d'Actions de Recherche Concertées" (ARC) 11/16-039 of the "Communauté française de Belgique" (granted by the "Académie universitaire Louvain").

1 Introduction

In this paper we consider the problem of estimating a semiparametric single-index regression model, when it is assumed that (some of) the explanatory variables are endogenous. Endogeneity is a central issue when modeling statistical data coming from human or medical sciences, and occurs when some of the independent variables in a regression model are correlated with the error term. It can arise when relevant explanatory variables are omitted from the model, as a result of sample selection errors or when unobserved subject selection occurs in experimental studies. The textbook by Hayashi (2000) is an excellent introduction into the problem of endogeneity and how to cope with it in identification, estimation or testing problems.

When endogeneity is present, ordinary regression techniques produce biased and inconsistent estimators. A possible way out is to make use of so-called 'instrumental variables'. These are variables that are not part of the original model, they are correlated with the endogenous explanatory variables conditional on the other covariates, and they cannot be correlated with the error term in the model (i.e. the instruments do not suffer from the same problem as the original explanatory variables).

We illustrate this concept by means of a textbook example taken from Wooldridge (2008). Consider the following model to estimate the effects of several variables, including cigarette smoking, on the weight of newborns:

$$\log(bwght) = \theta_0 + \theta_1 male + \theta_2 parity + \theta_3 \log(faminc) + \theta_4 packs + U,$$

where *male* is a binary indicator equal to one if the child is male; *parity* is the birth order of this child; *faminc* is family income; *packs* is the average number of packs of cigarettes smoked per day during pregnancy. The variable *packs* is likely to be correlated with omitted but important factors to explain the weight *bwght*. Among the omitted variables we think of health factors that are not necessarily easy to measure in a statistical survey. Hence, *packs* and U might be correlated. If the coefficient β_4 is estimated by common least squares techniques, the resulting estimator might thus be biased. A possible instrumental variable for *packs* suggested in Wooldridge (2008) is the average price of cigarettes in the state of residence, *cigprice*. That variable is likely to be uncorrelated with e.g. individual's health factors but it is certainly correlated with individual decisions to consume some quantity of cigarette packs.

Many other examples can be found in the literature, see e.g. Angrist & Krueger (2001), Manzi et al. (2014) or Johannes et al. (2013). Detecting sources of endogeneity and appro-

priate instrumental variables is a difficult empirical issue and it sometimes leads to animated debates. The purpose of this paper is not to enter into this discussion. Instead it aims at studying the interesting statistical challenges encountered when endogeneity arises in *semi-parametric* regression.

Throughout the paper we suppose that a random vector (X, Y) satisfies the following single-index model:

$$Y = h(X^t\vartheta) + U, \tag{1}$$

where the vector of covariates X is endogenous, i.e. the error term U is correlated with X (or equivalently $\mathbb{E}(U|X) \neq 0$), but we assume that there exists a vector of instruments W such that $\mathbb{E}(U|W) = 0$. We suppose that Y is one-dimensional, X is k-dimensional and W is q-dimensional. The data consist of an i.i.d. sample (W_i, X_i, Y_i) $(i = 1, \ldots, n)$, having the same distribution as the vector (W, X, Y). The function $h : \mathbb{R} \to \mathbb{R}$ and the parameter vector $\theta \in \mathbb{R}^k$ are unknown. The true unknown link function is denoted by h_0, the true unknown parameter vector by ϑ.

A number of approaches exist in the literature to identify regression models with endogenous variables. We adopt here the 'inverse problem'-approach, and develop conditions under which a certain operator is invertible, leading to the existence and uniqueness of a solution of model (1). Recent references on this approach in a fully nonparametric setting include the work by Hall & Horowitz (2005), Cavalier & Golubev (2006), Cavalier (2008), Johannes (2009), Darolles et al. (2011), Johannes et al. (2011), Bissantz et al. (2013) and Hildebrandt et al. (2014) to name but a few.

The estimator of ϑ we propose in this paper will be the solution of a certain system of equations, depending on an estimator of the unknown link function h_0. To prove the weak consistency and asymptotic normality of this estimator, we will make use of Chen et al. (2003). In this paper high-level conditions are developed under which a parameter estimator that is defined via an estimating equation depending on a nonparametric nuisance function, is consistent and asymptotically normal. Although some of these conditions require substantial amount of work when verified for particular models, their result offers the advantage of giving the framework of the proof. One does not need to start the proof from zero, but it suffices in fact to fill in the missing steps in the general proof. We will check each of these high-level conditions for our model.

The above single-index model has been studied very extensively in the absence of endogeneity, see e.g. Powell et al. (1989), Ichimura & Lee (1991), Ichimura (1993), Klein & Spady (1993), Härdle et al. (1993) and Carroll et al. (1997), for some of the fundamental papers on

estimation and inference for this model, and e.g. Hristache et al. (2001), Yin & Cook (2002), Delecroix et al. (2006), Kong & Xia (2007), Lin & Kulasekera (2007), Horowitz (2009), Liang et al. (2010), Wang et al. (2010), Zhang et al. (2010), Peng & Huang (2011), Xia et al. (2012) and Ma & Zhu (2013) for some of the more recent contributions. The literature is however limited when endogeneity is present in the explanatory variables. A general theory of inference using sieves for semiparametric models in the presence of endogeneity has been recently initiated by the seminal work of Ai & Chen (2003). The above single-index model belongs to the class of models considered in the latter paper. The results we derive below are different from existing work in several aspects. First, we use kernel-based estimators instead of sieves. Next, our estimator is exploiting the particular structure of the single-index model. Finally, our estimating view is original because inference relates to an ill-posed inverse problem for which we propose a regularization procedure. As far as we know, it is the first work where a regularization technique is combined with inference for endogenous single-index models.

We also note that there exists a (limited) literature on other semiparametric regression models with endogenous variables, e.g. Chen & Pouzo (2009) for semiparametric inference with nonsmooth residuals, Florens et al. (2012) for instrumental regression in partially linear models, and Vanhems & Van Keilegom (2013) for a control function approach to deal with endogeneity in semiparametric transformation models.

The paper is organized as follows. In the next section, we introduce some notations, and we propose estimators for the unknown link function h_0 and the unknown vector of regression parameters ϑ. In Section 3 the asymptotic normality of the estimator of ϑ is formulated, and we also give the conditions under which this result is valid. In Section 4 we present the results of a simulation study, in which we study the performance of the proposed estimator for small samples. Some general conclusions, possible extensions and lines of further research are discussed in Section 5. Finally, the proof of the main asymptotic result is given in the Appendix.

2 Estimation

Denote the densities of X and W by f_X and f_W respectively. The support of X, which is supposed to be a compact subset of $\mathbb{R}^k$, is denoted by $\mathcal{X}$. The parameter vector θ lives in a known compact set $\Theta \subset \mathbb{R}^k$. For identifiability reasons we suppose that $\theta_1 = 1$, which is by no means restrictive, since we can always arrange the order of the covariates in such a way that the first covariate has a non-zero effect on the response. We can therefore write

$\Theta = \{1\} \times \Theta_1$, where we assume that Θ_1 is a compact subset of $\mathbb{R}^{k-1}$. The link function h belongs to a Sobolev space $\mathcal{H}$ of degree 2, i.e.

$$\mathcal{H} = W_2(\Omega) = \Big\{ h : \Omega \to \mathbb{R} \,;\, h, h' \text{ are absolutely continuous, } h'' \in L_2(\Omega) \\ \text{and } h(\underline{\Omega}) = 0 = h'(\overline{\Omega}) \Big\},$$

where Ω is a compact subset of $\mathbb{R}$ containing the support of $X^t\theta$ for all $\theta \in \Theta$, $L_2(\Omega) = \{h : \Omega \to \mathbb{R} \,;\, \int_\Omega |h(z)|^2 dz < \infty\}$, and $\underline{\Omega}$ and $\overline{\Omega}$ are the lower and upper endpoint of Ω. We equip the space $\mathcal{H}$ with the following norm:

$$\|h\|_{\mathcal{H}}^2 = \int_\Omega h^2(z)\, dz.$$

Let $r(w) = \mathbb{E}(Y|W = w) f_W(w)$ for $w \in \mathbb{R}^q$, and for $\theta \in \Theta$ and $h \in \mathcal{H}$ define the operator

$$\begin{aligned} T_\theta : \mathcal{H} \to L_2(\mathbb{R}^q) : h \mapsto T_\theta h &= \mathbb{E}\Big(h(X^t\theta) | W = \cdot \Big) f_W(\cdot) \\ &= \int_\Omega h(z) f_{X^t\theta, W}(z, \cdot) dz. \end{aligned}$$

For each $\theta \in \Theta$, define the following functions:

$$\begin{aligned} h^\circ_{\theta,\alpha} &= \arg\min_{h \in \mathcal{H}} \Delta(h, \theta, \alpha) \\ h^\circ_\theta &= \arg\min_{h \in \mathcal{H}} \Delta(h, \theta, 0), \end{aligned}$$

where

$$\Delta(h, \theta, \alpha) = \int_{\mathbb{R}^q} \Big(T_\theta h(w) - r(w) \Big)^2 dw + \alpha \int_\Omega |h''(z)|^2 dz,$$

and α is a sequence of positive real numbers (possibly depending on n). Note that by convexity of the maps $h \mapsto \Delta(h, \theta, \alpha)$ and $h \mapsto \Delta(h, \theta, 0)$, the above functions are uniquely defined on $\mathcal{H}$. Remark also that for $\theta = \vartheta$, $h^\circ_\vartheta = h_0$ where ϑ is the true parameter.

Next, we will propose an estimator for ϑ. First of all, an estimator of the unknown operator T_θ can be obtained by kernel smoothing:

$$\widehat{T}_\theta h(w) = \frac{1}{n} \sum_{i=1}^n \int_{-\infty}^{+\infty} k_{b_Z}(X_i^t\theta - z) K_{b_W}(W_i - w) h(z) dz,$$

where b_Z and b_W are appropriate bandwidth sequences, k is a one-dimensional kernel, $k_{b_Z}(u) = b_Z^{-1} k(u/b_Z)$, $K(w) = \prod_{j=1}^q k(w_j)$ is a product kernel of dimension q, and $K_{b_W}(w) = b_W^{-q} K(w/b_W)$. An estimator of r is given by

$$\widehat{r}(w) = \frac{1}{n} \sum_{i=1}^n Y_i K_{b_W}(W_i - w).$$

For $h \in \mathcal{H}$ and $\theta \in \Theta$ define the criterion function:

$$\Delta_n(h, \theta, \alpha) = \int_{\mathbb{R}^q} \Big(\widehat{T}_\theta h(w) - \widehat{r}(w) \Big)^2 dw + \alpha \int_\Omega |h''(z)|^2 dz, \tag{2}$$

and let

$$\widehat{h}_{\theta,\alpha} = \arg\min_{h \in \mathcal{H}} \Delta_n(h, \theta, \alpha).$$

We are now ready to define the estimator of ϑ, which will be expressed as a Z-estimator as in Chen et al. (2003). Define the following criterion function for $h \in \mathcal{H}$ and $\theta \in \Theta$:

$$M(h, \theta) = \mathbb{E}\Big[m(W, X, Y, h, \theta) \Big],$$

where

$$m(W, X, Y, h, \theta) = g_\theta(W) \Big(Y - h(X^t \theta) \Big),$$

and where $g_\theta(W) = (g_{1,\theta}(W), \ldots, g_{\ell,\theta}(W))^t$ is a suitable ℓ-dimensional vector of weights, with $\ell \geq k$. Note that the weight function is fixed and not chosen from the data, but we allow a dependence on θ. Although the weight function g_θ can be any function, it can be chosen in an optimal way; see Remark 3 below for more details. Furthermore, for any choice of g_θ, we have that

$$M(h_0, \vartheta) = 0,$$

since $\mathbb{E}(U|W) = 0$. Now, define the empirical counterpart of this criterion function:

$$M_n(h, \theta) = \frac{1}{n} \sum_{i=1}^n m(W_i, X_i, Y_i, h, \theta).$$

Finally, let

$$\widehat{\theta} = \arg\min_{\theta \in \Theta} \| M_n(\widehat{h}_{\theta,\alpha}, \theta) \|, \tag{3}$$

where $\|A\| = (\mathrm{tr}(A^t A))^{1/2}$ is the Euclidean norm for any matrix (and in particular any vector) A.

3 Asymptotic results

We need to introduce a few additional notations. Let L be the second order derivative operator, defined by

$$L : \mathcal{H} \to L_2(\Omega) : g \mapsto Lg = -g'',$$

and let $S_\theta = T_\theta L^{-1}$ and $\widehat{S}_\theta = \widehat{T}_\theta L^{-1}$. It is well known, that on $\mathcal{H}$ L is one-to-one, see e.g. Florens, Johannes and van Bellegem (2011). Further, let

$$\begin{aligned} \Gamma &= -\mathbb{E}\Big[g_\vartheta(W)h_0'(X^t\vartheta)X^t\Big] - \mathbb{E}\Big[g_\vartheta(W)\Big(\frac{\partial}{\partial\theta^t}h_\theta^\circ\Big)\Big|_{\theta=\vartheta}(X^t\vartheta)\Big] \\ \Sigma &= \Big(\int \sigma^2(w)\xi_{j_1}(w)\xi_{j_2}(w)f_W(w)dw\Big)_{1\le j_1,j_2\le \ell} \end{aligned} \tag{4}$$

where $\sigma^2(w) = \mathrm{Var}(U|W=w)$, $\xi(w) = (\xi_1(w),\ldots,\xi_\ell(w))^t$,

$$\xi(w) = g_\vartheta(w) - \int g_\vartheta(\omega)\{(T_\vartheta^\star T_\vartheta)^{-1} f_{X^t\vartheta,W}(\cdot,w)\}(z) f_{X^t\vartheta,W}(z,\omega)\,dz\,d\omega,$$

and where $T_\vartheta^\star$ is the adjoint operator of T_ϑ. Also, for $\alpha \geqslant 0$, $s>0$ and $p \geq 1$, let $\mathfrak{G}^{s,\alpha}(\mathbb{R}^p)$ be the space of functions $f:\mathbb{R}^p\to\mathbb{R}$ satisfying:

1. f is everywhere $(m-1)$ times partially differentiable for $m-1 < s \leqslant m$ and $m\in\mathbb{N}$;

2. for some $\kappa>0$ and for all x, the inequality

$$\sup_{y:\|y-x\|\le\kappa} \frac{|f(y)-f(x)-Q_x(y-x)|}{\|y-x\|^s} \leqslant \psi(x), \tag{5}$$

   holds true where $Q_x \equiv 0$ when $m=1$ and $Q_x(z) = \sum_{0<j_1+\ldots+j_p\le m-1} \frac{\partial^{j_1+\ldots+j_p}f(x)}{\partial x_1^{j_1}\ldots\partial x_p^{j_p}} \big(\prod_{i=1}^p z_i^{j_i}\big)$ for any z when $m>1$;

3. ψ is uniformly bounded by a constant when $\alpha=0$ and the functions f and ψ satisfy $\int f^\alpha(x)\,dx<\infty$ and $\int \psi^\alpha(x)\,dx<\infty$ when $\alpha>0$.

The asymptotic results of this section will be valid under the following assumptions:

(A.1) If $T_{\theta_1}h_1 = T_{\theta_2}h_2$ for some $\theta_1,\theta_2\in\Theta$ and $h_1,h_2\in\mathcal{H}$, then $\theta_1=\theta_2$ and $h_1=h_2$ (identification condition).

(A.2) For all $\theta\in\Theta$, there exist $\gamma_\theta>0$ and a function $\psi_\theta\in L_2(\mathbb{R})$ such that $h_\theta^\circ = (T_\theta^\star T_\theta)^{\gamma_\theta/2}\psi_\theta$, $\sup_\theta\int\psi_\theta^2(z)dz<\infty$ and $\gamma = \inf_\theta\gamma_\theta<\infty$ (source condition).

(A.3) Each explanatory variable X_j, $j=1,\ldots,k$ has a density belonging to $\mathfrak{G}^{1,1}(\mathbb{R})\cap\mathfrak{G}^{s_1,2}(\mathbb{R})$ for some $s_1\geqslant 1$. Moreover, $\sup_{\theta,z} f_{X^t\theta}(z)<\infty$ and $f_{X^t\theta,W}$ belongs to $\mathfrak{G}^{1,1}(\mathbb{R}^{q+1})\cap\mathfrak{G}^{s_2,2}(\mathbb{R}^{q+1})$ for all $\theta\in\Theta$ and for some $s_2\geqslant 1$.

(A.4) The kernel function k is a symmetric, twice continuously differentiable probability density of order $p\geq 2$.

(A.5) $nb_W^{2\rho} \to 0$, $nb_Z^{2p} \to 0$, $nb_W^{2q} b_Z \to \infty$, $n\alpha^{2\gamma \wedge 4} \to 0$, $(b_W \vee b_Z)^{2\rho}\alpha^{-2} \to 0$ and $nb_W^q b_Z \alpha^2 \to \infty$, where γ is defined in assumption (A.2), $\rho = p \wedge s_2$, and p and s_2 are defined in assumptions (A.4) and (A.3) respectively.

(A.6) (a) $h_\theta^\circ \in \mathcal{H}$ for all $\theta \in \Theta$ and h_θ° is p times continuously differentiable with respect to θ, where p is defined in assumption (A.4).
(b) The matrix Γ is of full rank and the matrix Σ is positive definite.

(A.7) The function g_θ satisfies $\sup_\theta \mathbb{E}\|g_\theta(W)\|^2 < \infty$ and is continuously differentiable with respect to θ. Moreover, for all $\delta > 0$ there exists $\epsilon > 0$ such that $\inf_{\|\theta-\vartheta\|>\delta} \|M(h_\theta^\circ, \theta)\| > \epsilon$.

(A.8) For all $\theta \in \Theta$, the operators S_θ and $\widehat{S}_\theta$ are compact, and the derivatives of their eigenfunctions are uniformly bounded with probability 1. Furthermore, for all $\theta \in \Theta$ and $s = 0, 1, 2$, the operators $L^{s/2}T_\theta$ and $L^{s/2}\widehat{T}_\theta$ are compact with singular systems $\{(\lambda_{\theta,s,k}, \phi_{\theta,s,k}, \psi_{\theta,s,k})\}_{k\in\mathbb{N}_0}$ respectively $\{(\widehat{\lambda}_{\theta,s,k}, \widehat{\phi}_{\theta,s,k}, \widehat{\psi}_{\theta,s,k})\}_{k\in\mathbb{N}_0}$ such that

(i) $||L^{s/2}h_\theta^\circ|| < \infty$, $\mathrm{P}(||L^{s/2}\widehat{h}_{\theta,\alpha}|| < C_{s,1}) \to 1$ for $n \to \infty$, some constants $C_{s,1} > 0$, and $s = 0, 1$, and $||L\widehat{h}_{\theta,\alpha}|| = O_P(1)$.

(ii) $\sup_{z,\theta} \sum_k |\psi_{\theta,s,k}(z)| < \infty$ and $\mathrm{P}(\sup_{z,\theta} \sum_k |\widehat{\psi}_{\theta,s,k}(z)| < C_{s,2}) \to 1$ for $n \to \infty$, some constant $C_{s,2} > 0$ and $s = 0, 1$.

(iii) $\sup_{z,\theta}(\sum_k |\psi_{\theta,2,k}(z)|^2) < \infty$ and $\sup_{z,\theta}(\sum_k |\widehat{\psi}_{\theta,2,k}(z)|^2) = O_P(1)$.

(iv) There exist functions ℓ_s, $\widehat{\ell}_s \in L^1(\mathbb{N})$, $s = 0, 1$, independent of θ, bounded and monotone decreasing such that

$$\begin{aligned} |\lambda_{\vartheta,s,l}\langle\phi_{\theta,s,k}, \phi_{\vartheta,s,l}\rangle| &\leq C_{s,3}\widehat{\lambda}_{\theta,s,k}\ell_s(|k-l|) \\ \mathrm{P}\Big(|\lambda_{\vartheta,s,l}\langle\widehat{\phi}_{\theta,s,k}, \phi_{\vartheta,s,l}\rangle| &\leq C_{s,3}\widehat{\lambda}_{\theta,s,k}\widehat{\ell}_s(|k-l|)\Big) \to 1 \text{ for } n \to \infty, \end{aligned}$$

and for some $C_{s,3} < \infty$.

Remark 1 Note that the source condition in (A.2) can be seen as an assumption on the smoothness of the density from which the expectation operator is defined in T_θ. A thorough discussion of the source condition and its connection to smoothness assumptions can be found in Johannes et al. (2011). The essential point is that the smoothness of the density of $X^t\theta$ matters here. But the smoothness of the density of $X^t\theta$ is itself related to the smoothness of the density of the covariates. Assumptions (A.3)–(A.8) are rather classical regularity conditions on the smoothness of certain underlying functions, on the bandwidth sequences,

on the kernel function, and on the boundedness, non-singularity and compactness of certain quantities, matrices and operators, respectively. To check the compactness of S_θ or $\hat{S}_\theta$ in (A.8), note that they are both of the form

$$\int_\Omega g(z,\cdot)(L^{-1}h)(z)dz$$

with either $g(z,w) = f_{X^t\theta,W}(z,w)$ or $g(z,w) = \hat{f}_{X^t\theta,W}(z,w) = n^{-1}\sum_{i=1}^n k_{b_Z}(X_i^t\theta - z)K_{b_W}(W_i - w)$. So both operators are similar to Hilbert-Schmidt operators and it can be shown, that similar to Hilbert-Schmidt operators they are compact if

$$\int\int |g(z,w)|^2 dzdw < \infty.$$

For $g = f_{X^t\theta,W}$ this simply means that $f_{X^t\theta,W} \in L_2(\Omega \times \mathbb{R}^q)$ while for $g = \hat{f}_{X^t\theta,W}$ the condition can be reduced to $k \in L_2(\Omega)$ and $K \in L_2(\mathbb{R}^q)$.

Finally, the identification condition for θ in (A.1) is common in semiparametric models. In our case with endogenous variables, it means that $T_{\theta_1}h_1 = T_{\theta_2}h_2$ implies that $\theta_1 = \theta_2$ and $h_1 = h_2$, or equivalently that

$$\int h_1(x^t\theta_1) f_{X|W}(x|\cdot)\, dx = \int h_2(x^t\theta_2) f_{X|W}(x|\cdot)\, dx \tag{6}$$

implies that $\theta_1 = \theta_2$ and $h_1 = h_2$. First, note that (6) implies that $h_1(\cdot^t\theta_1) = h_2(\cdot^t\theta_2)$ if the family $\{f_{X|W}(\cdot|w) : w \in R_W\}$ is complete (see Section 2 in Newey & Powell (2003) and Hu & Shiu (2011) for the definition of a complete family of conditional densities). Examples of conditional densities that are complete, are the exponential family and the conditional normal family and are also presented in the above references. Next, thanks to the identifiability of the single-index model we know that $h_1(\cdot^t\theta_1) = h_2(\cdot^t\theta_2)$ implies that $h_1 = h_2$ and $\theta_1 = \theta_2$.

We are now ready to give an i.i.d. expansion of the estimator $\widehat{\theta}$, from which its asymptotic normality will follow immediately.

Theorem 1. Assume (A.1)–(A.8). Then, we have:

$$\widehat{\theta} - \vartheta = \frac{1}{n}\sum_{i=1}^n (\Gamma^t\Gamma)^{-1}\Gamma^t U_i \xi(W_i) + o_P(n^{-1/2}),$$

where $U_i = Y_i - h_0(X_i^t\vartheta)$. Hence,

$$n^{1/2}(\widehat{\theta} - \vartheta) \xrightarrow{d} N(0, V),$$

where

$$V = (\Gamma^t\Gamma)^{-1}\Gamma^t\Sigma\Gamma(\Gamma^t\Gamma)^{-1},$$

and where the function $\xi(\cdot)$ and the matrices Γ and Σ are defined in (4).

Remark 2 Note that when $\ell = k$ (i.e. when the function g_θ contains as many components as there are parameters in the model), the formula of the asymptotic variance reduces to $V = \Gamma^{-1}\Sigma(\Gamma^t)^{-1}$, since Γ is a square (invertible) matrix in that case.

Also note that the estimation of the asymptotic variance might be cumbersome in practice. The estimation of the matrix Γ is still manageable (although it involves the estimation of derivatives of the function h_0), but the estimation of the function $\xi(\cdot)$, which appears in the formula of Σ, is more problematic. In practice, it might therefore be more convenient to estimate the matrix V by means of a bootstrap procedure. Chen et al. (2003) give sufficient high level conditions under which a naive bootstrap procedure is consistent for the estimation of the distribution of $\widehat{\theta}$. We refer to their paper for more details.

A nice feature of our result is that the (first order) asymptotic distribution does not depend on the bandwidths b_W and b_Z. This is similar to the exogenous case (see e.g. Härdle et al. (1993)).

Remark 3 It can be easily seen that the optimal (theoretical) choice of the function $g_\theta(\cdot)$ is given by $g_\theta(\cdot) = \mathbb{E}(h'(X^t\theta)X|W = \cdot)\mathrm{Var}(U|W = \cdot)^{-1}$; see e.g. Florens et al. (2004), Section 17.5.3 p. 440 for a detailed derivation.

4 Numerical aspects and simulations

In this section we discuss the numerical aspects and the finite sample behavior of the proposed estimator. We first define a data generating process on which the performance of the estimator will be based.

The nonparametric function considered in this analysis is $h_0(z) = \sin[2\pi(1-z)^2]$ and is defined over $\Omega = [0, 1]$. This function is twice differentiable on the interval $[0, 1]$ and satisfies the border conditions $h_0(0) = 0$ and $h_0'(1) = 0$. A single-index model with five covariates is considered and constructed as follows. First we generate iid joint samples $X = (X_1, \ldots, X_5)^\top$

from

$$\begin{aligned} X_1 &= W + U + V_1 \\ X_2 &= W + 2U + V_2 \\ X_3 &= W^2 + U - 1 + V_3 \\ X_4 &= W^2 + 2U + V_4 \\ X_5 &= W^2 + \frac{1}{2}U + V_5, \end{aligned}$$

where $U, W, V_1, \ldots, V_5$ are independent zero-mean Normal random variables with variance 0.5. We fix the vector with parameters of interest to $\vartheta = (1, -.3, .8, .5, .9)^\top$ and consider the random variable $Z = X^\top \vartheta$. Since the realizations of this variable are not concentrated into the interval $[0, 1]$, we consider a linear transformation $\tilde{X}$ of the vector X such that the random variable $\tilde{Z} = \tilde{X}^\top \vartheta$ belongs to $[0, 1]$ with probability higher than 0.95. To achieve this, we first consider the vector X as if it was a Normal vector (which is only true in approximation). By doing so we approximate the random variable Z by a Normal distribution with mean μ_Z and standard deviation σ_Z. After a calculation assuming the Normal approximation, we arrive at the following transformation of X :

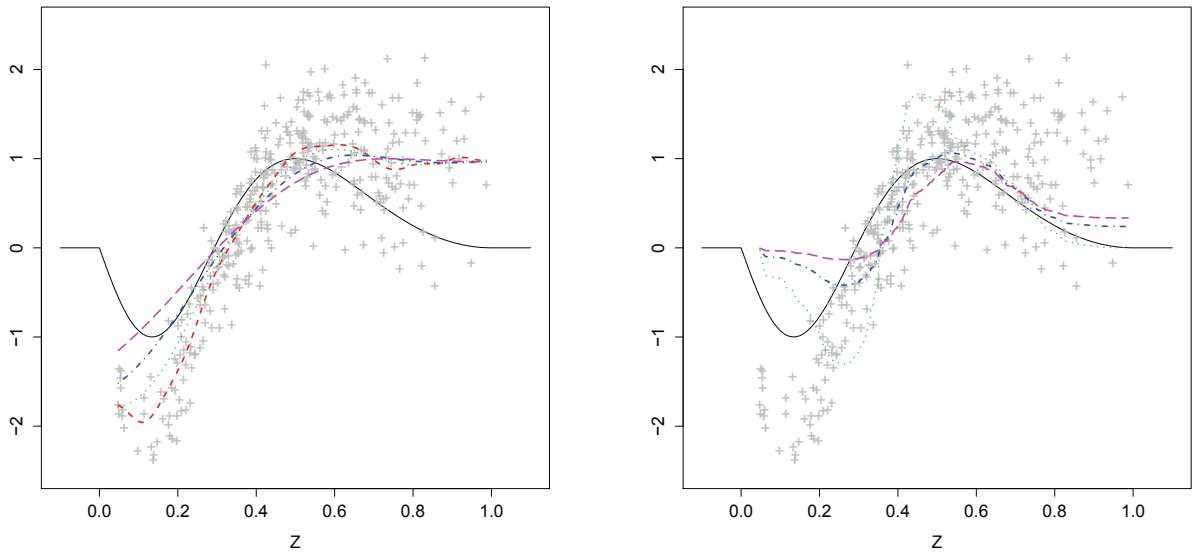
$$\begin{aligned} \tilde{X}_1 &= \tfrac{1}{2} - \tfrac{\mu_Z}{2\times 1.96\times \sigma_Z} + \tfrac{X_1}{2\times 1.96\times \sigma_Z} \\ \tilde{X}_i &= \tfrac{X_i}{2\times 1.96\times \sigma_Z} \qquad i = 2, \ldots, 5, \end{aligned}$$

which is such that most of the realizations of $\tilde{Z} = \tilde{X}^\top \vartheta$ belong to $[0, 1]$. In practice, we discard the values that are not in $[0, 1]$. The dependent variable Y is generated from the single-index model $Y = h(\tilde{X}^\top \vartheta) + U$, see (1). The component U that is the error term in the single-index model also appears in the generation of X and $\tilde{X}$, introducing non-zero correlation between the covariates and the error term.

Figure 1(a) illustrates this setting with a sample of $n = 400$ data points. In the figure the variable $\tilde{Z}$ is on the x-axis, and Y is on the y-axis. The solid line represents the true function $h(z)$. The endogeneity of $\tilde{Z}$ is apparent from this figure, since the cloud of data points is not equally located around the function h. Figure 1(a) also shows the result of the common kernel estimator of h from the observations of $(\tilde{Z}, Y)$. The Nadaraya-Watson estimator is used with a Gaussian kernel. The results for various bandwidth choices are superimposed in the figure. This estimator ignores endogeneity and is therefore biased.

In contrast, the same data are used in Figure 1(b), now with the nonparametric estimator studied in the previous sections that is minimizing the penalized discrepancy function (2). The figure superimposes our estimator for various choices of the regularization parameter α.

Figure 1: Nonparametric estimation of the function h from an i.i.d. sample of Y and $\tilde{Z}$ (with observed $\tilde{Z}$). In the figure at the left a Nadaraya-Watson estimator is used with bandwidth values $0.1, 0.2, 0.3$ and 0.4, whereas the figure at the right is based on a penalized estimator with regularization parameter $\alpha = 10^{-9}, 10^{-10}$ and 10^{-11}.



(a) Nadaraya-Watson estimator

(b) Nonparametric instrumental regression estimator

The estimation also includes the choice of the bandwidths b_Z and b_W (cf Section 2). Little is known about the optimal choice of those nuisance parameters in such a complex model. In the simulations we therefore consider Silverman (1998)'s rule of thumb and set b_W (resp. b_Z) equal to $1.06\widehat{\sigma}_W n^{-1/5}$ (resp. $1.06\widehat{\sigma}_Z n^{-1/5}$). We also found empirically that it might be better to undersmooth the density of Z. Accordingly we consider three data driven choices for b_Z in our Monte Carlo study below, where we also study the sensitivity of the estimator for a range of values of α. Note also that, by construction, the estimator in Figure 1(b) satisfies the constraints imposed by the penalty term (2) so that the estimator is twice differentiable and is such that $\widehat{h}(0) = 0$ and $\widehat{h}'(1) = 0$.

We now turn to the estimation of the vector of parameters ϑ. We report below the results for ϑ_2. To construct the functions $\widehat{h}_{\theta,\alpha}$ for each θ several choices are considered for the regularization parameter α (going from 10^{-9} to 10^{-5}) and for the bandwidth b_Z. The three choices for b_Z are: Silverman's rule of thumb (denoted $b_Z(1)$), $b_Z(2) = b_Z(1)/2$ and

Table 1: Each cell presents the bias and standard error (in parentheses) of the estimator $\widehat{\theta}$ from 500 simulations. Various choices for α are tested and three data driven choices of b_Z are considered: $b_Z(1)$ is Silverman's rule of thumb, $b_Z(2) = b_Z(1)/2$ and $b_Z(3) = b_Z(1)/4$. Two sample sizes are considered : $n = 200$ and 400. The function g equals $g(W) = (W, W^2, |W|, sgn(W)\sqrt{|W|}, \log(|W|))^\top$.

	$n = 200$			$n = 400$		
	$b_Z(1)$	$b_Z(2)$	$b_Z(3)$	$b_Z(1)$	$b_Z(2)$	$b_Z(3)$
$\alpha = 10^{-9}$	-0.1015 *(.34)*	-0.1213 *(.30)*	-0.0984 *(.31)*	-0.0997 *(.32)*	-0.1196 *(.28)*	-0.0987 *(.29)*
$\alpha = 10^{-10}$	-0.1008 *(.31)*	-0.1213 *(.35)*	-0.1105 *(.34)*	-0.1535 *(.33)*	-0.1435 *(.32)*	-0.1084 *(.32)*
$\alpha = 10^{-11}$	-0.1045 *(.39)*	-0.1148 *(.32)*	-0.1083 *(.33)*	-0.1044 *(.29)*	-0.1098 *(.28)*	-0.0973 *(.32)*

$b_Z(3) = b_Z(1)/4$. Estimating ϑ as in (3) also requires to choose a multivariate function g of the instruments. In our simulations, we study several options. In a first set of simulations we consider $g(W) = (W, W^2, |W|, sgn(W)\sqrt{|W|}, \log(|W|))^t$, which is simple to implement but not optimal. In theory, the optimal choice for this function is $g_\theta(W) = \mathbb{E}(h'(X^t\theta)X|W)$ (see Remark 3 together with the fact that U and W are independent, which implies that $\mathrm{Var}(U|W)$ is constant). In a second set of simulations, we consider that last function at the true value of the parameter ϑ. That estimator is theoretically optimal but unfeasible. The conditional expectation appearing in the optimal function g is computed in practice by the Nadaraya-Watson estimator with a Gaussian kernel and a bandwidth provided by Silverman's rule of thumb.

The minimization of (3) is performed by a discretization of the parameter space Θ. For each configuration the results for 500 Monte Carlo simulations are given in Tables 1 and 2 for various sample sizes.

Monte Carlo simulations show relatively stable results over the considered range of α. The tables also show that the bias is generally smaller when the bandwidth b_Z is smaller than Silverman's rule of thumb. Undersmoothing the density of Z is therefore a recommendation

Table 2: Each cell presents the bias and standard error (in parentheses) of the estimator $\widehat{\theta}$ from 500 simulations. Various choices for α are tested and three data driven choices of b_Z are considered: $b_Z(1)$ is Silverman's rule of thumb, $b_Z(2) = b_Z(1)/2$ and $b_Z(3) = b_Z(1)/4$. Two sample sizes are considered : $n = 200$ and 400. The function g equals $g_\theta(W) = \mathbb{E}(h'(X^t\theta)X|W)$.

	$n = 200$			$n = 400$		
	$b_Z(1)$	$b_Z(2)$	$b_Z(3)$	$b_Z(1)$	$b_Z(2)$	$b_Z(3)$
$\alpha = 10^{-9}$	-0.0933 *(.29)*	-0.1155 *(.33)*	-0.0744 *(.32)*	-0.0814 *(.30)*	-0.1056 *(.25)*	-0.0972 *(.28)*
$\alpha = 10^{-10}$	-0.0942 *(.31)*	-0.0931 *(.29)*	-0.1238 *(.29)*	-0.1354 *(.30)*	-0.1343 *(.31)*	-0.0899 *(.33)*
$\alpha = 10^{-10}$	-0.119 *(.30)*	-0.097 *(.32)*	-0.111 *(.29)*	-0.091 *(.27)*	-0.113 *(.31)*	-0.080 *(.30)*

for practical implementation. The same exercise with the density of the instrumental variable W (not reported here) showed that the procedure is less sensitive to changes of b_W.

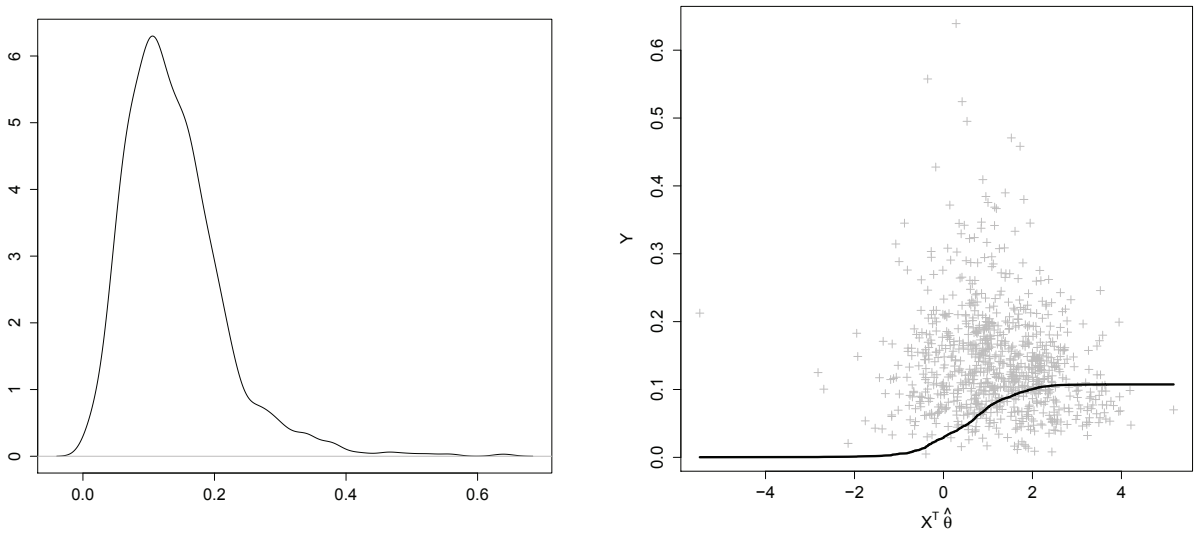
5 Application to the estimation of Engle curves

Following the work of the German statistician Ernst Engle (1821-1896), economists refer to *Engle curves* when they study the relationship how the household expenditure on a particular good or service varies according to the income or expense structure, see e.g. Lewbel (2008). In this section we use data coming from the Family Expenditure Survey (FES) of the UK government. Data have been studied in previous reports, and our application uses the cross-section that has been considered in Blundell et al. (2007), see also Kim et al. (2013).

We consider the sample of 861 families with one child. The dependent variable Y is the share of household expenses in food excluding catering and alcoholic drinks. The variable Y is a ratio between 0 and 1. A kernel density estimation of Y is displayed in Figure 2(a). The single index model considered in our study allows to analyse the share as function of various expenses. We consider the explanatory variables (X_1, X_2, X_3) given by the log of expenses

in household goods, household services and other expenses, respectively. Following the empirical work of Blundell et al. (2007), we consider the total income as the instrument W. We also consider a three-dimensional function $g(W)$ given by $g(W) = (W, \log(W), \Phi(\log(W)))^\top$.

Figure 2: Nonparametric estimation of the density of household share in food (Y) and of the link function $h_{\theta,\alpha}$.



(a) Kernel density estimation of Y

(b) Nonparametric instrumental regression of the function h

In the process of estimating the single index vector, we selected the bandwidths according to the conclusion of the above Monte Carlo simulations, that is the Silverman's rule of thumb for b_W and half the rule of thumb for b_Z. A grid of values of θ is used to minimize the discrepancy function. Several values of the regularization parameters have been used, and we report the result for $\hat{\alpha} = 10^{-4}$. The estimated vector of parameters is $\hat{\theta} = (1, -0.2, -0.7)$ and the estimated nonparametric link function $\hat{h}_{\hat{\theta},\hat{\alpha}}$ is shown in Figure 2(b). The monotonicity of the function $h_{\theta,\alpha}$ is remarkable. Observing that the values of $\hat{\theta}_2$ and $\hat{\theta}_3$ are negative, it suggests that, all other things being equal, the share of expenditure in food decreases with an increase of the expenses in services or other expenses than services and goods.

6 Conclusions

In this paper we have studied the estimation of a semiparametric single-index model when endogeneity is present in the explanatory variables, and a vector of instruments is available that is non-correlated with the error term. Under this model, an estimator of the parametric component of the model is proposed, which is the solution of an ill-posed inverse problem. The $\sqrt{n}$-consistency and asymptotic normality of the proposed parameter estimator $\widehat{\theta}$ is established using delicate results on the asymptotic theory for general semiparametric estimators. As a by-product we also obtain the asymptotic properties of the estimator $\widehat{h}$ of the link function, which is smooth and twice differentiable. Therefore meaningful quantities such as the marginal effect of a covariate, which involves the derivative of h, can be easily estimated. The finite sample performance of the parameter estimator is also studied via a simulation study. The simulations show the benefits of undersmoothing the density of $X^t\theta$, which is an interesting aspect to notice as well.

Although some indications are given in the simulation study about how to choose the smoothing parameters b_W and b_Z and the regularization parameter α in practice, the optimal choice of these parameters remains an open issue, which is worth to be studied in the future. Another open problem is the selection of the function g in the estimating equation. It is expected that the function g has an impact on the variance and the efficiency of the parameter estimator. This important issue of the method merits further attention, but is beyond the scope of this paper, since it necessitates an elaborated, lengthy and detailed efficiency study of the proposed method.

References

Ai, C. & Chen, X. (2003), 'Efficient estimation of models with conditional moment restrictions containing unknown functions', *Econometrica* **71**, 1795–1843.

Angrist, J. D. & Krueger, A. B. (2001), 'Instrumental variables and the search for identification: From supply and demand to natural experiments', *Journal of Economic Perspectives* **15**, 69–85.

Bissantz, N., Dette, H. & Hildebrandt, T. (2013), 'Smooth backfitting in additive inverse regression', *Preprint* .

Blundell, R., Chen, X. & Kristensen, D. (2007), 'Semi-nonparametric iv estimation of shape-invariant engel curves', *Econometrica* **75**, 1613–1669.

Carroll, R., Fan, J., Gijbels, I. & Wand, M. (1997), 'Generalized partially linear single-index models', *Journal of the American Statistical Association* **92**, 477–489.

Cavalier, L. (2008), 'Nonparametric statistical inverse problems', *Inverse Problems* **24**, 1–19.

Cavalier, L. & Golubev, G. (2006), 'Risk hull method and regularization by projections of ill-posed inverse problems', *Annals of Statistics* **34**, 1653–1677.

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

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

Darolles, S., Fan, Y., Florens, J.-P. & Renault, E. (2011), 'Nonparametric instrumental regression', *Econometrica* **79**, 1541–1565.

Delecroix, M., Hristache, M. & Patilea, V. (2006), 'On semiparametric m-estimation in single-index regression', *Journal of Statistical Planning and Inference* **136**, 730–769.

Florens, J.-P., Johannes, J. & Van Bellegem, S. (2011), 'Identification and estimation by penalization in nonparametric instrumental regression', *Econometric Theory* **27**, 472–496.

Florens, J.-P., Johannes, J. & Van Bellegem, S. (2012), 'Instrumental regression in partially linear models', *The Econometrics Journal* **15**, 304–324.

Florens, J.-P., Marimoutou, V. & Peguin, A. (2004), *Econométrie - Modélisation et Inférence*, Armand Colin.

Hall, P. & Horowitz, J. L. (2005), 'Nonparametric methods for inference in the presence of instrumental variables', *Annals of Statistics* **33**, 2904–2929.

Härdle, W., Hall, P. & Ichimura, H. (1993), 'Optimal smoothing in single-index models', *Annals of Statistics* **21**, 157–178.

Hayashi, F. (2000), *Econometrics*, Princeton University Press.

Hildebrandt, T., Bissantz, N. & Dette, H. (2014), 'Additive inverse regression models with convolution-type operators', *Electronic Journal of Statistics* **8**, 1–40.

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

Hristache, M., Juditsky, A. & Spokoiny, V. (2001), 'Direct estimation of the index coefficient in a single-index model', *Annals of Statistics* **29**, 595–623.

Hu, Y. & Shiu, J.-L. (2011), 'Nonparametric identification using instrumental variables: Sufficient conditions for completeness', *Working papers – the Johns Hopkins University, Department of Economics, No. 581* .

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

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

Johannes, J. (2009), 'Deconvolution with unknown error distribution', *Annals of Statistics* **37**, 2301–2323.

Johannes, J., Van Bellegem, S. & Vanhems, A. (2011), 'Convergence rates for ill-posed inverse problems with an unknown operator', *Econometric Theory* **27**, 522–545.

Johannes, J., Van Bellegem, S. & Vanhems, A. (2013), 'Iterative regularization in nonparametric instrumental regression', *Journal of Statistical Planning and Inference* **143**, 24–39.

Kim, N. H., Saart, P. & Gao, J. (2013), 'Semi-parametric analysis of shape-invariant Engel curves with control function approach', *Available at SSRN: http://dx.doi.org/10.2139/ssrn.2213067* .

Klein, R. & Spady, R. (1993), 'An efficient semiparametric estimator for binary response models', *Econometrica* **61**, 387–421.

Kong, E. & Xia, Y. (2007), 'Variable selection for the single-index model', *Biometrika* **94**, 217–229.

Lewbel, A. (2008), Engel curves, *in* 'The New Palgrave Dictionary of Economics', 2nd edn, Macmillan.

Liang, H., Liu, X., Li, R. & Tsai, C. (2010), 'Estimation and testing for partially linear single-index models', *Annals of Statistics* **38**, 3811–3836.

Lin, W. & Kulasekera, K. B. (2007), 'Identifiability of single-index models and additive-index models', *Biometrika* **94**, 496–501.

Ma, Y. & Zhu, L. (2013), 'Efficient estimation in sufficient dimension reduction', *Annals of Statistics* **41**, 250–268.

Manzi, J., San Martín, E. & Van Bellegem, S. (2014), 'School system evaluation by value-added analysis under endogeneity', *Psychometrika* **79**, 130–153.

Newey, W. & Powell, J. (2003), 'Instrumental variable estimation of nonparametric models', *Econometrica* **71**, 1565–1578.

Peng, H. & Huang, T. (2011), 'Penalized least squares for single index models', *Journal of Statistical Planning and Inference* **141**, 1362–1379.

Powell, J., Stock, J. & Stoker, T. (1989), 'Semiparametric estimation of index coefficients', *Econometrica* **57**, 1403–1430.

Silverman, B. (1998), *Density Estimation for Statistics and Data Analysis*, Chapman & Hall/CRC.

Van der Vaart, A. & Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, Springer.

Vanhems, A. & Van Keilegom, I. (2013), 'Semiparametric transformation model with endogeneity: a control function approach'. Submitted.

Wang, J.-L., Xue, L., Zhu, L. & Chong, Y. S. (2010), 'Estimation for a partial-linear single-index model', *Annals of Statistics* **38**, 246–274.

Wooldridge, J. (2008), *Introductory Econometrics: A Modern Approach*, 4th edn, South-Western College Pub.

Xia, Y., Härdle, W. K. & Linton, O. (2012), Optimal smoothing for a computationally and statistically efficient single index estimator, *in* I. Van Keilegom & P. W. Wilson, eds, 'Exploring Research Frontiers in Contemporary Statistics and Econometrics', Physica-Verlag HD, pp. 229–261.

Yin, X. & Cook, R. (2002), 'Dimension reduction for the conditional k-th moment in regression', *Journal of the Royal Statistical Society - Series B* **64**, 159–175.

Zhang, R., Huang, Z. & Lv, Y. (2010), 'Statistical inference for the index parameter in single-index models', *Journal of Multivariate Analysis* **101**, 1026–1041.

Corresponding author:
Ingrid Van Keilegom
Institute of Statistics, Biostatistics and Actuarial Sciences
Université catholique de Louvain
Voie du Roman Pays 20
1348 Louvain-la-Neuve
Belgium
ingrid.vankeilegom@uclouvain.be

Appendix: Proofs

We start with a definition and a number of technical lemmas, needed in the proof of the main result.

Lemma 1. 1. If f and g are two probability densities that belong to $\mathfrak{G}^{s,\alpha}(\mathbb{R})$ with $\alpha = 1$ or 2, then the convolution $f \star g \in \mathfrak{G}^{s,\alpha}(\mathbb{R})$.

2. If f is a probability density that belongs to $\mathfrak{G}^{s,\alpha}(\mathbb{R})$ with $\alpha = 1$ or 2, and if $\beta \neq 0$ then $(1/\beta)f(\cdot/\beta) \in \mathfrak{G}^{s,\alpha}(\mathbb{R})$.

Proof. We prove the first assertion and restrict attention to the case where $s > 1$. First

note that for all x, y,

$$Q_{f\star g,x}(y-x) = \sum_{j=1}^{m-1} \frac{\partial^j (f\star g)(x)}{\partial x^j}(y-x)^j$$
$$= \sum_{j=1}^{m-1} \frac{\partial^j}{\partial x^j} \int f(x-z)g(z)dz\,(y-x)^j = \int Q_{f,x-z}(y-x)g(z)dz.$$

Hence, for all $|y-x| \le \rho_f$,

$$\frac{|(f\star g)(y) - (f\star g)(x) - Q_{f\star g,x}(y-x)|}{|y-x|^s}$$
$$= \frac{\int \big|f(y-z) - f(x-z) - Q_{f,x-z}(y-x)\big| g(z)dz}{|(y-z)-(x-z)|^s}$$
$$\le \int \psi_f(x-z)g(z)dz = (\psi_f \star g)(x).$$

Moreover if $\alpha = 1$, $\int (f\star g)(x)dx = \int \big[\int f(x-z)dx\big]g(z)dz = \int f(y)dy \cdot \int g(z)dz = 1$ and similarly $\int (\psi_f \star g)(x)dx < \infty$. For $\alpha = 2$ we have:

$$\int (f\star g)^2(x)dx = \int \Big[\int f(x-z)g(z)dz\Big]^2 dx$$
$$\le \int \Big[\int f^2(x-z)g(z)dz \int g(z)dz\Big]dx$$
$$= \int \Big[\int f^2(x-z)dx\Big]g(z)dz = \int f^2(y)dy < \infty,$$

since $\int g(z)dz = 1$. In a similar way we can show that $\int (\psi_f \star g)^2(x)dx < \infty$. □

The previous lemma has the following consequence: if each variable X_j, $j = 1, \ldots, k$, has a density in $\mathfrak{G}^{s,\alpha}(\mathbb{R})$ ($\alpha = 1$ or 2), then for any $\theta \in \Theta$, $f_{X^t\theta} \in \mathfrak{G}^{s,\alpha}(\mathbb{R})$. This property will be used in the proofs below.

The next lemma gives a closed form expression for the functions h°_θ, $h^\circ_{\theta,\alpha}$ and $\widehat{h}_{\theta,\alpha}$.

Lemma 2. The functions h°_θ, $h^\circ_{\theta,\alpha}$ and $\widehat{h}_{\theta,\alpha}$ satisfy:

$$h^\circ_\theta = L^{-1}(S^\star_\theta S_\theta)^{-1} S^\star_\theta r, \qquad\qquad h^\circ_{\theta,\alpha} = L^{-1}(\alpha I + S^\star_\theta S_\theta)^{-1} S^\star_\theta r,$$

and

$$\widehat{h}_{\theta,\alpha} = L^{-1}(\alpha I + \widehat{S}^\star_\theta \widehat{S}_\theta)^{-1} \widehat{S}^\star_\theta \widehat{r},$$

where I is the identity operator.

Proof. We prove the last statement. The first and the second one can be derived in a similar way. First, note that the estimator $\widehat{h}_{\theta,\alpha}$ minimizes the functional

$$\langle \widehat{T}_\theta h - \widehat{r}, \widehat{T}_\theta h - \widehat{r} \rangle + \alpha \langle Lh, Lh \rangle$$

over all $h \in \mathcal{H}$. The minimizer of this functional is the element in $\mathcal{H}$ for which the Fréchet derivative of this functional in all possible directions $\widetilde{h}$ equals zero. Consider

$$\begin{aligned}
&\lim_{\varrho \to 0} \frac{1}{\varrho} \Big\{ \langle \widehat{T}_\theta h + \varrho \widehat{T}_\theta \widetilde{h} - \widehat{r}, \widehat{T}_\theta h + \varrho \widehat{T}_\theta \widetilde{h} - \widehat{r} \rangle + \alpha \langle Lh + \varrho L\widetilde{h}, Lh + \varrho L\widetilde{h} \rangle \\
&\qquad - \langle \widehat{T}_\theta h - \widehat{r}, \widehat{T}_\theta h - \widehat{r} \rangle - \langle Lh, Lh \rangle \Big\} \\
&= 2 \Big\{ \langle \widehat{T}_\theta \widetilde{h}, \widehat{T}_\theta h - \widehat{r} \rangle + \alpha \langle Lh, L\widetilde{h} \rangle \Big\} \\
&= 2 \Big\{ \langle \widetilde{h}, \widehat{T}_\theta^\star \widehat{T}_\theta h - \widehat{T}_\theta^\star \widehat{r} \rangle + \alpha \langle \widetilde{h}, L^\star Lh \rangle \Big\}
\end{aligned}$$

for all $\widetilde{h}$, and therefore

$$\widehat{h}_{\theta,\alpha} = (\alpha L^\star L + \widehat{T}_\theta^\star \widehat{T}_\theta)^{-1} \widehat{T}_\theta^\star \widehat{r} = L^{-1} (\alpha I + \widehat{S}_\theta^\star \widehat{S}_\theta)^{-1} \widehat{S}_\theta^\star \widehat{r},$$

where the last equality follows from the definition of $\widehat{S}_\theta$. □

The next lemma gathers useful results on the norm of (a function of) operators. It is quoted from Florens et al. (2011), see their Lemma A.1, p. 489, where a formal proof can be found.

Lemma 3 (Florens et al. (2011))**.** Let $K : \mathbb{H} \to \mathbb{G}$ be a linear operator defined between the two Hilbert spaces $\mathbb{H}$ and $\mathbb{G}$, and let $K^\star$ be the adjoint operator of K. Then, for all $\alpha > 0$, the following bounds on the operator norm hold true:

$$\|\alpha(\alpha I + K^\star K)^{-1} (K^\star K)^\gamma\| \leqslant \begin{cases} \alpha^\gamma & \text{if } 0 < \gamma \leqslant 1 \\ \|K^\star K\|^{\gamma - 1} \alpha & \text{if } \gamma > 1 \end{cases},$$

$$\|(\alpha I + K^\star K)^{-1} K^\star\| = \|K(\alpha I + K^\star K)^{-1}\| \lesssim 1/\sqrt{\alpha}\,,$$

$$\|(\alpha I + K^\star K)^{-1}\| \leqslant 1/\alpha\,,$$

$$\|K(\alpha I + K^\star K)^{-1} K^\star\| \leqslant 1\,,$$

$$\|K[I - (\alpha I + K^\star K)^{-1} K^\star K]\| \lesssim \sqrt{\alpha}\,,$$

$$\|I - (\alpha I + K^\star K)^{-1} K^\star K\| \leqslant 1\,.$$

The distance between the operators $\widehat{S}_\theta$ and S_θ and the corresponding distance between the adjoint operators is of crucial importance for the proof of the main result. We give its rate of convergence in the next lemma.

Lemma 4. Assume (A.3) and (A.4). Then,

$$\sup_{\theta\in\Theta}\|\widehat{S}_\theta - S_\theta\|^2 = \mathcal{O}_P((nb_W^q b_Z)^{-1} + (b_W \vee b_Z)^{2\rho})$$

and

$$\sup_{\theta\in\Theta}\|\widehat{S}_\theta^\star - S_\theta^\star\|^2 = \mathcal{O}_P((nb_W^q b_Z)^{-1} + (b_W \vee b_Z)^{2\rho}),$$

where ρ is defined in assumption (A.5). The same holds true when the operator S is replaced by T.

Proof. The proof follows from Lemma A.1 in Florens et al. (2012) combined with Markov's inequality (except that our result is uniform over θ), and is therefore omitted. □

The next proposition considers the rate of convergence of $\widehat{h}_{\theta,\alpha} - h_\theta^\circ$ with respect to the $||\cdot||_{\mathcal{H}}$-norm uniformly over θ.

Proposition 1. Assume (A.2), (A.3), (A.4) and (A.8). Then,

$$\sup_{\theta\in\Theta}\|\widehat{h}_{\theta,\alpha} - h_\theta^\circ\|_{\mathcal{H}}^2 = \mathcal{O}_P\Big(\frac{(b_W \vee b_Z)^{2\rho}}{\alpha^2} + \frac{1}{\alpha^2 n b_W^q b_Z} + \alpha^{\gamma\wedge 2}\Big),$$

where γ and ρ are defined in assumption (A.5).

Proof. First, consider

$$\widehat{h}_{\theta,\alpha} - h_{\theta,\alpha}^\circ = \mathrm{I} + \mathrm{II} + \mathrm{III},$$

where

$$\begin{aligned}
\mathrm{I} &= L^{-1}(\alpha I + \widehat{S}_\theta^\star\widehat{S}_\theta)^{-1}\widehat{S}_\theta^\star(\widehat{r} - r),\\
\mathrm{II} &= L^{-1}(\alpha I + \widehat{S}_\theta^\star\widehat{S}_\theta)^{-1}(\widehat{S}_\theta^\star - S_\theta^\star)r,\\
\mathrm{III} &= L^{-1}(\alpha I + \widehat{S}_\theta^\star\widehat{S}_\theta)^{-1}(S_\theta^\star S_\theta - \widehat{S}_\theta^\star\widehat{S}_\theta)(\alpha I + S_\theta^\star S_\theta)^{-1}S_\theta^\star r.
\end{aligned}$$

By Lemma 3 we have that

$$\|\mathrm{I}\|_{\mathcal{H}}^2 \le \frac{K}{\alpha}\|\widehat{r} - r\|_{\mathcal{H}}^2 = \mathcal{O}_P\Big(\frac{b_W^{2\rho}}{\alpha} + \frac{1}{\alpha n b_W^q}\Big),$$

under assumptions (A.3) and (A.4). Next, consider II:

$$\|\mathrm{II}\|_{\mathcal{H}}^2 \leq \frac{K}{\alpha^2}\|\widehat{S}_\theta^\star - S_\theta^\star\|^2 = \mathcal{O}_P\Big(\frac{1}{\alpha^2 n b_W^q b_Z} + \frac{(b_W \vee b_Z)^{2\rho}}{\alpha^2}\Big),$$

under assumption (A.3), by Lemmas 3 and 4 above. It remains to consider III. Note that

$$S_\theta^* S_\theta - \widehat{S}_\theta^* \widehat{S}_\theta = (S_\theta^* - \widehat{S}_\theta^*)S_\theta - \widehat{S}_\theta^*(\widehat{S}_\theta - S_\theta),$$

and hence,

$$\|\mathrm{III}\|_{\mathcal{H}}^2 \leq \frac{K}{\alpha^2}\|S_\theta^\star S_\theta - \widehat{S}_\theta^\star \widehat{S}_\theta\|^2 \|r\|_{\mathcal{H}}^2 = \mathcal{O}_P\Big(\frac{1}{\alpha^2 n b_W^q b_Z} + \frac{(b_W \vee b_Z)^{2\rho}}{\alpha^2}\Big),$$

again by Lemmas 3 and 4. It remains to consider the bias part $\|h_{\theta,\alpha}^\circ - h_\theta^\circ\|_{\mathcal{H}}^2$, which is of order $\alpha^{\gamma\wedge 2}$ under assumption (A.2), by using standard arguments (e.g. Florens et al. (2012)).□

In the next lemma we consider the rate of convergence of the L_2-norm of $[\widehat{h}_{\vartheta,\alpha} - h_{\vartheta,\alpha}^\circ](X^t\vartheta)$ and of its first derivative as well as uniform bounds with respect to z and θ for the estimator and its derivatives.

Lemma 5. Assume (A.3) and (A.4). Then,

$$\mathbb{E}_X\Big(\widehat{h}_{\vartheta,\alpha}(X^t\vartheta) - h_{\vartheta,\alpha}^\circ(X^t\vartheta)\Big)^2 = o_P(1) \quad \text{and} \quad \mathbb{E}_X\Big(\widehat{h}'_{\vartheta,\alpha}(X^t\vartheta) - h_{\vartheta,\alpha}^{\circ\prime}(X^t\vartheta)\Big)^2 = o_P(1),$$

where $\mathbb{E}_X$ denotes the expectation with respect to the variable X only.

Proof. Let $\widehat{d}_{\vartheta,\alpha} = \widehat{h}_{\vartheta,\alpha} - h_{\vartheta,\alpha}^\circ$. The first expectation equals

$$\mathbb{E}_X\Big(\widehat{d}_{\vartheta,\alpha}^2(X^t\vartheta)\Big) = \int \widehat{d}_{\vartheta,\alpha}^2(z) f_{X^t\vartheta}(z)dz,$$

and since $f_{X^t\vartheta}$ is uniformly bounded this is bounded up to a constant by $\int \widehat{d}_{\vartheta,\alpha}^2(z)dz$. The first result now follows by using similar arguments as in the proof of Proposition 1.

For the second statement the proof is similar. Indeed, we can write

$$\widehat{h}'_{\vartheta,\alpha} - h_{\vartheta,\alpha}^{\circ\prime} = L^{-1/2}(\alpha I + \widehat{S}_\vartheta^\star \widehat{S}_\vartheta)^{-1}\widehat{S}_\vartheta^\star \widehat{r} - L^{-1/2}(\alpha I + S_\vartheta^\star S_\vartheta)^{-1} S_\vartheta^\star r.$$

The control of this difference is similar to what we did in Proposition 1. We omit the details. □

Lemma 6. Assume (A.2), (A.4), (A.5) and (A.8). Then, there exists a constant $M > 0$ such that

$$\mathrm{P}\Big(\sup_{z\in\Omega,\theta\in\Theta} |\widehat{h}^{(s)}_{\theta,\alpha}(z)| < M \Big) \to 1 \text{ for } n \to \infty,\ s = 0, 1.$$

Furthermore,

$$\sup_{z\in\Omega,\theta\in\Theta} |\widehat{h}''_{\theta,\alpha}(z) - h^{\circ\prime\prime}_{\theta,\alpha}(z)| = O_P(1).$$

Proof. Write

$$\begin{aligned} |L^{s/2}\widehat{h}_{\theta,\alpha}| &\leq |(\alpha I + L^{s/2}\widehat{T}^{\star}_{\theta}\widehat{T}_{\theta}L^{s/2})^{-1}L^{s/2}\widehat{T}^{\star}_{\theta}r| + |(\alpha I + L^{s/2}\widehat{T}^{\star}_{\theta}\widehat{T}_{\theta}L^{s/2})^{-1}L^{s/2}\widehat{T}^{\star}_{\theta}(\widehat{r} - r)| \\ &= R_1 + R_2. \end{aligned}$$

For the first term we have

$$\begin{aligned} R_1 &\leq \sum_k \left| \frac{\widehat{\lambda}_{\theta,s,k}}{\alpha + \widehat{\lambda}^2_{\theta,s,k}} \right| |\langle T_\vartheta h_0, \widehat{\phi}_{\theta,s,k}\rangle|\, |\widehat{\psi}_{\theta,s,k}| \\ &= \sum_k \left| \frac{\widehat{\lambda}_{\theta,s,k}}{\alpha + \widehat{\lambda}^2_{\theta,s,k}} \right| |\langle h_0, T^{\star}_\vartheta \widehat{\phi}_{\theta,s,k}\rangle|\, |\widehat{\psi}_{\theta,s,k}| \\ &\leq \sum_{k,l} \left| \frac{\widehat{\lambda}_{\theta,s,k}\lambda_{\vartheta,s,l}}{\alpha + \widehat{\lambda}^2_{\theta,s,k}} \right| |\langle \widehat{\phi}_{\theta,s,k}, \phi_{\vartheta,s,l}\rangle| |\langle L^{-s/2}h_0, \psi_{\vartheta,s,l}\rangle|\, |\widehat{\psi}_{\theta,s,k}| \\ &\leq C_{s,3} \sum_{k,l} \left| \frac{\widehat{\lambda}^2_{\theta,s,k}}{\alpha + \widehat{\lambda}^2_{\theta,s,k}} \right| \widehat{\ell}_s(|k-l|) |\langle L^{-s/2}h_0, \psi_{\vartheta,s,l}\rangle|\, |\widehat{\psi}_{\theta,s,k}| \\ &\leq C_{s,3} \sum_k ||L^{-s/2}h^{\circ}_\vartheta||\, |\widehat{\psi}_{\theta,s,k}| \sum_l \widehat{\ell}_s(|k-l|) \\ &\leq C_{s,3} \sum_k ||L^{s/2}h^{\circ}_\vartheta||\, |\widehat{\psi}_{\theta,s,k}| \sum_l \widehat{\ell}_s(|k-l|) \end{aligned}$$

by using the basis expansion of $\widehat{\phi}_{\theta,s,k}$ with respect to $\{\phi_{\vartheta,s,l}\}_{l\geq 0}$ in the third line, the fact that $z^2/(\alpha + z^2) \leq 1$ for all $\alpha > 0$ and assumption (A.8) (iv) in the fourth line, and using that $||L^{-s/2}||$ is bounded. R_1 is therefore bounded, using assumption (A.8) (ii). For the second term we see that

$$\begin{aligned} R_2 &\leq \sum_k \left| \frac{\widehat{\lambda}_{\theta,s,k}}{\alpha + \widehat{\lambda}^2_{\theta,s,k}} \right| |\langle \widehat{r} - r, \widehat{\phi}_{\theta,s,k}\rangle|\, |\widehat{\psi}_{\theta,s,k}(z)| \\ &\leq \frac{1}{\alpha}||\widehat{r} - r||_2 \sum_k |\widehat{\psi}_{\theta,s,k}(z)| = \Big[O\Big(\frac{b^{\rho}_W}{\alpha}\Big) + O_{a.s.}\Big(\frac{1}{\alpha\sqrt{nb^q_W}}\Big)\Big] \sum_k |\widehat{\psi}_{\theta,s,k}(z)| \end{aligned}$$

Taking the supremum over all $\theta \in \Theta$ and $z \in \Omega$ and using the fact that $O_{a.s.}((nb_W^q\alpha^2)^{-1}) = o_{a.s.}(b_Z) = o_{a.s.}(1)$, gives the first result. For the second statement, we first write $|\widehat{d}''_{\vartheta,\alpha}(z)| \leqslant |\widehat{h}''_{\vartheta,\alpha}(z)| + |h^{\circ\prime\prime}_{\vartheta,\alpha}(z)|$, and then follow the lines above by using the weaker conditions for $s = 2$. □

Lemma 7. Assume (A.3) and (A.4). Then, for $\delta_n = o(1)$,

$$\sup_{\|\theta-\vartheta\|\leq\delta_n} \mathbb{E}\Big[\Big\|\Big(\frac{\partial}{\partial\gamma}[\widehat{h}_{\gamma,\alpha} - h^{\circ}_{\gamma,\alpha}]\Big)\Big|_{\gamma=\vartheta}(X^t\theta)\Big\|^2\Big] = o_P(1),$$

and

$$\sup_{\|\theta-\vartheta\|\leq\delta_n}\sup_{\|\eta-\vartheta\|\leq\delta_n} \mathbb{E}\Big[\Big\|\Big(\frac{\partial^2}{\partial\gamma\partial\gamma^t}[\widehat{h}_{\gamma,\alpha} - h^{\circ}_{\gamma,\alpha}]\Big)\Big|_{\gamma=\eta}(X^t\theta)\Big\|^2\Big] = O_P(1).$$

Proof. Let $\widehat{d}_{\theta,\alpha} = \widehat{h}_{\theta,\alpha} - h^{\circ}_{\theta,\alpha}$ as in the previous proof. Using a Taylor expansion and Lemma 2 we can write

$$\frac{\partial}{\partial\gamma}\widehat{d}_{\gamma,\alpha}\Big|_{\gamma=\vartheta} = \widehat{Q}_{\vartheta,\alpha}\widehat{r} - Q_{\vartheta,\alpha}r = (\widehat{Q}_{\vartheta,\alpha} - Q_{\vartheta,\alpha})r + \widehat{Q}_{\vartheta,\alpha}(\widehat{r} - r),$$

where $D = L^\star L$,

$$\widehat{Q}_{\theta,\alpha} = \Big(\alpha D + \widehat{T}^\star_\theta\widehat{T}_\theta\Big)^{-1}\Big\{\widehat{T}^{\star(1)}_\theta - \Big[\widehat{T}^\star_\theta\widehat{T}^{(1)}_\theta + \widehat{T}^{\star(1)}_\theta\widehat{T}_\theta\Big]\Big(\alpha D + \widehat{T}^\star_\theta\widehat{T}_\theta\Big)^{-1}\widehat{T}^\star_\theta\Big\}$$
$$Q_{\theta,\alpha} = \Big(\alpha D + T^\star_\theta T_\theta\Big)^{-1}\Big\{T^{\star(1)}_\theta - \Big[T^\star_\theta T^{(1)}_\theta + T^{\star(1)}_\theta T_\theta\Big]\Big(\alpha D + T^\star_\theta T_\theta\Big)^{-1}T^\star_\theta\Big\},$$

the operators $T^{(1)}_\theta : L^2(\mathbb{R}^k) \to L^2(\mathbb{R}^q)$ and $T^{\star(1)}_\theta : L^2(\mathbb{R}^q) \to L^2(\mathbb{R}^k)$ are defined by

$$T^{(1)}_\theta h = \frac{\partial}{\partial\theta}(T_\theta h), \qquad T^{\star(1)}_\theta g = \frac{\partial}{\partial\theta}(T^\star_\theta g),$$

and

$$\widehat{T}_\theta h = \int_{\mathbb{R}} \frac{1}{n}\sum_{i=1}^n k_{b_Z}(X_i^t\theta - z)K_{b_W}(W_i - \cdot)h(z)dz,$$
$$\widehat{T}^{(1)}_\theta h = \int_{\mathbb{R}} \frac{1}{nb_Z}\sum_{i=1}^n k'_{b_Z}(X_i^t\theta - z)X_i^tK_{b_W}(W_i - \cdot)h(z)dz,$$
$$\widehat{T}^\star_\theta g = \int_{\mathbb{R}^q} \frac{1}{n}\sum_{i=1}^n k_{b_Z}(X_i^t\theta - \cdot)K_{b_W}(W_i - w)g(w)dw,$$
$$\widehat{T}^{\star(1)}_\theta g = \int_{\mathbb{R}^q} \frac{1}{nb_Z}\sum_{i=1}^n k'_{b_Z}(X_i^t\theta - \cdot)X_i^tK_{b_W}(W_i - w)g(w)dw.$$

It follows that

$$
\begin{aligned}
&\big(\widehat{Q}_{\theta,\alpha} - Q_{\theta,\alpha}\big)r \\
&= \Big[\big(\alpha D + \widehat{T}_\theta^\star \widehat{T}_\theta\big)^{-1} - \big(\alpha D + T_\theta^\star T_\theta\big)^{-1}\Big]\Big\{\widehat{T}_\theta^{\star(1)} - \big[\widehat{T}_\theta^\star \widehat{T}_\theta^{(1)} + \widehat{T}_\theta^{\star(1)}\widehat{T}_\theta\big]\big(\alpha D + \widehat{T}_\theta^\star \widehat{T}_\theta\big)^{-1}\widehat{T}_\theta^\star\Big\}r \\
&\quad + \big(\alpha D + T_\theta^\star T_\theta\big)^{-1}\big[\widehat{T}_\theta^{\star(1)} - T_\theta^{\star(1)}\big]r \\
&\quad - \big(\alpha D + T_\theta^\star T_\theta\big)^{-1}\Big\{\big[\widehat{T}_\theta^\star \widehat{T}_\theta^{(1)} + \widehat{T}_\theta^{\star(1)}\widehat{T}_\theta\big]\big(\alpha D + \widehat{T}_\theta^\star \widehat{T}_\theta\big)^{-1}\widehat{T}_\theta^\star \\
&\qquad\qquad - \big[T_\theta^\star T_\theta^{(1)} + T_\theta^{\star(1)} T_\theta\big]\big(\alpha D + T_\theta^\star T_\theta\big)^{-1} T_\theta^\star\Big\}r.
\end{aligned}
$$

Note that

$$
\begin{aligned}
&\big(\alpha D + \widehat{T}_\theta^\star \widehat{T}_\theta\big)^{-1} - \big(\alpha D + T_\theta^\star T_\theta\big)^{-1} \\
&= -\big(\alpha D + \widehat{T}_\theta^\star \widehat{T}_\theta\big)^{-1}\Big(\big[\widehat{T}_\theta^\star - T_\theta^\star\big]T_\theta + \widehat{T}_\theta^\star\big[\widehat{T}_\theta - T_\theta\big]\Big)\big(\alpha D + T_\theta^\star T_\theta\big)^{-1}
\end{aligned}
$$

and that

$$
\begin{aligned}
&\big[\widehat{T}_\theta^\star \widehat{T}_\theta^{(1)} + \widehat{T}_\theta^{\star(1)}\widehat{T}_\theta\big]\big(\alpha D + \widehat{T}_\theta^\star \widehat{T}_\theta\big)^{-1}\widehat{T}_\theta^\star - \big[T_\theta^\star T_\theta^{(1)} + T_\theta^{\star(1)} T_\theta\big]\big(\alpha D + T_\theta^\star T_\theta\big)^{-1} T_\theta^\star \\
&= \big[\widehat{T}_\theta^\star \widehat{T}_\theta^{(1)} + \widehat{T}_\theta^{\star(1)}\widehat{T}_\theta\big]\big(\alpha D + \widehat{T}_\theta^\star \widehat{T}_\theta\big)^{-1}\big[\widehat{T}_\theta^\star - T_\theta^\star\big] \\
&\quad + \Big(\big[\widehat{T}_\theta^\star - T_\theta^\star\big]\widehat{T}_\theta^{(1)} + T_\theta^\star\big[\widehat{T}_\theta^{(1)} - T_\theta^{(1)}\big] + \big[\widehat{T}_\theta^{\star(1)} - T_\theta^{\star(1)}\big]\widehat{T}_\theta + T_\theta^{\star(1)}\big[\widehat{T}_\theta - T_\theta\big]\Big)\big(\alpha D + T_\theta^\star T_\theta\big)^{-1} T_\theta^\star \\
&\quad - \big[\widehat{T}_\theta^\star \widehat{T}_\theta^{(1)} + \widehat{T}_\theta^{\star(1)}\widehat{T}_\theta\big]\big(\alpha D + \widehat{T}_\theta^\star \widehat{T}_\theta\big)^{-1}\Big(\big[\widehat{T}_\theta^\star - T_\theta^\star\big]\widehat{T}_\theta + T_\theta^\star\big[\widehat{T}_\theta - T_\theta\big]\Big)\big(\alpha D + T_\theta^\star T_\theta\big)^{-1} T_\theta^\star.
\end{aligned}
$$

Hence, it suffices to control the differences $\|\widehat{T}_\vartheta - T_\vartheta\|$, $\|\widehat{T}_\vartheta^\star - T_\vartheta^\star\|$, $\|\widehat{T}_\vartheta^{(1)} - T_\vartheta^{(1)}\|$ and $\|\widehat{T}_\vartheta^{\star(1)} - T_\vartheta^{\star(1)}\|$. The former two expressions are $o_P(1)$ by Lemma 4. The order of the latter two can be obtained by using similar arguments. The first part of the statement of the lemma now follows.

For the second part, write

$$
\frac{\partial^2}{\partial\theta\partial\theta^t}\big(\widehat{h}_{\theta,\alpha} - h^\circ_{\theta,\alpha}\big)(z) = \big(\widehat{Q}^{(1)}_{\theta,\alpha} - Q^{(1)}_{\theta,\alpha}\big)r(z) + \widehat{Q}^{(1)}_{\theta,\alpha}\big(\widehat{r}(z) - r(z)\big), \tag{7}
$$

with

$$
\begin{aligned}
\widehat{Q}^{(1)}_{\theta,\alpha} &= \big(\alpha D + \widehat{T}_\theta^\star \widehat{T}_\theta\big)^{-1}\Big\{\widehat{T}_\theta^{\star(2)} - \big[2\widehat{T}_\theta^{(1)}\widehat{T}_\theta^{\star(1)t} + \widehat{T}_\theta^\star \widehat{T}_\theta^{(2)} + \widehat{T}_\theta^{\star(2)}\widehat{T}_\theta\big]\big(\alpha D + \widehat{T}_\theta^\star \widehat{T}_\theta\big)^{-1}\widehat{T}_\theta^\star \\
&\qquad - 2\big[\widehat{T}_\theta^\star \widehat{T}_\theta^{(1)} + \widehat{T}_\theta^{\star(1)}\widehat{T}_\theta\big]\big(\alpha D + \widehat{T}_\theta^\star \widehat{T}_\theta\big)^{-1}\Big[\widehat{T}_\theta^{\star(1)} - \big[\widehat{T}_\theta^\star \widehat{T}_\theta^{(1)} + \widehat{T}_\theta^{\star(1)}\widehat{T}_\theta\big]\big(\alpha D + \widehat{T}_\theta^\star \widehat{T}_\theta\big)^{-1}\widehat{T}_\theta^\star\Big]^t\Big\}, \\
Q^{(1)}_{\theta,\alpha} &= \big(\alpha D + T_\theta^\star T_\theta\big)^{-1}\Big\{T_\theta^{\star(2)} - \big[2T_\theta^{(1)}T_\theta^{\star(1)t} + T_\theta^\star T_\theta^{(2)} + T_\theta^{\star(2)} T_\theta\big]\big(\alpha D + T_\theta^\star T_\theta\big)^{-1} T_\theta^\star \\
&\qquad - 2\big[T_\theta^\star T_\theta^{(1)} + T_\theta^{\star(1)} T_\theta\big]\big(\alpha D + T_\theta^\star T_\theta\big)^{-1}\Big[T_\theta^{\star(1)} - \big[T_\theta^\star T_\theta^{(1)} + T_\theta^{\star(1)} T_\theta\big]\big(\alpha D + T_\theta^\star T_\theta\big)^{-1} T_\theta^\star\Big]^t\Big\},
\end{aligned}
$$

where $T_\theta^{(2)}h = \frac{\partial^2}{\partial\theta\partial\theta^t}\big(T_\theta h\big)$, $T_\theta^{\star(2)}g = \frac{\partial^2}{\partial\theta\partial\theta^t}\big(T_\theta^\star g\big)$,

$$\widehat{T}_\theta^{(2)}h = \int_{\mathbb{R}} \frac{1}{nb_Z^2}\sum_{i=1}^n k_{b_Z}''\big(X_i^t\theta - z\big)X_iX_i^tK_{b_W}\big(W_i - \cdot\big)h(z)dz,$$

$$\widehat{T}_\theta^{\star(2)}g = \int_{\mathbb{R}^q} \frac{1}{nb_Z^2}\sum_{i=1}^n k_{b_Z}''\big(X_i^t\theta - \cdot\big)X_iX_i^tK_{b_W}\big(W_i - w\big)g(w)dw.$$

Using similar arguments as above, we can show that $\|\widehat{T}_\theta^{\star(2)} - T_\theta^{\star(2)}\| = O_P(1)$ and $\|\widehat{T}_\theta^{(2)} - T_\theta^{(2)}\| = O_P(1)$ uniformly over a neighborhood around ϑ. Hence, expression (7) is bounded in probability uniformly over that neighborhood. □

The proof of the main result will be based on results in Chen et al. (2003). In the latter paper high-level conditions are given under which a semiparametric Z-estimator (i.e. any parameter estimator that is obtained as the solution of a system of equations involving a nonparametric nuisance function) is weakly consistent (Theorem 1) and asymptotically normal (Theorem 2). For the asymptotic normality we need a small modification of their result, which we state below. We omit the proof.

To present this modified result, we need to introduce a number of additional notations. As before we assume that θ belongs to a parameter space Θ, and we will assume that the nuisance functions h belong to a certain space $\mathbb{H}$. This space is not necessarily equal to the space $\mathcal{H}$ introduced before, and will be chosen in the proof of the main result, in such a way that the high level conditions of Proposition 2 below will be satisfied. The space $\mathbb{H}$ is endowed with a pseudo-norm $\|\cdot\|_{\mathbb{H}}$. The functions h in $\mathbb{H}$ will often be indexed by the parameter vector θ and we will identify h with $(h_\theta)_\theta$.

Since in the result below we assume that $\widehat{\theta}$ and $\widehat{h}$ are weakly consistent, we can restrict the spaces Θ and $\mathbb{H}$ to shrinking neighborhoods around the true ϑ and h_0. Let $\Theta_\delta = \{\theta \in \Theta : \|\theta - \vartheta\| \leq \delta_n\}$ and $\mathbb{H}_\delta = \{h \in \mathbb{H} : \sup_{\theta\in\Theta_\delta}\|h_\theta - h_\theta^\circ\|_{\mathbb{H}} \leq \delta_n\}$ for some $\delta_n = o(1)$. Moreover, for any $\theta \in \Theta_\delta$, we say that $M(h_\theta, \theta)$ is pathwise differentiable at $h_\theta \in \mathbb{H}_\delta$ in the direction $[\overline{h}_\theta - h_\theta]$ if $\{h_\theta + \tau(\overline{h}_\theta - h_\theta) : \tau \in [0,1]\} \subset \mathbb{H}$ and $\lim_{\tau\to 0}[M(h_\theta + \tau(\overline{h}_\theta - h_\theta), \theta) - M(h_\theta, \theta)]/\tau$ exists. We denote the limit by $\Lambda(h_\theta, \theta)[\overline{h}_\theta - h_\theta]$.

The modification with respect to Theorem 2 in Chen et al. (2003) consists in the following two changes. First of all we suppose in the result below that $M(h_\theta, \theta)$ is linear in h_θ, which implies that condition (2.3)(i) in Theorem 2 in Chen et al. (2003) is automatically satisfied. Moreover, it can be easily seen from the proof of the latter result that the condition $\sup_{\theta\in\Theta_\delta}\|\widehat{h}_{\theta,\alpha} - h_\theta^\circ\|_{\mathbb{H}} = o_P(n^{-1/4})$ in (2.4) can in that case be replaced by $\sup_{\theta\in\Theta_\delta}\|\widehat{h}_{\theta,\alpha} - h_\theta^\circ\|_{\mathbb{H}} = o_P(1)$. The second change with respect to Theorem 2 is related

to their condition (2.3)(ii). The proof of their theorem shows that this condition can be replaced by $\|\Lambda(h_\theta^\circ,\theta)[\widehat{h}_{\theta,\alpha}-h_\theta^\circ]-\Lambda(h_0,\vartheta)[\widehat{h}_{\vartheta,\alpha}-h_0]\|\leq o_P(1)\delta_n+O_P(n^{-1/2})$, i.e. instead of taking all h_θ in a neighbourhood of h_θ°, we only consider $\widehat{h}_{\theta,\alpha}$.

Proposition 2. Suppose that $M(h_0,\vartheta)=0$ and $\widehat{\theta}-\vartheta=o_P(1)$. In addition, assume that

(C.1) $\|M_n(\widehat{h}_{\widehat{\theta},\alpha},\widehat{\theta})\|=\inf_{\theta\in\Theta_\delta}\|M_n(\widehat{h}_{\theta,\alpha},\theta)\|+o_P(n^{-1/2})$.

(C.2) The ordinary derivative $\Gamma(h_\theta^\circ,\theta):=\frac{\partial}{\partial\theta}M(h_\theta^\circ,\theta)$ exists for all $\theta\in\Theta_\delta$ and is continuous at $\theta=\vartheta$. Moreover, the matrix $\Gamma:=\Gamma(h_0,\vartheta)$ is of full (column) rank.

(C.3) For all $\theta\in\Theta_\delta$ the pathwise derivative $\Lambda(h_\theta^\circ,\theta)[h_\theta-h_\theta^\circ]$ exists in all directions $[h_\theta-h_\theta^\circ]\in\mathbb{H}$. Moreover, for all $(h_\theta,\theta)\in\mathbb{H}_\delta\times\Theta_\delta$, $M(h_\theta,\theta)$ is linear in h_θ, i.e. $M(h_\theta,\theta)-M(h_\theta^\circ,\theta)-\Lambda(h_\theta^\circ,\theta)[h_\theta-h_\theta^\circ]=0$, and $\|\Lambda(h_\theta^\circ,\theta)[\widehat{h}_{\theta,\alpha}-h_\theta^\circ]-\Lambda(h_0,\vartheta)[\widehat{h}_{\vartheta,\alpha}-h_0]\|\leq o_P(1)\delta_n+O_P(n^{-1/2})$.

(C.4) For all $\theta\in\Theta_\delta$, $P(\widehat{h}_{\theta,\alpha}\in\mathbb{H})\to1$ and $\sup_{\theta\in\Theta_\delta}\|\widehat{h}_{\theta,\alpha}-h_\theta^\circ\|_{\mathbb{H}}=o_P(1)$.

(C.5) For all sequences $\epsilon_n=o(1)$,

$$\sup_{\|\theta-\vartheta\|\leq\epsilon_n,\sup_{\theta\in\Theta_\delta}\|h_\theta-h_\theta^\circ\|_{\mathbb{H}}\leq\epsilon_n}\|M_n(h_\theta,\theta)-M(h_\theta,\theta)-M_n(h_0,\vartheta)\|=o_P(n^{-1/2}).$$

(C.6) For some positive definite matrix Σ, $n^{1/2}\{M_n(h_0,\vartheta)+\Lambda(h_0,\vartheta)[\widehat{h}_{\vartheta,\alpha}-h_0]\}\xrightarrow{d}N(0,\Sigma)$.

Then,

$$n^{1/2}(\widehat{\theta}-\vartheta)\xrightarrow{d}N(0,V),$$

where

$$V=(\Gamma^t\Gamma)^{-1}\Gamma^t\Sigma\Gamma(\Gamma^t\Gamma)^{-1}.$$

We are now ready to prove the main result of the paper.

Proof of Theorem 1. First of all, let us define the space $\mathbb{H}$ by

$$\mathbb{H}=\Big\{h:\Omega\to\mathbb{R};h,h'\text{ are absolutely continuous,}\\ \sup_{z\in\Omega}|h(z)|\leq M\text{ and }\sup_{z\in\Omega}|h'(z)|\leq M\Big\}\tag{8}$$

where M is defined in Lemma 6, and let $\|h\|_{\mathbb{H}}^2 = \|h\|_{\mathcal{H}}^2 = \int_\Omega h^2(z)\,dz$.

We will first show that $\widehat{\theta} - \vartheta = o_P(1)$ by checking the conditions of Theorem 1 in Chen et al. (2003). Condition (1.1) in the latter paper is automatically satisfied by construction, whereas (1.2) follows from assumption (A.7). For condition (1.3), write

$$\begin{aligned}\sup_\theta \|M(h_\theta,\theta) - M(h_\theta^\circ,\theta)\| &= \sup_\theta \Big\| E\Big[g(W)\Big\{h_\theta(X^t\theta) - h_\theta^\circ(X^t\theta)\Big\}\Big]\Big\| \\ &\leq \Big\|E(g^2(W))^{1/2}\Big\| \sup_\theta E\Big((h_\theta - h_\theta^\circ)^2(X^t\theta)\Big)^{1/2}.\end{aligned}$$

The former expected value is finite by assumption (A.7), whereas the latter one is bounded by $C\sup_\theta \|h_\theta - h_\theta^\circ\|_{\mathbb{H}}^2$ for some $C < \infty$, since $\sup_{\theta,z} f_{X^t\theta}(z) < \infty$ by assumption (A.3). Next, condition (1.4) follows from Proposition 1, whereas condition (1.5) can be verified in a similar way as condition (C.5) from Proposition 2, which we show below. Hence, we have shown that the conditions of Theorem 1 in Chen et al. (2003) are satisfied, except for condition (1.5) of which we postpone the verification to later.

We are now ready to show the asymptotic normality of $\widehat{\theta}$, using Proposition 2 above.

(C.1) This is automatically satisfied by construction of the estimator $\widehat{\theta}$.

(C.2) The derivative with respect to θ is

$$\begin{aligned}&\Gamma(h_\theta^\circ,\theta)(\bar\theta - \theta) \\ &= \lim_{\tau\to 0}\frac{1}{\tau}\Big\{\mathbb{E}\Big[g_{\theta+\tau(\bar\theta-\theta)}(W)\big(Y - h^\circ_{\theta+\tau(\bar\theta-\theta)}(X^t\{\theta+\tau(\bar\theta-\theta)\})\big)\Big] \\ &\qquad\qquad -\mathbb{E}\Big[g_\theta(W)\big(Y - h_\theta^\circ(X^t\theta)\big)\Big]\Big\} \\ &= \Big\{\mathbb{E}\Big[\Big(\frac{\partial}{\partial\gamma}g_\gamma(W)\Big)\Big|_{\gamma=\theta}(Y - h_\theta^\circ(X^t\theta))\Big] - \mathbb{E}\Big[g_\theta(W)h_\theta^{\circ\prime}(X^t\theta)X^t\Big] \\ &\qquad -\mathbb{E}\Big[g_\theta(W)\Big(\frac{\partial}{\partial\gamma^t}h_\gamma^\circ\Big)\Big|_{\gamma=\theta}(X^t\theta)\Big]\Big\}(\bar\theta - \theta).\end{aligned}$$

The first part of (C.2) is therefore fulfilled if $h_\theta^{\circ\prime}$, $\partial h_\theta^\circ/\partial\theta$ and $\partial g_\theta/\partial\theta$ exist and are continuous in θ (guaranteed by assumption (A.6) (a) and (A.7)). The second part is fulfilled by assumption (A.6) (b). Note that for $\theta = \vartheta$ we have $h_\vartheta^\circ = h_0$ and the expression reduces to

$$\Gamma(h_\vartheta^\circ,\vartheta) = -\mathbb{E}\Big[g_\vartheta(W)h_0'(X^t\vartheta)X^t\Big] - \mathbb{E}\Big[g_\vartheta(W)\Big(\frac{\partial}{\partial\gamma^t}h_\gamma^\circ\Big)\Big|_{\gamma=\vartheta}(X^t\vartheta)\Big]$$

since

$$\mathbb{E}\Big[\Big(\frac{\partial}{\partial\gamma}g_\gamma(W)\Big)\Big|_{\gamma=\vartheta}(Y - h_\vartheta^\circ(X^t\vartheta))\Big] = \mathbb{E}\Big[\Big(\frac{\partial}{\partial\gamma}g_\gamma(W)\Big)\Big|_{\gamma=\vartheta}\mathbb{E}[(Y - h_0(X^t\vartheta))|W]\Big] = 0.$$

(C.3) We calculate the functional derivative as

$$\begin{aligned}\Lambda(h_\theta^\circ,\theta)[h_\theta - h_\theta^\circ] &= \lim_{\tau\to 0}\frac{1}{\tau}\Big\{\mathbb{E}\Big[g_\theta(W)\big(Y - \{h_\theta^\circ + \tau(h_\theta - h_\theta^\circ)\}(X^t\theta)\big)\Big] \\ &\qquad -\mathbb{E}\Big[g_\theta(W)\big(Y - h_\theta^\circ(X^t\theta)\big)\Big]\Big\} \\ &= -\mathbb{E}\Big[g_\theta(W)(h_\theta - h_\theta^\circ)(X^t\theta)\Big].\end{aligned}$$

It follows that the first part of (C.3) is fulfilled, i.e.

$$M(h_\theta,\theta) - M(h_\theta^\circ,\theta) - \Lambda(h_\theta^\circ,\theta)[h_\theta - h_\theta^\circ] = 0.$$

For the second part we have

$$\begin{aligned}&\Big\|\Lambda(h_\theta^\circ,\theta)[\widehat{h}_{\theta,\alpha} - h_\theta^\circ] - \Lambda(h_\vartheta^\circ,\vartheta)[\widehat{h}_{\vartheta,\alpha} - h_\vartheta^\circ]\Big\| \\ &= \Big\|\mathbb{E}\Big[g_\theta(W)\Big\{(\widehat{h}_{\vartheta,\alpha} - h_\vartheta^\circ)(X^t\theta) - (\widehat{h}_{\vartheta,\alpha} - h_\vartheta^\circ)(X^t\vartheta) + (\widehat{h}_{\theta,\alpha} - h_\theta^\circ - (\widehat{h}_{\vartheta,\alpha} - h_\vartheta^\circ))(X^t\theta)\Big\}\Big]\Big\| \\ &\le C\Big(\mathbb{E}\Big\|g_\theta(W)\Big\|^2\Big)^{1/2}\Big(\mathbb{E}[I_1^2] + \mathbb{E}[I_2^2] + \mathbb{E}[I_3^2] + \mathbb{E}[I_4^2]\Big)^{1/2},\end{aligned}$$

with $\widehat{d}_{\theta,\alpha} = \widehat{h}_{\theta,\alpha} - h_{\theta,\alpha}^\circ$, $d_{\theta,\alpha} = h_{\theta,\alpha}^\circ - h_\theta^\circ$, and

$$\begin{aligned}I_1 &= \widehat{d}_{\vartheta,\alpha}(X^t\theta) - \widehat{d}_{\vartheta,\alpha}(X^t\vartheta), & I_2 &= \widehat{d}_{\theta,\alpha}(X^t\theta) - \widehat{d}_{\vartheta,\alpha}(X^t\theta), \\ I_3 &= d_{\vartheta,\alpha}(X^t\theta) - d_{\vartheta,\alpha}(X^t\vartheta), & I_4 &= d_{\theta,\alpha}(X^t\theta) - d_{\vartheta,\alpha}(X^t\theta).\end{aligned}$$

The terms $\mathbb{E}[I_3^2]$ and $\mathbb{E}[I_4^2]$ are $O(\alpha^{\gamma\wedge 2})$ (as was shown at the end of the proof of Proposition 1), which is $O(n^{-1/2})$ under assumption (A.5). We obtain by a Taylor expansion of $\widehat{d}_{\theta,\alpha}(X^t\theta)$ around $X^t\vartheta$ that

$$I_1 \;=\; \widehat{d}'_{\vartheta,\alpha}(X^t\vartheta)X^t(\theta-\vartheta) + \frac{1}{2}\widehat{d}''_{\vartheta,\alpha}(\xi_n)(\theta-\vartheta)^t XX^t(\theta-\vartheta).$$

Here ξ_n denotes a random variable between $X^t\theta$ and $X^t\vartheta$. Because X belongs to some compact set we have $\|XX^t\| \le \kappa$ for some $\kappa < \infty$. Hence,

$$\begin{aligned}\mathbb{E}[I_1^2] &\le C_1\Big|(\theta-\vartheta)^t\mathbb{E}\Big[\widehat{d}'^2_{\vartheta,\alpha}(X^t\vartheta)XX^t\Big](\theta-\vartheta) \\ &\qquad +(\theta-\vartheta)^t\mathbb{E}\Big[\widehat{d}''^2_{\vartheta,\alpha}(\xi_n)^2XX^t(\theta-\vartheta)(\theta-\vartheta)^tXX^t\Big](\theta-\vartheta)\Big| \\ &\le C_2\|\theta-\vartheta\|^2\mathbb{E}\Big[\widehat{d}'^2_{\vartheta,\alpha}(X^t\vartheta)\Big] + C_3\|\theta-\vartheta\|^4\mathbb{E}\Big[\widehat{d}''^2_{\vartheta,\alpha}(\xi_n)\Big] \\ &\le o_P(1)\delta_n^2 + O_P(\delta_n^4)\end{aligned}$$

by Lemma 6, uniformly over all $\|\theta-\vartheta\| \le \delta_n$, and where $C_2 = C_1\kappa$ and $C_3 = C_1\kappa^2$. Using a similar development for I_2, we obtain that

$$\mathbb{E}[I_2^2] \;\le\; C_4\Big(\|\theta-\vartheta\|^2\mathbb{E}\Big[\Big\|\frac{\partial}{\partial\gamma}\widehat{d}_{\gamma,\alpha}\Big|_{\gamma=\vartheta}(X^t\theta)\Big\|^2\Big] + \|\theta-\vartheta\|^4\mathbb{E}\Big[\Big\|\frac{\partial^2}{\partial\gamma\partial\gamma^t}\widehat{d}_{\gamma,\alpha}\Big|_{\gamma=\eta}(X^t\theta)\Big\|^2\Big]\Big),$$

for some η on the line segment between θ and ϑ. The latter is of the order $o_P(1)\delta_n^2 + O_P(\delta_n^4)$ by Lemma 7. This shows that the second part of (C.3) is satisfied.

(C.4) Using the definition of the space $\mathbb{H}$ given in equation (8) and using Lemma 6, it is easily seen that for all $\theta \in \Theta_\delta$ we have that $P(\widehat{h}_{\theta,\alpha} \in \mathbb{H}) \to 1$ as n tends to infinity. Moreover, $\sup_\theta \|\widehat{h}_{\theta,\alpha} - h_\theta^\circ\|_{\mathbb{H}} = o_P(1)$ by Proposition 1 and the definition of the norm $\|\cdot\|_{\mathbb{H}}$.

(C.5) For proving condition (C.5), we make use of Theorem 3 in Chen et al. (2003). If conditions (3.2) and (3.3) in the latter theorem are verified, then (C.5) holds true. Condition (3.2) is easily seen to be valid for $r = 2$ and $s_j = 1$ $(j = 1, \ldots, \ell)$ (using the notation of Chen et al. (2003)). For (3.3), it follows from e.g. Theorem 2.7.1 in Van der Vaart & Wellner (1996) that $\log N(\epsilon, \mathbb{H}, \|\cdot\|_{\mathbb{H}}) \leq K\epsilon^{-1}$ for all $0 \leq \epsilon \leq M$, where $N(\epsilon, \mathbb{H}, \|\cdot\|_{\mathbb{H}})$ is the covering number, i.e. the smallest number of balls of $\|\cdot\|_{\mathbb{H}}$-radius ϵ needed to cover the space $\mathbb{H}$. Condition (3.3) now follows.

(C.6) We need to show the asymptotic normality of

$$M_n(h_\vartheta^\circ, \vartheta) + \Lambda(h_\vartheta^\circ, \vartheta)\Big(\widehat{h}_{\vartheta,\alpha} - h_\vartheta^\circ\Big).$$

Now, $M_n(h_\vartheta^\circ, \vartheta)$ is already a sum of independent identically distributed random variables with zero mean, since

$$\mathbb{E}[M_n(h_\vartheta^\circ, \vartheta)] = M(h_\vartheta^\circ, \vartheta) = 0.$$

From the first part of (C.3) we have

$$\begin{aligned}
\Lambda(h_\vartheta^\circ, \vartheta)[\widehat{h}_{\vartheta,\alpha} - h_\vartheta^\circ] &= M(\widehat{h}_{\vartheta,\alpha}, \vartheta) - M(h_\vartheta^\circ, \vartheta) \\
&= -\mathbb{E}[g_\vartheta(W)(\widehat{h}_{\vartheta,\alpha}(X^t\vartheta) - h_\vartheta^\circ(X^t\vartheta))] \\
&= -\int_{\mathbb{R}^2} g_\vartheta(w)(\widehat{h}_{\vartheta,\alpha}(z) - h_\vartheta^\circ(z)) f_{X^t\vartheta,W}(z,w)d(z,w).
\end{aligned}$$

We can write

$$\begin{aligned}
&\int_{\mathbb{R}^2} g_\vartheta(w)(\widehat{h}_{\vartheta,\alpha}(z) - h_\vartheta^\circ(z)) f_{X^t\vartheta,W}(z,w)d(z,w) = \int_{\mathbb{R}} g_\vartheta(w) S_\vartheta L(\widehat{h}_{\vartheta,\alpha} - h_\vartheta^\circ)(w)dw \\
&= \langle g_\vartheta, S_\vartheta(S_\vartheta^\star S_\vartheta)^{-1} S_\vartheta^\star(\widehat{r} - \widehat{S}_\vartheta L h_\vartheta^\circ)\rangle + \langle g_\vartheta, S_\vartheta[(\alpha I + \widehat{S}_\vartheta^\star \widehat{S}_\vartheta)^{-1}\widehat{S}_\vartheta^\star - (S_\vartheta^\star S_\vartheta)^{-1} S_\vartheta^\star]\widehat{r}\rangle \\
&\quad + \langle g_\vartheta, S_\vartheta(S_\vartheta^\star S_\vartheta)^{-1}[\widehat{S}_\vartheta^\star - S_\vartheta^\star] L h_\vartheta^\circ\rangle \\
&= I_{n,1} + I_{n,2} + I_{n,3}.
\end{aligned}$$

First, note that

$$
\begin{aligned}
&(S_\vartheta^\star S_\vartheta)^{-1} S_\vartheta^\star (\widehat{r} - \widehat{S}_\vartheta L h_\vartheta^\circ)(z) \\
&= \frac{1}{n b_W} \sum_{i=1}^n \Big\{ (T_\vartheta^\star T_\vartheta)^{-1} \int K\Big(\frac{W_i - w}{b_W}\Big) \Big[Y_i - \int \frac{1}{b_Z} k\Big(\frac{X_i^t \vartheta - z}{b_Z}\Big) h_\vartheta^\circ(z) dz \Big] f_{X^t\vartheta, W}(\cdot, w) dw \Big\}(z) \\
&= \frac{1}{n} \sum_{i=1}^n U_i \Big\{ (T_\vartheta^\star T_\vartheta)^{-1} \int K(w) f_{X^t\vartheta, W}(\cdot, W_i - b_W w) dw \Big\}(z) \\
&\quad - \mu_m(k) \frac{b_Z^p}{n} \sum_{i=1}^n h_\vartheta^{\circ(p)}(X_i^t \vartheta) \Big\{ (T_\vartheta^\star T_\vartheta)^{-1} \int K(w) f_{X^t\vartheta, W}(\cdot, W_i - b_W w) dw \Big\}(z) \\
&\quad + o_P(b_Z^p) \frac{1}{n} \sum_{i=1}^n \Big\{ (T_\vartheta^\star T_\vartheta)^{-1} \int K(w) f_{X^t\vartheta, W}(\cdot, W_i - b_W w) dw \Big\}(z) \\
&= \frac{1}{n} \sum_{i=1}^n U_i \{ (T_\vartheta^\star T_\vartheta)^{-1} f_{X^t\vartheta, W}(\cdot, W_i) \}(z) + O_P(b_Z^p) + O_P(b_W^\rho),
\end{aligned}
$$

where both remainder terms are of order $o_P(n^{-1/2})$ since $n b_Z^{2p} \to 0$ and $n b_W^{2\rho} \to 0$. Hence, we obtain for $I_{n,1}$:

$$
I_{n,1} = \frac{1}{n} \sum_{i=1}^n U_i \int g_\vartheta(w) \{ (T_\vartheta^\star T_\vartheta)^{-1} f_{X^t\vartheta, W}(\cdot, W_i) \}(z) f_{X^t\vartheta, W}(z, w) d(z, w) + o_P(n^{-1/2}),
$$

which gives the following contribution to the i.i.d. sum :

$$
-\frac{1}{n} \sum_{i=1}^n U_i \int g_\vartheta(w) \{ (T_\vartheta^\star T_\vartheta)^{-1} f_{X^t\vartheta, W}(\cdot, W_i) \}(z) f_{X^t\vartheta, W}(z, w) d(z, w).
$$

Further we get

$$
\begin{aligned}
I_{n,3} &= \langle (S_\vartheta^\star S_\vartheta)^{-1} S_\vartheta^\star g_\vartheta, [\widehat{S}_\vartheta^\star - S_\vartheta^\star] L h_\vartheta^\circ \rangle \\
&= \int \int (S_\vartheta^\star S_\vartheta)^{-1} S_\vartheta^\star g_\vartheta(w) L h_\vartheta^\circ(z) \Big[(\widehat{f}_{X^t\vartheta, W}(z, w) - E[\widehat{f}_{X^t\vartheta, W}(z, w)]) \\
&\qquad + (E[\widehat{f}_{X^t\vartheta, W}(z, w)] - f_{X^t\vartheta, W}(z, w)) \Big] dz\, dw,
\end{aligned}
$$

where, by standard calculations, the bias part is of order $O((b_Z \vee b_W)^{2\rho}) = o(n^{-1/2})$ and the stochastic part is of order $o_P(n^{-1/2})$ due to the additional integration. In a somewhat similar way we can show that also $I_{n,2}$ is asymptotically negligible. $\square$