Estimating the error distribution in nonlinear and nonparametric regression

Ursula U. Müller
Universität Bremen

Anton Schick *
Binghamton University

Wolfgang Wefelmeyer
Universität Siegen

Abstract

We consider estimation of linear functionals of the error distribution for two regression models: nonlinear and nonparametric, and for two types of errors: independent of the covariate (type I), and unspecified (type II). We show that the residual-based empirical estimators for the *nonparametric* type I model remain efficient in the type II model. For the *nonlinear* type I regression model, efficient estimators are obtained by correcting the empirical estimator using that the errors are centered, and using an efficient estimator for the regression parameter. Since such efficient parameter estimators are not consistent in the nonlinear type II model, the empirical estimator is not consistent either. We construct efficient estimators for linear functionals of the error distribution in the nonlinear type II regression model, starting from residual-based empirical estimators, correcting it for the fact that the errors are conditionally centered, and using an appropriate efficient weighted least squares estimator for the regression parameter.

Key words and Phrases. Plug-in estimator, local polynomial smoother, i.i.d. representation, constrained model, Donsker class.

1. Introduction

Suppose we have i.i.d. observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ from a regression model $Y = r(X)+\varepsilon$. We are interested in efficient estimation of a linear functional $E[h(\varepsilon)]$ of the error distribution. A natural estimator is the empirical estimator

$$\hat{H} = \frac{1}{n} \sum_{i=1}^{n} h(\hat{\varepsilon}_i)$$

based the residuals $\hat{\varepsilon}_i = Y_i - \hat{r}_i$, with $\hat{r}_i$ an estimator for $r(X_i)$. We discuss when $\hat{H}$ is efficient, and under which structural assumptions on the regression model it can be improved.

*Supported in part by NSF Grant DMS 0072174

We distinguish between two types of model. Historically, it was first assumed that the errors are independent of the covariates. For identifiability, we assume that they are centered, $E[\varepsilon] = 0$. We call such models type I models. Most of the literature makes these assumptions. In many applications, especially in the recent econometrics literature, independence of error and covariate is considered too strong an assumption. For identifiability, we assume that the error is now conditionally centered given the covariate, $E(\varepsilon \mid X) = 0$. We call such models type II models.

For each of the two types of model, we focus attention on two models for the regression function. In the *nonparametric* regression model, the function r is unspecified (up to smoothness). In the *nonlinear* regression model, the function $r = r_\vartheta$ is assumed known up to a finite-dimensional parameter ϑ. This includes of course also the *linear* regression model.

Type II models can also be characterized as bivariate models with a possible constraint on the conditional distribution of Y given X. The nonparametric regression model is just the full nonparametric bivariate model, with no structural constraint, but with smoothness assumptions on the conditional expectation $r(X) = E(Y \mid X)$. It depends on the problem at hand which description is more convenient. For the calculation of efficient influence functions in Sections 2 and 3 we find it convenient to use the same parametrization of the law $P(dx, dy)$ of (X, Y) for all four models: by the regression function r, the covariate distribution M, and the conditional density $f(x, z)$ of ε given $X = x$,

$$P(dx, dy) = M(dx) f(x, y - r(x))\, dy.$$

On the other hand, for the construction of efficient estimators it is better to use the simplest description for the specific model at hand, and to avoid introducing unnecessary parameters. The nonlinear type II regression model is now best described by the constraint $E(Y \mid X) = r_\vartheta(X)$, which suggests a weighted least squares estimator for ϑ. The nonparametric type II regression model is best described as the nonparametric bivariate model, with $E[h(Y - r(X))]$ as functional of interest. The residual-based empirical estimator $\hat{H}$ is then seen as a plug-in estimator for a specific functional on a nonparametric model.

The contrast between model descriptions convenient for efficiency considerations and for constructions of estimators is also reflected in the organization of the paper. Sections 2 and 3 calculate efficient influence functions for arbitrary type II and type I models, respectively, while in Sections 4 and 5 we distinguish between nonlinear and nonparametric regression models, the essential difference now lying in the estimation of the regression function, through local polynomial smoothers and through parameter estimators, respectively.

Specifically, we obtain the following results. For the *nonparametric* type II model we show in Theorem 1 that the residual-based empirical estimator has the i.i.d. representation

$$\hat{H} = \frac{1}{n} \sum_{i=1}^{n} \big(h(\varepsilon_i) + \mu'(X_i, 0) \varepsilon_i \big) + o_p(n^{-1/2}),$$

where $\mu'(X_i, 0)$ is the derivative of $v \mapsto \int h(z-v)f(X_i, z)\,dz$ at $v = 0$, and $\hat{\varepsilon}_i = Y_i - \hat{r}_i$ with $\hat{r}_i$ a leave-one-out polynomial smoother. In particular, we allow a *linear* smoother. In that case we assume that the regression function has one derivative fulfilling a Hölder condition, and we take a bandwidth of smaller order than the optimal bandwidth for estimating the regression function under these conditions, i.e. we undersmooth.

It follows from Section 2 that $\hat{H}$ is efficient. We note that efficiency is almost automatic here because the nonparametric type II model is just the full nonparametric bivariate model, and efficiency theory tells us that the i.i.d. representation of regular estimators in nonparametric models is unique. To use this argument, we would however need to show that $\hat{H}$ is regular, which is most easily done by checking that its influence function is the efficient one, i.e. by doing what we did anyway.

In the nonparametric regression model, the assumption $E(\varepsilon \mid X) = 0$ does not constitute a restriction: it is needed for identifiability, and for interpreting the regression function as the conditional mean of Y given X. This is reflected in the following observation. One might think of using $E(\varepsilon \mid X) = 0$ to improve $\hat{H}$ by subtracting a correction term $\frac{1}{n}\sum_{i=1}^n a(X_i)\hat{\varepsilon}_i$, where a is a possibly random weight function. However, we can estimate the regression function in such a way that

$$\frac{1}{n}\sum_{i=1}^{n} a(X_i)\hat{\varepsilon}_i = o_p(n^{-1/2}) \tag{1.1}$$

for all weight functions a; see the proof of Theorem 1. This is an empirical version of $E[a(X)\varepsilon] = 0$ for all a, which is another way of saying that $E(\varepsilon \mid X) = 0$. In view of (1.1), any possible correction term based on $E(\varepsilon \mid X) = 0$ would be negligible.

In the type I models, the conditional expectation $\mu'(X_i, 0)$ does not depend on X_i, and the i.i.d. representation reduces to that proved in Müller, Schick and Wefelmeyer (2003) (in the following: MSW) for the same estimator. Hence that estimator is robust against non-independence of error and covariate. As shown in MSW, the estimator is also efficient in the nonparametric type I model.

Theorem 1 excludes functions h with jumps, e.g. indicator functions. In particular, it does not cover the residual-based empirical distribution function. Akritas and Keilegom (2001) consider the heteroscedastic regression model $Y = r(X) + s(X)\varepsilon$, with ε and X independent, and use empirical process theory to obtain an i.i.d. representation for the empirical distribution function based on residuals $\hat{\varepsilon}_i = (Y_i - \hat{r}_i)/\hat{s}_i$. Their model is between our nonparametric regression models of types I and II. In Theorem 2 we indicate how to adapt their technique to our type II model. Note that our functional is different: In our model, their functional would be $E[h(\varepsilon/s(X))]$.

In Section 5 we consider the *nonlinear* type II regression model $Y = r_\vartheta(X) + \varepsilon$. Now we estimate r_ϑ by $r_{\hat{\vartheta}}$, where $\hat{\vartheta}$ is a $n^{1/2}$-consistent estimator of ϑ, and obtain the following i.i.d.

representation for the empirical estimator based on the residuals $\hat{\varepsilon}_i = Y_i - \hat{r}_\vartheta(X_i)$:

$$\hat{H} = \frac{1}{n}\sum_{i=1}^{n} h(\varepsilon_i) + E[\mu'(X,0)\dot{r}_\vartheta(X)^\top](\hat{\vartheta} - \vartheta) + o_p(n^{-1/2}).$$

This estimator is no longer efficient, even if an efficient estimator $\hat{\vartheta}$ for ϑ is used. The reason is that now $E(\varepsilon \mid X) = 0$ is, in general, a restriction on the model, and we can no longer achieve (1.1) with an estimator of the form $\hat{r}_i = r_{\hat{\vartheta}}(X_i)$. However, this allows us to improve the estimator $\hat{H}$. The efficient influence function is calculated in Section 2 and suggests subtracting a correction term from $\hat{H}$, namely

$$\hat{C} = \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{\rho}_i}{\hat{\tau}_i^2}\,\hat{\varepsilon}_i,$$

where $\hat{\rho}_i$ estimates $E(\varepsilon_i h(\varepsilon_i) \mid X_i)$, and $\hat{\tau}_i^2$ estimates $E(\varepsilon_i^2 \mid X_i)$. We show in Theorem 4 that $\hat{C}$ has a stochastic expansion

$$\hat{C} = \frac{1}{n}\sum_{i=1}^{n} \frac{E(\varepsilon_i h(\varepsilon_i) \mid X_i)}{E(\varepsilon_i^2 \mid X_i)}\,\varepsilon_i + E\Big[\frac{E(\varepsilon h(\varepsilon) \mid X)}{E(\varepsilon^2 \mid X)}\,\dot{r}_\vartheta(X)^\top\Big](\hat{\vartheta} - \vartheta) + o_p(n^{-1/2}).$$

The corrected estimator $\hat{H} - \hat{C}$ is efficient in the nonlinear type II model if an efficient estimator $\hat{\vartheta}$ for ϑ is used. Such an estimator $\hat{\vartheta}$ can be obtained as a weighted least squares estimator with (optimal) random weights. Various such weighted least squares estimators have already been constructed in the literature: see Carroll (1982), Müller and Stadtmüller (1987), Robinson (1987), Schick (1987) and Chiou and Müller (1999). Some of those authors have used additional structure on the conditional second moment, e.g. that $E(\varepsilon^2 \mid X)$ is a function of the regression function. These weighted least squares estimators are not efficient in our model, and also not in the models with additional structure; improvements under such additional structural assumptions are obtained in Wefelmeyer (1996), Schick (1999), and Li (2000) and (2001).

Now consider the nonlinear type I model, with ε and X independent. Then the constraint $E(\varepsilon \mid X) = 0$ reduces to $E[\varepsilon] = 0$, and the correction term $\hat{C}$ can be replaced by the simpler term

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\sum_{i=1}^n \hat{\varepsilon}_i h(\hat{\varepsilon}_i)}{\sum_{i=1}^n \hat{\varepsilon}_i^2}\,\hat{\varepsilon}_i.$$

In the type I model, the weighted least squares estimator $\hat{\vartheta}$ is asymptotically equivalent to the ordinary least squares estimator, which is known to be inefficient. For efficiency of $\hat{H} - \hat{C}$, we must replace $\hat{\vartheta}$ by an estimator that is efficient in the type I model; see Schick (1993) for a construction. The behavior of the residual-based empirical distribution function for linear type I regression models is studied among others by Koul (1969, 1970, 2002) and Loynes (1980); for increasing dimension see Portnoy (1986) and Mammen (1996). The nonlinear autoregressive

model $X_i = r_\vartheta(X_{i-1}) + \varepsilon_i$ with independent innovations is closely related to the nonlinear type I regression model; an efficient estimator for $E[h(\varepsilon)]$ in this autoregressive model is constructed in Schick and Wefelmeyer (2002).

2. Efficient influence functions for type II models

Consider the regression model $Y = r(X) + \varepsilon$ with $E(\varepsilon \mid X) = 0$. Denote the distribution of the covariate by $M(dx)$. Suppose that the conditional distribution of ε given $X = x$ has a Lebesgue density $z \mapsto f(x,z)$. In this section we calculate efficient influence functions of arbitrary real-valued functionals of (r, f, M) for general type II models, and specialize the result to functionals $E[h(\varepsilon)] = \int h(z) f(x,z)\, dz\, M(dx)$, and to nonparametric and nonlinear regression. Write $Q(dx,dz) = M(dx) f(x,z)\, dz$ for the joint law of (X,ε) and $P(dx,dy) = M(dx) f(x, y - r(x))\, dy$ for the joint law of (X,Y), and write $g(z) = \int f(x,z) M(dx)$ for the density of ε. We impose the following assumptions on the conditional density. They are natural extensions of the assumptions required in MSW for type I models and were already used in Koul and Schick (2003).

Assumption A1. There exist positive c_τ, C_τ such that

$$c_\tau \le \tau^2(x) = E(\varepsilon^2 \mid X = x) = \int z^2 f(x,z)\, dz \le C_\tau.$$

Assumption A2. The map $z \mapsto f(x,z)$ is uniformly integrable in the sense that

$$\sup_x \int_{|z|>A} z^2 f(x,z)\, dz \to 0 \quad \text{as } A \to \infty.$$

Assumption A3. The map $z \mapsto f(x,z)$ is absolutely continuous with a.e. derivative $z \mapsto f'(x,z)$ for each x, and

$$\iint L(x,z)^2 f(x,z)\, dz\, M(dx) < \infty \quad \text{with} \quad L = -f'/f.$$

Note that $L(x,\cdot)$ is the usual score function for the location model generated by the density $f(x,\cdot)$. Thus Assumptions A1 and A3 imply

$$\int L(x,z) f(x,z)\, dz = 0 \quad \text{and} \quad \int z L(x,z) f(x,z)\, dz = 1. \tag{2.1}$$

Under the above assumptions, we have local asymptotic normality of the regression model for local perturbations of the parameters r, f, M, described as follows.

For simplicity we treat the nonlinear and nonparametric models as special cases of some *arbitrary* model whose tangent space at the true covariate distribution M is some closed linear subset U of $L_2(M)$. This has the additional advantage that our calculations may later also be used for other regression problems, with *semiparametric* models for the regression function. For $u \in U$ we consider a perturbation r_{nu} such that

$$\int (r_{nu} - r - n^{-1/2}u)^2 \, dM = o(n^{-1}).$$

For the error distribution we restrict ourselves, again for simplicity, to the case where the only restriction is conditional centering $E(\varepsilon \mid X) = 0$, but the conditional density f is unspecified otherwise. This excludes, for example, *parametric* models for f. The calculations below could however be modified to cover such models. The tangent space at the true conditional density f is then

$$V = \{v \in L_2(Q) : \int v(x,z) f(x,z) \, dz = 0, \int z v(x,z) f(x,z) \, dz = 0\}.$$

This is the space of functions orthogonal to functions in $L_2(Q)$ of the form $a(x) + b(x)z$. For $v \in V$ we consider a perturbation f_{nv} such that

$$\iint \left(f_{nv}^{1/2}(x,z) - f^{1/2}(x,z) - \frac{1}{2} \, n^{-1/2} v(x,z) f^{1/2}(x,z) \right)^2 dz \, M(dx) = o(n^{-1}). \tag{2.2}$$

See Koul and Schick (2003) for a construction. For the covariate distribution we allow some arbitrary model, with tangent space W a closed linear subspace of

$$L_{2,0}(M) = \{w \in L_2(M) : \int w \, dM = 0\}.$$

The two cases of interest to us will be: known covariate distribution ("fixed design"), $W = \{0\}$, and completely unknown covariate distribution, $W = L_{2,0}(M)$. For $w \in W$ we consider a perturbation M_{nw} such that

$$\int \left(dM_{nw}^{1/2} - dM^{1/2} - \frac{1}{2} \, n^{-1/2} w \, dM^{1/2} \right)^2 = o(n^{-1}).$$

The tangent space of the law P of (X, Y) is now obtained as follows. Let $P_{nuvw}(dx, dy) = M_{nw}(dx) f_{nv}(x, y - r_{nu}(x)) \, dy$. Then P_{nuvw} has tangent

$$t_{uvw}(x,y) = u(x) L(x, y - r(x)) + v(x, y - r(x)) + w(x)$$

in the sense that

$$\int \left(dP_{nuvw}^{1/2} - dP^{1/2} - \frac{1}{2} \, n^{-1/2} t_{uvw} dP^{1/2} \right)^2 = o(n^{-1}). \tag{2.3}$$

Hence we have *local asymptotic normality*

$$\log \frac{dP^n_{nuvw}}{dP^n} = n^{-1/2} \sum_{i=1}^{n} t_{uvw}(X_i, Y_i) - \frac{1}{2} E[t_{uvw}(X,Y)^2] + o_p(1). \tag{2.4}$$

In order to express the tangent space of the model as sum of orthogonal subspaces, we rewrite the tangent t_{uvw} as follows. The projection k_V of $k \in L_2(Q)$ onto V is

$$k_V(x,z) = k(x,z) - \int k(x,s) f(x,s)\, ds - \frac{\int sk(x,s) f(x,s)\, ds}{\tau^2(x)}\, z.$$

In particular, by (2.1), the projection L_V of L onto V is

$$L_V(x,z) = L(x,z) - \frac{z}{\tau^2(x)}.$$

For better comparison with (2.4) it is convenient to express the tangent as a random variable:

$$\begin{aligned} t_{uvw}(X,Y) &= u(X)L(X,\varepsilon) + v(X,\varepsilon) + w(X) \\ &= \frac{u(X)}{\tau^2(X)}\,\varepsilon + \big(u(X)L_V(X,\varepsilon) + v(X,\varepsilon)\big) + w(X). \end{aligned}$$

We see that the tangent space

$$T = \{t_{uvw}(X,Y) : u \in U, v \in V, w \in W\}$$

is the sum of the pairwise orthogonal spaces

$$\overline{U} = \Big\{ \frac{u(X)}{\tau^2(X)}\,\varepsilon : u \in U \Big\}, \quad \overline{V} = \{v(X,\varepsilon) : v \in V\}, \quad \overline{W} = \{w(X) : w \in W\}.$$

Now consider a real-valued functional χ of (r,f,M). Suppose that χ is differentiable at (r,f,M) with *natural gradient* $(\overline{u},\overline{v},\overline{w}) \in U \times V \times W$ in the sense that for all $(u,v,w) \in U \times V \times W$,

$$n^{1/2}(\chi(r_{nu}, f_{nv}, M_{nw}) - \chi(r,f,M)) \to \int \overline{u} u\, dM + \int \overline{v} v\, dQ + \int \overline{w} w\, dM.$$

The *efficient influence function* is defined as the element

$$t^*(X,Y) = \frac{u^*(X)}{\tau^2(X)}\,\varepsilon + v^*(X,\varepsilon) + w^*(X)$$

of T that expresses the derivative of χ in terms of the inner product inherited from local asymptotic normality (2.4): For all $(u,v,w) \in U \times V \times W$,

$$\int t^* t_{uvw}\, dP = \int \overline{u} u\, dM + \int \overline{v} v\, dQ + \int \overline{w} w\, dM. \tag{2.5}$$

For short, we call (u^*, v^*, w^*) the *LAN gradient* of χ.

The reason for calling t^* the efficient influence function lies in the following semiparametric version of Hájek's (1970) convolution theorem. Call an estimator $\hat{\chi}$ *regular* for χ at (r, f, M) with *limit* K if K is a random variable such that, for all $(u, v, w) \in U \times V \times W$,

$$n^{1/2}(\hat{\chi} - \chi(r_{nu}, f_{nv}, M_{nw})) \Rightarrow K \quad \text{under } P_{nuvw}.$$

By the convolution theorem, K is the convolution of a normal random variable with mean zero and variance $E[t^*_{uvw}(X, Y)^2]$, and another random variable. This justifies calling a regular estimator $\hat{\chi}$ *efficient* if it is asymptotically normal with this variance. Also, an estimator $\hat{\chi}$ is regular and efficient for χ at (r, f, M) if and only if it is *asymptotically linear* with *influence function* equal to t^*, i.e.,

$$n^{1/2}(\hat{\chi} - \chi(r, f, M)) = n^{-1/2} \sum_{i=1}^{n} t^*(X_i, Y_i) + o_p(1).$$

See Bickel, Klaassen, Ritov and Wellner (1998, Section 3.3) for these results.

To calculate the efficient influence function t^*_{uvw}, we introduce the following notation. Let ψ be the function defined by

$$\psi(X) = E(\overline{v}(X, \varepsilon)L(X, \varepsilon) \mid X).$$

Let Π denote the projection operator from $L_2(M)$ onto $\{u/\tau : u \in U\}$. We show now that the LAN gradient of χ is $(u^*, \overline{v}, \overline{w})$ with

$$u^* = \tau\Pi(\tau(\overline{u} - \psi)).$$

For these choices, the left-hand side of (2.5) becomes

$$\int \frac{u^*}{\tau^2} u \, dM + \int \psi u \, dM + \int \overline{v} v \, dQ + \int \overline{w} w \, dM.$$

Here we have used (2.1) and the orthogonality properties of $v(X, \varepsilon)$, $w(X)$, and ε. Since for $a \in L_2(M)$ and $u \in U$,

$$\int au \, dM = \int \tau a \frac{u}{\tau} \, dM = \int \Pi(\tau a) \frac{u}{\tau} \, dM,$$

we see that

$$\int \frac{u^*}{\tau^2} u \, dM = \int \Pi(\tau(\overline{u} - \psi)) \frac{u}{\tau} \, dM = \int (\overline{u} - \psi) u \, dM,$$

and (2.5) is immediate.

We are interested in linear functionals of the error distribution,

$$\chi(r, f, M) = \iint h(z) f(x, z) \, dz \, M(dx) = E[h(\varepsilon)].$$

The natural gradient of such a functional is $(\overline{u}, \overline{v}, \overline{w})$, where now $\overline{u} = 0$; $\overline{v}$ is the projection onto V of the function $(x, z) \mapsto h(z)$, so that

$$\overline{v}(X, \varepsilon) = h_V(X, \varepsilon) = h(\varepsilon) - E(h(\varepsilon) \mid X) - \frac{E(\varepsilon h(\varepsilon) \mid X)}{\tau^2(X)}\, \varepsilon;$$

and $\overline{w}$ is the projection of $\tilde{h}$ onto W, where

$$\tilde{h}(X) = E(h(\varepsilon) \mid X).$$

Since $\overline{u} = 0$, we have $u^* = -\tau\Pi(\tau\psi)$. Thus the LAN gradient for $\chi(r, f, M) = E[h(\varepsilon)]$ is $(u^*, v^*, w^*) = (-\tau\Pi(\tau\psi), \overline{v}, \overline{w})$. For this functional, the function ψ can be expressed as

$$\psi(X) = E(h(\varepsilon)L_V(X, \varepsilon) \mid X) = E(h(\varepsilon)L(X, \varepsilon) \mid X) - \frac{E(\varepsilon h(\varepsilon) \mid X)}{\tau^2(X)}. \tag{2.6}$$

Two models for the covariate distribution M are of interest: completely known covariate distribution, in which case $W = \{0\}$; and completely unknown covariate distribution (up to regularity), in which case $W = L_{2,0}(M)$. In the first case, $\overline{w} = 0$. In the second case,

$$\overline{w}(X) = \tilde{h}(X) - E[\tilde{h}(X)] = E(h(\varepsilon) \mid X) - E[h(\varepsilon)].$$

From now on we restrict ourselves to the second case. In this case, the influence function for $\chi(r, f, M) = E[h(\varepsilon)]$ is

$$h(\varepsilon) - E[h(\varepsilon)] - \frac{E(\varepsilon h(\varepsilon) \mid X) - u^*(X)}{\tau^2(X)}\, \varepsilon.$$

We are interested in two models for the regression function, namely nonparametric regression and nonlinear regression. In the *nonparametric regression model*, r is unspecified up to smoothness, so that the regression functions are dense in $L_2(M)$. Hence $u^* = -\tau\psi$. In view of (2.6), the efficient influence function simplifies to

$$t_{np} = h(\varepsilon) - E[h(\varepsilon)] - E(h(\varepsilon)L(X, \varepsilon) \mid X)\, \varepsilon.$$

The subscript np stands for 'nonparametric'.

If X and ε happen to be independent, then $f(x, z) = g(z)$ and $L(x, z) = \ell(z) = -g'(z)/g(z)$, so that the influence function simplifies further to

$$h(\varepsilon) - E[h(\varepsilon)] - E[h(\varepsilon)\ell(\varepsilon)]\, \varepsilon.$$

This coincides with the efficient influence function in the smaller model I where X and ε are known to be independent, which was obtained in MSW. This means that an efficient estimator for $E[h(\varepsilon)]$ in model II remains efficient in model I. We show later that, conversely, the efficient estimator constructed there for model I remains efficient in model II.

The second model is the *nonlinear regression model*, in which $r = r_\vartheta$ depends on a finite-dimensional parameter ϑ.

Assumption A4. The function $r_{\vartheta+t}$ is differentiable at $t = 0$ in $L_2(M)$,

$$\int (r_{\vartheta+t} - r_\vartheta - t^\top \dot{r}_\vartheta)^2 \, dM = o(\|t\|^2),$$

and $R_\vartheta = \int \dot{r}_\vartheta \dot{r}_\vartheta^\top \, dM$ is positive definite.

Then $U = [\dot{r}_\vartheta]$, the span of the components of $\dot{r}_\vartheta$. Now

$$u^* = -\tau\Pi(\tau\psi) = -\tilde{c}_\vartheta^\top \tilde{R}_\vartheta^{-1} \dot{r}_\vartheta, \quad \text{with} \quad \tilde{c}_\vartheta = \int \psi \dot{r}_\vartheta \, dM, \quad \tilde{R}_\vartheta = \int \frac{r_\vartheta \dot{r}_\vartheta^\top}{\tau^2} \, dM.$$

By the definition of ψ, we have

$$\tilde{c}_\vartheta = E[h(\varepsilon) L_V(X, \varepsilon) \dot{r}_\vartheta(X)] = E[h_V(X, \varepsilon) L(X, \varepsilon) \dot{r}_\vartheta(X)].$$

Hence the efficient influence function for $E[h(\varepsilon)]$ in the nonlinear model is

$$t_{nl}(X, Y) = h(\varepsilon) - E[h(\varepsilon)] - \frac{E(\varepsilon h(\varepsilon) \mid X)}{\tau^2(X)} \, \varepsilon - \tilde{c}_\vartheta^\top t_{par}(X, Y), \tag{2.7}$$

where

$$t_{par}(X, Y) = \frac{\tilde{R}_\vartheta^{-1} \dot{r}_\vartheta(X)}{\tau^2(X)} \, \varepsilon \tag{2.8}$$

is the influence function of the weighted least squares estimator of ϑ.

We show now that t_{par} is also the efficient influence function for ϑ, understood componentwise. The functional for the j-th component of ϑ is $\chi_j(r_\vartheta, f, M) = \vartheta_j = e_j^\top \vartheta$. The natural gradient is $(u_j, 0, 0)$ with $u_j = R_\vartheta^{-1} \dot{r}_\vartheta$. Hence the LAN gradient is $(u_j^*, 0, 0)$ with $u_j^* = e_j^\top \tilde{R}_\vartheta^{-1} \dot{r}_\vartheta$.

3. Efficient influence functions for type I models

Consider the regression model $Y = r(X) + \varepsilon$, where X and ε are independent and ε has mean zero and finite variance $\sigma^2 = E[\varepsilon^2]$, and density g. Let G denote the corresponding distribution function. This is a submodel of the type II model considered in Section 2. Parallel to Section 2, we calculate efficient influence functions of arbitrary real-valued functionals of (r, g, M) for general type I models, and specialize the result to functionals $E[h(\varepsilon)] = \int h(z) g(z) \, dz$, and to nonparametric and nonlinear regression. The joint distribution Q of (X, ε) now factors as $Q(dx, dz) = M(dx) g(z) \, dz$. Assumptions A1 to A3 are replaced by the following assumption.

Assumption A5. The density g has finite Fisher information for location: It is absolutely continuous with a.e. derivative g' fulfilling

$$J = \int \ell(z)^2 g(z)\, dz < \infty, \quad \text{with } \ell = -g'/g.$$

Assumptions A1 to A3 are then satisfied with $f(x,z) = g(z)$. We take perturbations of r and M as before, and replace V by the following unconditional version:

$$V_0 = \{v \in L_2(G) : \int v(z)\, G(dz) = 0, \int zv(z)\, G(dz) = 0\}.$$

For $v \in V_0$ we consider a perturbation g_{nv} such that

$$\int \Big(g_{nv}^{1/2}(z) - g^{1/2}(z) - \frac{1}{2}\, n^{-1/2} v(z) f^{1/2}(z)\Big)^2 dz = o(n^{-1}).$$

This is a special case of (2.2). Then P_{nuvw} has tangent

$$t_{uvw}(x,y) = u(x)\ell(y - r(x)) + v(y - r(x)) + w(x)$$

in the sense of (2.3). Thus we have local asymptotic normality (2.4). We rewrite the tangent t_{uvw} as

$$t_{uvw}(X,Y) = u(X)\ell(\varepsilon) + v(\varepsilon) + w(X).$$

Now let χ be a real-valued functional of (r,g,M) which is differentiable at (r,g,M) with natural gradient $(\overline{u}, \overline{v}_0, \overline{w}) \in U \times V_0 \times W$ in the sense that for all $(u,v,w) \in U \times V_0 \times W$,

$$n^{1/2}(\chi(r_{nu}, g_{nv}, M_{nw}) - \chi(r,g,M)) \to \int \overline{u} u\, dM + \int \overline{v}_0 v\, dG + \int \overline{w} w\, dM.$$

To describe the efficient influence function of χ, we need additional notation. We let ℓ_0 denote the projection of ℓ onto V_0, and J_0 its variance. It is easily seen that $\ell_0(\varepsilon) = \ell(\varepsilon) - \varepsilon/\sigma^2$ and $J_0 = J - 1/\sigma^2$. We let π denote the projection of the constant function 1 onto U. Finally let

$$\beta = \frac{E[\overline{v}_0(\varepsilon)\ell(\varepsilon)] - E[\bar{u}(X)]J_0/J}{J - J_0 E[\pi(X)]}.$$

The efficient influence function is now

$$t_0^*(X,Y) = \Big(\frac{\overline{u}(X)}{J} - \beta\pi(X)\Big)\ell(\varepsilon) + \overline{v}_0(\varepsilon) - \frac{E[\overline{u}(X)]}{J}\,\ell_0(\varepsilon) + \beta E[\pi(X)]\ell_0(\varepsilon) + \overline{w}(X).$$

Indeed, one verifies that for all $(u,v,w) \in U \times V_0 \times W$,

$$\int t_0^* t_{uvw}\, dP = \int \overline{u} u\, dM + \int \overline{v}_0 v\, dG + \int \overline{w} w\, dM.$$

To check this, use $E[u(X)\pi(X)] = E[u(X)]$ and $E[v(\varepsilon)\ell_0(\varepsilon)] = E[v(\varepsilon)\ell(\varepsilon)]$ to get

$$\begin{aligned}
E[t_0^*(X,Y)u(X)\ell(\varepsilon)] &= E[\overline{u}(X)u(X)] - E[u(X)]\beta J + E[u(X)]E[\overline{v}_0(\varepsilon)\ell(\varepsilon] \\
&\quad -E[u(X)]E[\overline{u}(X)]\frac{J_0}{J} + E[u(X)]\beta E[\pi(X)]J_0 \\
&= E[\overline{u}(X)u(X)], \\
E[t_0^*(X,Y)v(\varepsilon)] &= \Big(\frac{E[\overline{u}(X)]}{J} - \beta E[\pi(X)]\Big)E[v(\varepsilon)\ell(\varepsilon)] + E[\overline{v}_0(\varepsilon)v(\varepsilon)] \\
&\quad -\frac{E[\overline{u}(X)]}{J}\,E[v(\varepsilon)\ell(\varepsilon)] + \beta E[\pi(X)]E[v(\varepsilon)\ell(\varepsilon)] \\
&= E[\overline{v}_0(\varepsilon)v(\varepsilon)], \\
E[t_0^*(X,Y)w(X)] &= E[\overline{w}(X)w(X)].
\end{aligned}$$

We are interested in linear functionals of the error distribution,

$$\chi(r,g,M) = \int h(z)g(z)\,dz = E[h(\varepsilon)].$$

The natural gradient of such a functional is $(0, h_0, 0)$ with h_0 the projection of h onto V_0, which is

$$h_0(\varepsilon) = h(\varepsilon) - E[h(\varepsilon)] - \frac{E[\varepsilon h(\varepsilon)]}{\sigma^2}\,\varepsilon.$$

The efficient influence function for $E[h(\varepsilon)]$ now becomes

$$t_0^*(X,Y) = h_0(\varepsilon) - \beta_0\big(\pi(X) - E[\pi(X)]\big)\ell(\varepsilon) - \beta_0\,E[\pi(X)]\,\frac{\varepsilon}{\sigma^2}$$

with

$$\beta_0 = \frac{E[h_0(\varepsilon)\ell(\varepsilon)]}{J - J_0E[\pi(X)]}.$$

In the *nonparametric regression model*, $U = L_2(M)$ and $\pi = 1$, so that

$$t_{0,np} = h_0(\varepsilon) - E[h_0(\varepsilon)\ell(\varepsilon)]\,\varepsilon = h(\varepsilon) - E[h(\varepsilon)] - E[h(\varepsilon)\ell(\varepsilon)]\,\varepsilon.$$

This result was already obtained in MSW.

In the *nonlinear regression model* we have $r = r_\vartheta$. Under Assumption A4, $U = [\dot{r}_\vartheta]$, and $\pi = a_\vartheta^\top R_\vartheta^{-1}\dot{r}_\vartheta$ with $a_\vartheta = E[r_\vartheta(X)]$. In this case the efficient influence function for $E[h(\varepsilon)]$ becomes

$$t_{0,nl}(X,Y) = h_0(\varepsilon) - c_\vartheta^\top\big((J - J_0a_\vartheta^\top R_\vartheta^{-1}a_\vartheta)R_\vartheta\big)^{-1}S_\vartheta(X,Y),$$

where

$$\begin{aligned}
c_\vartheta &= E[h_0(\varepsilon)\ell(\varepsilon)]a_\vartheta, \\
S_\vartheta(X,Y) &= (\dot{r}_\vartheta(X) - a_\vartheta)\ell(\varepsilon) + a_\vartheta\frac{\varepsilon}{\sigma^2} = \dot{r}_\vartheta(X)\ell(\varepsilon) - a_\vartheta\ell_0(\varepsilon).
\end{aligned}$$

The covariance matrix

$$\Lambda_\vartheta = JR_\vartheta - J_0a_\vartheta a_\vartheta^\top$$

of $S_\vartheta(X,Y)$ has inverse

$$\Lambda_\vartheta^{-1} = (JR_\vartheta)^{-1} + \frac{J_0}{J - J_0 a_\vartheta^\top R_\vartheta^{-1} a_\vartheta} R_\vartheta^{-1} a_\vartheta a_\vartheta^\top (JR_\vartheta)^{-1}.$$

It is now easily checked that

$$c_\vartheta^\top \Lambda_\vartheta^{-1} = \big(J - J_0 a_\vartheta^\top R_\vartheta^{-1} a_\vartheta\big)^{-1} c_\vartheta^\top R_\vartheta^{-1},$$

yielding

$$t_{0,nl}(X,Y) = h_0(\varepsilon) - c_\vartheta^\top \Lambda_\vartheta^{-1} S_\vartheta(X,Y).$$

As in Schick (1993) one checks that $\Lambda_\vartheta^{-1} S_\vartheta(X,Y)$ is the efficient influence function for ϑ. Indeed, the functional for the j-th component of ϑ is $\chi_j(r_\vartheta, g, M) = \vartheta_j = e_j^\top \vartheta$. The natural gradient is $(\overline{u}_j, 0, 0)$ with $\overline{u}_j = e_j^\top R_\vartheta^{-1} \dot r_\vartheta$. Then the efficient influence function of ϑ_j is $b_j(X)\ell(\varepsilon) - E[b_j(X)]\ell_0(\varepsilon)$, where

$$b_j = \frac{\bar u_j}{J} + \frac{E[\bar u_j(X)] J_0/J}{J - J_0 a_\vartheta^\top R_\vartheta^{-1} a_\vartheta} a_\vartheta^\top R_\vartheta^{-1} \dot r_\vartheta = e_j^\top \Lambda_\vartheta^{-1} \dot r_\vartheta.$$

This shows that $e_j^\top \Lambda_\vartheta^{-1} S_\vartheta(X,Y)$ is the efficient influence function for ϑ_j. Similar results for nonlinear autoregression are obtained in Schick and Wefelmeyer (2002).

4. Constructions of estimators for nonparametric regression

Let $(X_1,Y_1),\dots,(X_n,Y_n)$ be independent observations from the nonparametric regression model $Y = r(X) + \varepsilon$ of type II. For simplicity we take X one-dimensional. We want to construct an efficient estimator for $E[h(\varepsilon)]$. We consider the empirical estimator

$$\hat H = \frac{1}{n} \sum_{i=1}^n h(\hat\varepsilon_i)$$

based on residuals $\hat\varepsilon_i = Y_i - \hat r_i$, where $\hat r_i$ estimates $r(X_i)$. To be specific, we take the leave-one-out local polynomial smoother of degree L, which is the first component $\hat r_i = \hat\beta_{i0}$ of the vector $(\hat\beta_{i0},\dots,\hat\beta_{iL})$ that minimizes

$$\sum_{j:j\neq i} \left(Y_j - \sum_{\lambda=0}^L \beta_\lambda \Big(\frac{X_j - X_i}{b_n}\Big)^\lambda \right)^2 K\Big(\frac{X_j - X_i}{b_n}\Big).$$

Here K is a symmetric and bounded density with support $[-1,1]$, and b_n is a bandwidth. If $L = 0$, then $\hat r_i$ is the usual leave-one-out kernel estimator:

$$\hat r_i = \frac{\sum_{j:j\neq i} Y_j K_{b_n}(X_i - X_j)}{\sum_{j:j\neq i} K_{b_n}(X_i - X_j)} \quad \text{with} \quad K_{b_n}(x) = \frac{1}{b_n} K\Big(\frac{x}{b_n}\Big).$$

MSW use the leave-one-out local polynomial smoother to obtain an i.i.d. representation of $\hat{H}$ in model I. More precisely, they require that h, M and r are smooth in the following sense.

Assumption B1. There are positive numbers $\alpha \leq 1$, c, C_1, C_2 such that

$$\int (h(z+w+v) - h(z+w))^2 g(z)\, dz \leq C_1 |v|^{1+\alpha}, \quad |v|, |w| \leq c,$$

and $\mu(v) = \int h(z-v) g(z)\, dz$ is differentiable at $v = 0$ with

$$|\mu(v) - \mu(0) - \mu'(0)v| \leq C_2 |v|^{1+\alpha}, \quad |v| \leq c.$$

Assumption B2. The covariate distribution M has compact support $[0,1]$ and admits a density that is continuous and positive on its support.

Assumption B3. The regression function r is L times differentiable, and its L-th derivative is Hölder with positive exponent β.

Let α_* be the smaller of $1/3$ and $\alpha/(1+\alpha)$. Under the above assumptions, with $L + \beta > 1/(2\alpha_*)$, MSW obtain the following i.i.d. representation:

$$\hat{H} = \frac{1}{n} \sum_{i=1}^{n} \big(h(\varepsilon_i) + \mu'(0)\varepsilon_i \big) + o_p(n^{-1/2})$$

if the bandwidth fulfills

$$n^{1/2} b_n^{L+\beta} \to 0 \quad \text{and} \quad b_n n^{\alpha_*} \to \infty. \tag{4.1}$$

The condition $L + \beta > 1/(2\alpha_*)$ is needed to guarantee the existence of such a bandwidth. Indeed, we can pick b_n proportional to $n^{-\psi}$ with ψ in the open interval $(1/(2L+2\beta), \alpha_*)$. Since $L + \beta$ is a measure of smoothness of r, the condition $L + \beta > 1/(2\alpha_*)$ demands a certain amount of smoothness of the regression function. For example, if $\alpha \geq 1/2$, then $L + \beta$ must be larger than $3/2$.

A similar i.i.d. representation of $\hat{H}$ holds in the larger model II under the following modified smoothness condition on h.

Assumption C1. There are positive numbers $\alpha \leq 1$, c, C_1, C_2 such that, for each x in $[0,1]$,

$$\int (h(z+w+v) - h(z+w))^2 f(x,z)\, dz \leq C_1 |v|^{1+\alpha}, \quad |v|, |w| \leq c,$$

and $v \mapsto \mu(x,v) = \int h(z-v) f(x,z)\, dz$ is differentiable at $v = 0$ with

$$|\mu(x,v) - \mu(x,0) - \mu'(x,0)v| \leq C_2 |v|^{1+\alpha}, \quad |v| \leq c.$$

Note that Assumption C1 is equivalent to Assumption B1 if $f(x,z) = g(z)$. An appropriate modification of the arguments in MSW now gives the following result in the larger model II.

Theorem 1. *Suppose Assumptions A1, C1, B2 and B3 hold with $L + \beta > 1/(2\alpha_*)$, and b_n fulfills (4.1). Then*

$$\hat{H} = \frac{1}{n}\sum_{i=1}^{n}\big(h(\varepsilon_i) + \mu'(X_i, 0)\varepsilon_i\big) + o_p(n^{-1/2}).$$

Proof. Let $\Delta_i = \hat{r}_i - r(X_i)$. It suffices to show

$$n^{-1/2}\sum_{i=1}^{n}\big(h(\varepsilon_i - \Delta_i) - h(\varepsilon_i) - \mu(X_i, \Delta_i) + \mu(X_i, 0)\big) = o_p(1), \tag{4.2}$$

$$n^{-1/2}\sum_{i=1}^{n}\big(\mu(X_i, \Delta_i) - \mu(X_i, 0) - \mu'(X_i, 0)\Delta_i\big) = o_p(1), \tag{4.3}$$

$$n^{-1/2}\sum_{i=1}^{n}\mu'(X_i, 0)(\Delta_i - \varepsilon_i) = o_p(1). \tag{4.4}$$

Note that the conditional variance τ^2 is bounded by Assumption A1. Thus the conditions (2.3) to (2.6) of MSW remain true in model II. The first two statements, (4.2) and (4.3), are then proved as in Theorem 1 of MSW. It remains to prove (4.4). We prove it for $\mu'(\cdot, 0)$ replaced by an arbitrary function a in $L_2(M)$. This means we prove (1.1). As shown in MSW, the estimator $\hat{r}_i$ is a linear smoother $\hat{r}_i = \sum_{j=1}^{n} A_{ij}Y_j$ with weights $A_{ii} = 0$ and

$$A_{ij} = \frac{1}{n-1}\sum_{\lambda=0}^{L} q_{i\lambda}\frac{(X_j - X_i)^\lambda}{b_n^{\lambda+1}}K\Big(\frac{X_j - X_i}{b_n}\Big), \quad i \neq j,$$

where the $q_{i\lambda}$ are functions of $X_1, \ldots, X_n$ only, with $\max_\lambda \max_i |q_{i\lambda}| = O_p(1)$. MSW also show that

$$\max_{i=1,\ldots,n}\Big|\sum_{j=1}^{n} A_{ij}r(X_j) - r(X_i)\Big| = o_p(n^{-1/2}). \tag{4.5}$$

Thus we can write

$$n^{-1/2}\sum_{i=1}^{n} a(X_i)(\Delta_i - \varepsilon_i) = n^{-1/2}\sum_{i=1}^{n} a(X_i)\Big(\sum_{j=1}^{n} A_{ij}\varepsilon_j - \varepsilon_i\Big) + R,$$

where

$$R = n^{-1/2}\sum_{i=1}^{n} a(X_i)\Big(\sum_{j=1}^{n} A_{ij}r(X_j) - r(X_i)\Big) = o_p(1).$$

Now write

$$n^{-1/2}\sum_{i=1}^{n} a(X_i)\Big(\sum_{j=1}^{n} A_{ij}\varepsilon_j - \varepsilon_i\Big) = n^{-1/2}\sum_{j=1}^{n}\varepsilon_j\sum_{i=1}^{n}\big(a(X_i)A_{ij} - a(X_j)\big).$$

Its conditional second moment given $X_1, \ldots, X_n$ is bounded by $C_\tau L(a)$, where C_τ is the bound on τ^2 and

$$L(a) = \frac{1}{n}\sum_{j=1}^{n}\Big(a(X_j) - \sum_{i:i\neq j} a(X_i)A_{ij}\Big)^2.$$

It remains to show that $L(a) = O_p(1)$. The special case $a = 1$ was already obtained in MSW. Note that

$$(n-1)|A_{ij}| \leq QK_{b_n}(X_j - X_i),$$
$$S = \max_{j=1,\ldots,n}\frac{1}{n-1}\sum_{i:i\neq j} K_{b_n}(X_j - X_i) = O_p(1).$$

This shows that

$$\begin{aligned} L(a) &\leq \frac{1}{n}\sum_{j=1}^{n}\Big(2a(X_j)^2 + 2Q^2\Big(\frac{1}{n-1}\sum_{j=1}^{n}|a(X_i)|K_{b_n}(X_j - X_i)\Big)^2\Big) \\ &\leq \frac{1}{n}\sum_{j=1}^{n}\Big(2a(X_j)^2 + 2Q^2 S\sum_{i:i\neq j} a(X_i)^2 K_{b_n}(X_j - X_i)\Big) \\ &\leq (2 + 2Q^2S^2)\frac{1}{n}\sum_{j=1}^{n} a(X_j)^2. \end{aligned}$$

Moreover, for a uniformly continuous $\overline{a}$ in $L_2(M)$,

$$\begin{aligned} L(a) &\leq 2L(\overline{a}) + 2L(a-\overline{a}) \leq 2L(\overline{a}) + (4 + 4Q^2S^2)\frac{1}{n}\sum_{j=1}^{n}(a(X_j) - \overline{a}(X_j))^2, \\ L(\overline{a}) &\leq \frac{1}{n}\sum_{j=1}^{n}\Big(2\overline{a}(X_j)^2\Big(1 - \sum_{i:i\neq j} A_{ij}\Big)^2 + 2\Big(\sum_{i:i\neq j}(\overline{a}(X_i) - \overline{a}(X_j))A_{ij}\Big)^2\Big) \\ &\leq \sup_{x\in[0,1]}\overline{a}(x)^2 L(1) + 2Q^2S^2 \sup_{|x-y|\leq b_n}(\overline{a}(x) - \overline{a}(y))^2. \end{aligned}$$

Since $L(1) = o_p(1)$ as shown in MSW, we obtain $L(\overline{a}) = o_p(1)$ and thus

$$L(a) \leq (4 + 4Q^2S^2)\frac{1}{n}\sum_{j=1}^{n}(a(X_j) - \overline{a}(X_j))^2 + o_p(1).$$

Since the uniformly continuous functions are dense in $L_2(M)$, the result follows. □

We can weaken the smoothness assumptions on h at the expense of smoothness assumptions on M and f. This is necessary if we want to estimate the distribution function of ε. An i.i.d. representation for the empirical distribution function based on residuals from a related heteroscedastic model $Y = r(X) + \tau(X)\varepsilon$ with X and ε *independent* is obtained by Akritas and Van Keilegom (2001). They use empirical process theory to establish their result. We show

that their approach can be adapted to estimating expectations $E[h(\varepsilon)]$ in our type II model $Y = r(X) + \varepsilon$, which includes their model as a submodel. Note that the two expectations are different: in our model, their functional would be written $E[h(\varepsilon/\tau(X))]$.

We give a version of their approach in our model. For this we assume that $\hat{r}_i$ is now of the form $\hat{r}(X_i)$ for some estimator $\hat{r}$ of r. We assume also that h is nondecreasing. Our results then apply to functions that can be written as differences of nondecreasing functions. We need some assumptions on $\hat{r}$ and h. Let D be a nonnegative function in $L_2(M)$, and let $\mathcal{D}_0$ be a subset of $\mathcal{D} = \{a \in L_2(M) : |a| \le D\}$.

Assumption D1. There is a finite constant C_D such that

$$\iint \big(h(z + b(x)) - h(z + a(x))\big)^2 f(x, z)\, dz M(dx) \le C_D \Big(\int (b - a)^2\, dM \Big)^{1/2}, \quad a, b \in \mathcal{D}_0.$$

This assumption requires less smoothness of h than the first part of Assumption C1. Consider the function $h(x) = 1(x > t)$, for which we get $E[h(\varepsilon)] = 1 - G(t)$, the survival function of ε at t. Then the first condition of Assumption 1 is not fulfilled, because it requires an exponent greater than $1/2$. Assumption D1 holds under Assumption A3. Indeed, the latter implies that

$$f(x, z)^2 \le J(x) = \int L(x, u)^2 f(x, u)\, du. \tag{4.6}$$

Hence the left-hand side in Assumption D1 is bounded by

$$\int \int_{t - a(x) \vee b(x)}^{t - a(x) \wedge b(x)} f(x, z)\, dz\, M(dx) \le \int J(x)^{1/2} |b(x) - a(x)|\, M(dx),$$

and Assumption D1 follows by an application of the Cauchy–Schwarz inequality.

Recall that the *bracketing number* $N_{[\,]}(\eta, \mathcal{D}_0, L_2(M))$ is the smallest integer N for which there are $a_1^L \le a_1^U, \ldots, a_N^L \le a_N^U$ such that $\int (a_j^U - a_j^L)^2\, dM \le \eta^2$, and each $a \in \mathcal{D}_0$ belongs to one bracket, i.e. $a_j^L \le a \le a_j^U$ for some j.

Theorem 2. *Suppose h is nondecreasing and satisfies Assumption D1. Suppose that the bracketing number of $\mathcal{D}_0$ satisfies*

$$\int_0^\infty \sqrt{\log N_{[\,]}(\eta^2, \mathcal{D}_0, L_2(M))}\, d\eta < \infty. \tag{4.7}$$

Suppose $\hat{r}$ satisfies $P(\hat{r} - r \in \mathcal{D}_0) \to 1$ and $\int (\hat{r} - r)^2\, dM = o_p(1)$. Then

$$\hat{H} = \frac{1}{n} \sum_{i=1}^n \big(h(\varepsilon_i) + \mu(X_i, \hat{r}(X_i) - r(X_i)) - \mu(X_i, 0)\big) + o_p(n^{-1/2}),$$

with $\mu(x, s) = \int h(z - s) f(x, z)\, dz$.

Proof. The result follows if we show that the class $\mathcal{F} = \{\psi_a : a \in \mathcal{D}_0\}$ is Donsker, where

$$\psi_a(\varepsilon, X) = h(\varepsilon - a(X)) - h(\varepsilon) - \mu(X, a(X)) + \mu(X, 0).$$

Let $\mathcal{F}_1 = \{h_a : a \in \mathcal{D}_0\}$, with $h_a(\varepsilon, X) = h(\varepsilon - a(X))$. We show that $\mathcal{F}_1$ is Donsker. This implies that $\mathcal{F}_2 = \{\mu_a : a \in \mathcal{D}_0\}$ is Donsker, where $\mu_a(X) = E(h_a(\varepsilon, X) \mid X) = \mu(X, a(X))$. Since sums of Donsker classes are Donsker, see e.g. van der Vaart and Wellner (1996, Example 2.10,7), it follows that $\mathcal{F}$ is Donsker. We show that

$$N_{[\,]}(\eta, \mathcal{F}_1, L_2(Q)) \le N_{[\,]}(\eta^2, \mathcal{D}_0, L_2(M)). \tag{4.8}$$

The desired result then follows from the integrability condition (4.7) and Ossiander (1987), see also van der Vaart and Wellner (1996, Theorem 2.5.6). Since h is nondecreasing, we have $h(\varepsilon - a^U(x)) \le h(\varepsilon - a(x)) \le h(\varepsilon - a^L(x))$ whenever $a^L \le a \le a^U$. Let $a_1^L \le a_1^U, \ldots, a_N^L \le a_N^U$ be brackets for $\mathcal{D}_0$ with $\int (a_j^U - a_j^L)^2 \, dM \le \eta^4$. Then $h_{a_1^U} \le h_{a_1^L}, \ldots, h_{a_N^U} \le h_{a_N^L}$ are brackets for $\mathcal{F}_1$ with

$$\int (h_{a_j^U} - h_{a_j^L})^2 \, dQ \le C_1 \Big(\int (a_j^U - a_j^L)^2 \, dM \Big)^{1/2} \le \eta^2.$$

This establishes (4.8) and completes the proof. □

Under the assumptions of Theorem 2, to get the desired i.i.d. representation for $\hat{H}$, one needs to show that

$$\frac{1}{n} \sum_{i=1}^{n} \big(\mu(X_i, \hat{r}(X_i) - r(X_i)) - \mu(X_i, 0) - \mu'(X_i, 0)\varepsilon_i \big) = o_p(n^{-1/2}).$$

This is implied by

$$\frac{1}{n} \sum_{i=1}^{n} \big(\mu(X_i, \hat{r}(X_i) - r(X_i)) - \mu(X_i, 0) - \mu'(X_i, 0)(\hat{r}(X_i) - r(X_i)) \big) = o_p(n^{-1/2}),$$

$$\frac{1}{n} \sum_{i=1}^{n} \mu'(X_i, 0)(\hat{r}(X_i) - r(X_i)) = \frac{1}{n} \sum_{i=1}^{n} \mu'(X_i, 0)\varepsilon_i + o_p(n^{-1/2}).$$

The latter was already established for (undersmoothed) local polynomial smoothers in Theorem 1. A sufficient condition for the former is that there are a positive number α_B and a nonnegative function B such that

$$\big| \mu(x, a(x)) - \mu(x, 0) - \mu'(x, 0)a(x) \big| \le B(x)|a(x)|^{1+\alpha_B}, \quad a \in \mathcal{D}_0, \tag{4.9}$$

$$\frac{1}{n} \sum_{i=1}^{n} B(X_i) \big| \hat{r}(X_i) - r(X_i) \big|^{1+\alpha_B} = o_p(n^{-1/2}). \tag{4.10}$$

Note that if B is in $L_1(M)$, then (4.10) follows from

$$\max_{i=1,\ldots,n} |\hat{r}(X_i) - r(X_i)|^{1+\alpha_B} = o_p(n^{-1/2}).$$

For $h(x) = 1(x > t)$, Assumption A3 yields (4.9) with $\mu'(x,0) = f(x,t)$, $B(x) = J(x)$, and $\alpha_B = 1/2$. Indeed, we find with the help of (4.6) that, for $t < z$,

$$|f(x,z) - f(x,t)| = \Big| \int_t^z L(x,s) f(x,s)\, ds \Big| \le J(x)^{1/2} \Big(\int_t^z f(x,s)\, ds \Big)^{1/2} \le J(x)(z-t)^{1/2}.$$

Thus for positive $a(x)$ the left-hand side of (4.9) is bounded by

$$\int_t^{t+a(x)} |f(x,z) - f(x,t)|\, dz \le J(x) a(x)^{3/2}.$$

Similarly for $a(x)$ negative.

Corollary. *Let Assumption A3 hold. Suppose that $\hat{r}$ satisfies $P(\hat{r} - r \in \mathcal{D}_0) \to 1$,*

$$\sup_x |\hat{r}(x) - r(x)| = o_p(n^{-1/3})$$

and, for all a in $L_2(M)$,

$$\frac{1}{n} \sum_{i=1}^n a(X_i)(\hat{r}(X_i) - r(X_i)) = \frac{1}{n} \sum_{i=1}^n a(X_i)\varepsilon_i + o_p(n^{-1/2}).$$

Then, for each t in $\mathbf{R}$,

$$\frac{1}{n} \sum_{i=1}^n 1(\hat{\varepsilon}_i > t) = \frac{1}{n} \sum_{i=1}^n \big(1(\varepsilon_i > t) - f(X_i)\varepsilon_i \big) + o_p(n^{-1/2}).$$

It would be interesting to find spaces $\mathcal{D}_0$ appropriate for existing nonparametric regression estimators. One such result is in Akritas and Van Keilegom (2001).

5. Constructions of estimators for nonlinear regression

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent observations from the nonlinear regression model $Y = r_\vartheta(X) + \varepsilon$ of type II, with d-dimensional parameter ϑ. Throughout the section, we suppose that Assumptions A1 and A2 hold. We want to construct an efficient estimator for $E[h(\varepsilon)]$. As in Section 4, a natural estimator is the empirical estimator $\hat{H} = \frac{1}{n} \sum_{i=1}^n h(\hat{\varepsilon}_i)$, now however based on residuals $\hat{\varepsilon}_i = Y_i - r_{\hat{\vartheta}}(X_i)$, where $\hat{\vartheta}$ is an estimator of ϑ.

The estimator $\hat{H}$ does not use the information that the errors are conditionally centered, $E(\varepsilon \mid X) = 0$. The form of the efficient influence function (2.7) for the nonlinear type II model suggests to correct $\hat{H}$ as

$$\hat{H}^* = \hat{H} - \frac{1}{n} \sum_{i=1}^n \frac{\hat{\rho}_i}{\hat{\tau}_i^2} \hat{\varepsilon}$$

with estimators $\hat{\tau}_i^2$ of $\tau^2(X_i)$ and $\hat{\rho}_i$ of $\rho(X_i)$, where $\rho(X) = E(\varepsilon h(\varepsilon) \mid X)$. In Theorem 3 below we give a stochastic expansion of $\hat{H}$, while Theorem 4 gives one for the correction term. Together, those results give the expansion

$$\hat{H}^* = \frac{1}{n}\sum_{i=1}^{n}\Big(h(\varepsilon_i) - \frac{\rho(X_i)}{\tau^2(X_i)}\varepsilon_i\Big) + \tilde{a}_\vartheta^\top(\hat{\vartheta} - \vartheta) + o_p(n^{-1/2}),$$

where

$$\tilde{a}_\vartheta = \int \Big(\mu'(x,0) + \frac{\rho(x)}{\tau^2(x)}\Big)\dot{r}_\vartheta(x)\, M(dx).$$

Under Assumption A3 and the assumptions below,

$$\mu'(x,0) = -\int h(z)L(x,z)f(x,z)\,dz \quad \text{for } M\text{-a.a. } x$$

and thus $\tilde{a}_\vartheta = \tilde{c}_\vartheta$. As shown in Section 2, an efficient estimator $\hat{\vartheta}$ of ϑ has influence function (2.8). For such an estimator, $\hat{H}^*$ has influence function (2.7) and is therefore efficient for $E[h(\varepsilon)]$.

We now analyze $\hat{H}$ and the correction term under the assumption that $\hat{\vartheta}$ is $n^{1/2}$-consistent. In most cases of interest, h can be written as a difference of two nondecreasing functions. Then it suffices to consider the case of nondecreasing h. We also need assumptions on the regression function.

Assumption R. The function $t \mapsto r_{\vartheta+t}(x)$ is continuously differentiable for M-a.a. x with derivative $\dot{r}_{\vartheta+t}(x)$, the matrix

$$R_\vartheta = \int \dot{r}_\vartheta \dot{r}_\vartheta^\top\, dM$$

is positive definite, and

$$\sup_{\|t\|\le\delta} \|\dot{r}_{\vartheta+t} - \dot{r}_\vartheta\| \to 0 \quad \text{in } L_2(M) \text{ as } \delta \to 0.$$

This assumption implies $L_2(M)$-differentiability of $t \mapsto r_{\vartheta+t}$ at $t = 0$. This follows from the representation

$$r_{\vartheta+t}(x) - r_\vartheta(x) - t^\top \dot{r}_\vartheta(x) = t^\top \int_0^1 (\dot{r}_{\vartheta+vt}(x) - \dot{r}_\vartheta(x))\,dv.$$

Thus Assumption R implies Assumption A4. The above representation also yields

(R1) $$|r_{\vartheta+t}(x) - r_{t+s}(x)| \le \|t - s\| A_\delta(x) \quad \text{for } \|s\|, \|t\| \le \delta,$$

where

$$A_\delta(x) = \sup_{\|t\|\le\delta} \|\dot{r}_{\vartheta+t}(x)\|.$$

Furthermore, for every constant C,

$$\text{(R2)} \qquad \sup_{\|t\|\le C} \sum_{i=1}^{n} \big(r_{\vartheta+n^{-1/2}t}(X_i) - r_\vartheta(X_i) - \dot r_\vartheta(X_i)^\top n^{-1/2} t\big)^2 = o_p(1).$$

Assumption H. Set $\Delta_t = r_{\vartheta+t} - r_\vartheta$. The function h is nondecreasing and satisfies the following conditions.

$$\text{(H1)} \qquad E[h(\varepsilon - \Delta_t(X)) - h(\varepsilon)]^2 \to 0 \quad \text{as } t \to 0.$$

There are positive constants α, c_H, C_H such that

$$\text{(H2)} \qquad E\big[h(\varepsilon - \Delta_t(X) + vA_\delta(X)) - h(\varepsilon - \Delta_t(X))\big]^2 \le C_H|v|^\alpha \quad \text{for } |\delta|, |v|, \|t\| \le c_H.$$

The functions $v \mapsto \mu(x,v) = \int h(z-v) f(x,z)\,dz$ are continuously differentiable for M-a.a. x, and their derivatives $\mu'(x,v)$ satisfy $E[\mu'(X,0)^2] < \infty$ and

$$\text{(H3)} \qquad E\Big[\sup_{|v|\le\delta} (\mu'(X,v) - \mu'(X,0))^2\Big] \to 0 \quad \text{as } \delta \to 0.$$

This assumption implies that

$$|\mu(x,w) - \mu(x,v)| \le |w-v| B_\delta(x) \quad \text{for } |v|, |w| \le \delta,$$

where

$$B_\delta(x) = \sup_{|w|\le\delta} |\mu'(x,w)|.$$

Theorem 3. *Suppose Assumptions R and H hold, and $\hat\vartheta$ is $n^{1/2}$-consistent for ϑ. Then*

$$\hat H = \frac{1}{n}\sum_{i=1}^{n} h(\varepsilon_i) + \int \mu'(x,0)\dot r_\vartheta(x)^\top M(dx)(\hat\vartheta - \vartheta) + o_p(n^{-1/2}).$$

Proof. In view of the law of large numbers, it suffices to show

$$\hat H = \frac{1}{n}\sum_{i=1}^{n} \big(h(\varepsilon_i) + \mu'(X_i,0)\dot r_\vartheta(X_i)^\top(\hat\vartheta - \vartheta)\big) + o_p(n^{-1/2}).$$

With $\Delta_{nt} = \Delta_{n^{-1/2}t} = r_{\vartheta+n^{-1/2}t} - r_\vartheta$ let

$$H_{nt} = n^{-1/2}\sum_{i=1}^{n} \big(h(\varepsilon_i - \Delta_{nt}(X_i)) - h(\varepsilon_i) - \mu(X_i, \Delta_{nt}(X_i)) + \mu(X_i,0)\big).$$

We first show that, for every constant C,

$$\sup_{\|t\|\le C} |H_{nt}| = o_p(1). \tag{5.1}$$

For this, fix C and an integer D. Let $t_j = jC/D$, $j = (j_1, \ldots, j_d) \in J = \{-D, \ldots, D\}^d$. We have

$$\sup_{\|t\|\le C} |H_{nt}| \le \max_{j\in J} |H_{nt_j}| + \max_{j\in J} \sup_{\|t-t_j\|\le d^{1/2}C/D} |H_{nt} - H_{nt_j}|.$$

For $\eta > 0$ we have by Assumption (H1),

$$\begin{aligned} P\Big(\max_{j\in J} |H_{nt_j}| \ge \eta\Big) &\le \sum_{j\in J} P(|H_{nt_j}| \ge \eta) \le \eta^{-2} \sum_{j\in J} E[H_{nt_j}^2] \\ &\le \eta^{-2} \sum_{j\in J} E[(h(H(\varepsilon - \Delta_{nt_j}(X)) - h(\varepsilon))^2] \to 0. \end{aligned} \tag{5.2}$$

For $s, t \in [-C, C]^d$ with $\|t - s\| \le d^{1/2}C/D$ we have $|H_{nt} - H_{ns}| \le K_{nst} + \overline{K}_{nst}$ with

$$\begin{aligned} K_{nst} &= \frac{1}{n}\sum_{i=1}^n n^{1/2} \Big| h((\varepsilon_i - \Delta_{nt}(X_i)) - h(\varepsilon_i - \Delta_{ns}(X_i)) \Big|, \\ \overline{K}_{nst} &= \frac{1}{n}\sum_{i=1}^n n^{1/2} \Big| \mu(X_i, \Delta_{nt}(X_i)) - \mu(X_i, \Delta_{ns}(X_i)) \Big|. \end{aligned}$$

We obtain from Assumption (R1) with $\delta = \delta_n = n^{-1/2} d^{1/2} C$ that

$$\overline{K}_{nst} \le d^{1/2} \frac{C}{D} U_n, \quad \text{where} \quad U_n = \frac{1}{n}\sum_{i=1}^n B_{\delta_n}(X_i) A_{\delta_n}(X_i),$$

and

$$\begin{aligned} K_{nst} &\le \frac{1}{n}\sum_{i=1}^n n^{1/2} \Big| h\Big(\varepsilon_i - \Delta_{ns}(X_i) + \frac{\delta_n}{D} A_{\delta_n}(X_i)\Big) - h\Big(\varepsilon_i - \Delta_{ns}(X_i) - \frac{\delta_n}{D} A_{\delta_n}(X_i)\Big) \Big| \\ &\le \frac{1}{n}\sum_{i=1}^n n^{1/2} \Big\{ h\Big(\varepsilon_i - \Delta_{ns}(X_i) + \frac{\delta_n}{D} A_{\delta_n}(X_i)\Big) - \mu\Big(X_i, \Delta_{ns}(X_i) + \frac{\delta_n}{D} A_{\delta_n}(X_i)\Big) \Big\} \\ &\quad - \frac{1}{n}\sum_{i=1}^n n^{1/2} \Big\{ h\Big(\varepsilon_i - \Delta_{ns}(X_i) - \frac{\delta_n}{D} A_{\delta_n}(X_i)\Big) + \mu\Big(X_i, \Delta_{ns}(X_i) - \frac{\delta_n}{D} A_{\delta_n}(X_i)\Big) \Big\} \\ &\quad + 2d^{1/2} \frac{C}{D} U_n. \end{aligned}$$

Hence $|H_{nt} - H_{ns}| \le T_{ns} + 3d^{1/2} C U_n / D$, where T_{ns} is the sum of the two centered averages on the right-hand side of the last display. By Assumption (H2) we have $E[T_{ns}^2] \le 4 C_H \delta_n / D$ and thus $\max_{j\in J} T_{nt_j} = o_p(1)$. Since $U_n \le K + o_p(1)$ for some K,

$$\max_{j\in J} \sup_{\|t-t_j\|\le d^{1/2}C/D} |H_{nt} - H_{nt_j}| = o_p(1) + d^{1/2} \frac{C}{D}(K + o_p(1)). \tag{5.3}$$

Since D is arbitrary, the desired (5.1) now follows from (5.2) and (5.3).

Next we show that

$$\sup_{\|t\|\le C}\Big|n^{-1/2}\sum_{i=1}^{n}\big(\mu(X_i,\Delta_{nt}(X_i))-\mu(X_i,0)-\mu'(X_i,0)\dot r_\vartheta(X_i)^\top n^{-1/2}t\big)\Big|=o_p(1). \tag{5.4}$$

Because of Assumption (R2), $E[\mu'(X,0)^2]<\infty$, and the Cauchy–Schwarz inequality,

$$\sup_{\|t\|\le C}\Big|n^{-1/2}\sum_{i=1}^{n}\mu'(X_i,0)\big(\Delta_{nt}(X_i)-\dot r_\vartheta(X_i)^\top n^{-1/2}t\big)\Big|=o_p(1).$$

Thus it suffices to show

$$\sup_{\|t\|\le C}\Big|n^{-1/2}\sum_{i=1}^{n}\int_0^1\big(\mu'(X_i,v\Delta_{nt}(X_i))-\mu'(X_i,0)\big)\Delta_{nt}(X_i)\,dv\Big|=o_p(1). \tag{5.5}$$

In view of Assumption (R2) and $E[\|\dot r_\vartheta(X)\|^2]<\infty$,

$$\Delta_n=\max_{i=1,\dots,n}\sup_{\|t\|\le C}|\Delta_{nt}(X_i)|=O_p\Big(\max_{i=1,\dots,n}\|\dot r_\vartheta(X_i)\|n^{-1/2}\Big)=o_p(1).$$

Thus, by the Cauchy–Schwarz inequality, the square of the left-hand side of (5.5) is bounded by

$$\begin{aligned}&\sup_{\|t\|\le C}\sum_{i=1}^{n}\Delta_{nt}(X_i)^2\,\frac1n\sum_{i=1}^{n}\int_0^1\big(\mu'(X_i,v\Delta_{nt}(X_i))-\mu'(X_i,0)\big)^2\,dv\\&\quad\le\sup_{\|t\|\le C}\sum_{i=1}^{n}\Delta_{nt}(X_i)^2\,\frac1n\sum_{i=1}^{n}\sup_{|s|\le\Delta_n}(\mu'(X_i,s)-\mu'(X_i,0))^2,\end{aligned}$$

which is of order $o_p(1)$ in view of Assumptions (R2) and (H3).

The desired result now follows from (5.1), (5.4), and the $n^{1/2}$-consistency of $\hat\vartheta$. □

We now address the correction term to $\hat H$. We require Assumption B2 on the covariate distribution. We need estimators $\hat\tau_i^2$ of $\tau^2(X_i)$ and $\hat\rho_i$ of $\rho(X_i)$. To keep the argument simple, we use sample splitting and Le Cam's (1956) discretization of $\hat\vartheta$. Our estimators are kernel estimators based on half the sample. To avoid additional integrability assumptions, we truncate ε^2 and $\varepsilon h(\varepsilon)$ with the aid of the function $\psi_{c_n}(y)=(-c_n)\vee x\wedge c_n$, where c_n increases with n at a rate that will be specified later. Let $\hat\vartheta^*$ be a discretized version of $\hat\vartheta$ with values on a grid of mesh size $n^{-1/2}$. Define $\hat\varepsilon_j^*=Y_j-\hat r_{\hat\vartheta^*}(X_j)$. Set $q=[n/2]$. For $i=1,\dots,q$ define

$$\hat\tau_i^2=\frac{\sum_{j=q+1}^n\psi_{c_n}(\hat\varepsilon_j^{*2})K_{b_n}(X_i-X_j)}{\sum_{j=q+1}^nK_{b_n}(X_i-X_j)},\quad\hat\rho_i=\frac{\sum_{j=q+1}^n\psi_{c_n}(\hat\varepsilon_j^*h(\hat\varepsilon_j^*))K_{b_n}(X_i-X_j)}{\sum_{j=q+1}^nK_{b_n}(X_i-X_j)}.$$

Here again $K_{b_n}(x)=K(x/b_n)/b_n$ for a symmetric and bounded density K with support $[-1,1]$, and some bandwidth b_n. For $i=q+1,\dots,n$, define $\hat\tau_i^2$ and $\hat\rho_i$ correspondingly, with summation extending from 1 to q.

Theorem 4. *Suppose Assumptions A1, A2, R, H and B1 hold, and $\hat{\vartheta}$ is $n^{1/2}$-consistent for ϑ. Suppose that $b_n \to 0$, $c_n \to \infty$, and $nb_n/(c_n^2 \log n) \to \infty$. Then*

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\hat{\rho}_i}{\hat{\tau}_i^2}\hat{\varepsilon}_i = \frac{1}{n}\sum_{i=1}^{n} \frac{\rho(X_i)}{\tau^2(X_i)}\varepsilon_i - \int \frac{\rho}{\tau^2}\dot{r}_\vartheta^\top \, dM \cdot (\hat{\vartheta} - \vartheta) + o_p(n^{-1/2}). \tag{5.6}$$

Proof. By Le Cam's discretization argument, since $\hat{\vartheta}^*$ is discrete and $n^{1/2}$-consistent, we can and will assume that it is an arbitrary deterministic sequence $\hat{\vartheta}^* = \vartheta + n^{-1/2}t_n$ with t_n bounded. To simplify notation, we write $\gamma_i = \rho(X_i)/\tau(X_i)^2$ and $\hat{\gamma}_i = \hat{\rho}_i/\hat{\tau}_i^2$. We begin by showing that

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\gamma}_i - \gamma_i)^2 = o_p(1) \tag{5.7}$$

implies the desired (5.6). The error term in (5.6) is $R_{n1} + R_{n2} - (S_{n1} + S_{n2})(\hat{\vartheta} - \vartheta) - T_n$ with

$$\begin{gathered}
R_{n1} = \tfrac{1}{n}\textstyle\sum_{i=1}^{q}(\hat{\gamma}_i - \gamma_i)\varepsilon_i, \qquad R_{n2} = \tfrac{1}{n}\sum_{i=q+1}^{n}(\hat{\gamma}_i - \gamma_i)\varepsilon_i, \\
S_{n1} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\big(\gamma_i\dot{r}_\vartheta(X_i) - E[\gamma_i\dot{r}_\vartheta(X_i)]\big), \qquad S_{n2} = \tfrac{1}{n}\sum_{i=1}^{n}(\hat{\gamma}_i - \gamma_i)\dot{r}_\vartheta(X_i), \\
T_n = \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\hat{\gamma}_i\big(r_{\hat{\vartheta}^*}(X_i) - r_\vartheta(X_i) - \dot{r}_\vartheta(X_i)^\top(\hat{\vartheta} - \vartheta)\big).
\end{gathered}$$

We have $S_{n1} = o_p(1)$ by the law of large numbers. It follows from (5.7) and the Cauchy–Schwarz inequality that $S_{n2} = o_p(1)$. By (5.7), Assumption (R2) and the Cauchy–Schwarz inequality we obtain $T_n = o_p(n^{-1/2})$. Since $\hat{\vartheta}^*$ is taken to be deterministic, and τ^2 is bounded, we get from (5.7),

$$E(nR_{n1}^2 \mid X_1, \ldots, X_n, Y_{p+1}, \ldots, Y_n) = \frac{1}{n}\sum_{i=1}^{q}(\hat{\gamma}_i - \gamma_i)^2\tau^2(X_i) = o_p(1)$$

and hence $R_{n1} = o_p(n^{-1/2})$. Similarly, $R_{n2} = o_p(n^{-1/2})$.

Let us now show (5.7). We begin by showing

$$\frac{1}{n}\sum_{i=1}^{q}(\hat{\gamma}_i - \gamma_i)^2 = o_p(1). \tag{5.8}$$

Since $\|\dot{r}_\vartheta\|$ is in $L_2(M)$, it follows from Assumption (R2) that $\max_j \Delta_{nt_n}(X_j) \to 0$ in probability. Thus there is a sequence of functions $\overline{\Delta}_n$ with $|\overline{\Delta}_n| \le |\Delta_{nt_n}|$ and

$$\sup_{0 \le x \le 1} |\overline{\Delta}_n(x)| \to 0 \tag{5.9}$$

such that $P\big(\Delta_{nt_n}(X_j) \ne \overline{\Delta}_n(X_j)$ for some $j = q+1, \ldots, n\big) \to 0$. In view of this, it suffices to verify (5.8) with $\hat{\gamma}_i$ replaced by $\tilde{\gamma}_i = \Phi_{n1}(X_i)/\Phi_{n2}(X_i)$, where

$$\Phi_{nk} = \frac{1}{n-q}\sum_{j=q+1}^{n} \varphi_{nk}(\varepsilon_j - \overline{\Delta}_n(X_i))K_{b_n}(x - X_j)$$

with $\varphi_{n1}(z) = \psi_{c_n}(zh(z))$ and $\varphi_{n2}(z) = \psi_{c_n}(z^2)$. Both Φ_{n1} and Φ_{n2} are special cases of the statistic $\Phi_n(X_i)$ with

$$\Phi_n(x) = \frac{1}{n-q} \sum_{j=q+1}^{n} \varphi_n(\varepsilon_j - \overline{\Delta}_n(X_i)) K_{b_n}(x - X_j),$$

where φ_n is a sequence of bounded measurable functions with $\|\varphi_n\|_\infty$ bounded away from zero. The expected value of $\Phi_n(x)$ is

$$\overline{\Phi}_n(x) = E[\Phi_n(x)] = \int \overline{\varphi}_n(u) K_{b_n}(x-u)\, M(du)$$

with $\overline{\varphi}_n(u) = \int \varphi_n(u - \overline{\Delta}_n(z)) f(u,z)\, dz$. It follows that

$$\max_{i=1,\dots,q} |\Phi_n(X_i) - \overline{\Phi}_n(X_i)| = o_p(1) \quad \text{if} \quad \frac{nb_n}{\|\varphi_n\|_\infty^2 \log n} \to \infty. \tag{5.10}$$

Indeed we have for $0 < \eta < 1$,

$$\begin{aligned} P\Big(\max_{i=1,\dots,q} |\Phi_n(X_i) - \overline{\Phi}_n(X_i)| > \eta \Big) &\le \sum_{i=1}^{q} E\big[P\big(|\Phi_n(X_i) - \overline{\Phi}_n(X_i)| > \eta \mid X_i \big) \big] \\ &\le q \sup_{0 \le x \le 1} P\big(|\Phi_n(x) - \overline{\Phi}_n(x)| > \eta \big). \end{aligned}$$

The summands of $\Phi_n(x)$ are independent and bounded by $e_n = 2\|\varphi_n\|_\infty \|K\|_\infty / b_n$, have common mean zero, and common variance bounded by $d_n = \beta \|\varphi_n\|_\infty^2 \int K(u)^2\, du / b_n$, where β is a bound for the density of M. Thus we get from the Bernstein inequality, see e.g. Hoeffding (1963), that

$$q \sup_{0 \le x \le 1} P\big(|\Phi_n(x) - \overline{\Phi}_n(x)| > \eta \big) \le 2q \exp\Big(- \frac{(n-q)\eta^2}{2d_n + (2/3)e_n \eta} \Big) \le n \exp(-\eta^2 a_n \log n)$$

for some sequence $a_n \to \infty$. Hence the right-hand side tends to zero and gives the desired (5.10).

Since $\|\varphi_{nk}\|_\infty \le c_n$ for $k = 1, 2$, we get as special cases of (5.10) that

$$\max_{i=1,\dots,q} |\Phi_{nk}(X_i) - \overline{\Phi}_{nk}(X_i)| = o_p(1), \quad k = 1, 2.$$

Thus, if we show that $\overline{\Phi}_{n2}$ is uniformly bounded away from zero, we obtain

$$\frac{1}{n} \sum_{j=1}^{q} \Big(\frac{\Phi_{n1}(X_i)}{\Phi_{n2}(X_i)} - \frac{\overline{\Phi}_{n1}(X_i)}{\overline{\Phi}_{n2}(X_i)} \Big)^2 = o_p(1). \tag{5.11}$$

We have

$$\inf_{0 \le u \le 1} \overline{\varphi}_{n2}(u) m_n(x) \le \overline{\Phi}_{n2}(x) \le \sup_{0 \le u \le 1} \overline{\varphi}_{n2}(u) m_n(x)$$

with $m_n(x) = \int K_{b_n}(x-u)\,M(du)$. By Assumption B2, M has a density with values in a compact interval $[2\alpha,\beta]$ of $(0,\infty)$, and thus $m_n(x)$ takes values in $[\alpha,\beta]$ for $x \in [0,1]$. Since $\tau^2(x) + \overline{\Delta}_n(x)^2 = \int (z-\overline{\Delta}_n(x))^2 f(x,z)\,dz$, we have

$$\bigl|\overline{\varphi}_{n2}(x) - \tau^2(x) - \overline{\Delta}_n(x)^2\bigr| \le \int_{|z-\overline{\Delta}_n(x)|\ge c_n^{1/2}} (z-\overline{\Delta}_n(x))^2 f(x,z)\,dz.$$

We get from Assumption A2 and (5.9) that $\|\overline{\varphi}_{n2} - \tau^2\|_\infty \to 0$. Thus, in view of Assumption A1, for sufficiently large n we have $0 < c_* \le \|\overline{\varphi}_{n2}\| \le c_{**} < \infty$ for some c_*, c_{**} and hence $0 < \alpha c_* \le c_* m_n \le \overline{\Phi}_{n2} \le c_{**} m_n \le c_{**}\beta < \infty$. This completes the proof of (5.11). Thus (5.8) follows if we show

$$\frac{1}{n}\sum_{j=1}^{q} \Bigl(\frac{\rho_n(X_i)}{\tau_n^2(X_i)} - \frac{\rho(X_i)}{\tau^2(X_i)}\Bigr)^2 = o_p(1), \tag{5.12}$$

where $\rho_n = \overline{\Phi}_{n1}/m_n$ and $\tau_n^2 = \overline{\Phi}_{n2}/m_n$. We have

$$\Bigl|\frac{\rho_n}{\tau_n^2} - \frac{\rho}{\tau^2}\Bigr| \le \frac{1}{c_*}\Bigl(|\rho_n - \rho| + \frac{|\rho|}{\tau^2}|\tau_n^2 - \tau^2|\Bigr).$$

An application of the Cauchy–Schwarz inequality gives

$$\rho(X)^2 = \bigl(E(\varepsilon h(\varepsilon) \mid X)\bigr)^2 \le \tau^2(X) E(h(\varepsilon)^2 \mid X).$$

Hence ρ/τ^2 is in $L_2(M)$. Since $|\tau_n^2 - \tau^2| \le c_{**} + C_\tau$, we see from the above that (5.12) follows if we show that ρ_n converges to ρ and τ_n^2 to τ^2 in $L_2(M)$.

Let us now show that these are implied by the convergence of $\overline{\varphi}_{n1}$ to ρ and of $\overline{\varphi}_{n2}$ to τ^2 in $L_2(M)$. This follows from the following more general result. If $\overline{\varphi}_n$ converges to some $\overline{\varphi}$ in $L_2(M)$, then so does $\overline{\Phi}_n/m_n$. Indeed, using $\alpha \le m_n \le \beta$ and the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}
\int \Bigl(\frac{\overline{\Phi}_n}{m_n} - \overline{\varphi}\Bigr)^2 dM &= \int \Bigl(\frac{\int (\overline{\varphi}_n(u) - \overline{\varphi}(x)) K_{b_n}(x-u)\,M(du)}{m_n(x)}\Bigr)^2 M(dx) \\
&\le \frac{1}{\alpha}\int (\overline{\varphi}_n(u) - \overline{\varphi}(x))^2 K_{b_n}(x-u)\,M(du)\,M(dx) \\
&\le \frac{1\beta^2}{\alpha}\iint (\overline{\varphi}_n(u) - \overline{\varphi}(x))^2 K_{b_n}(x-u)\,du dx \\
&\le \frac{1\beta^2}{\alpha}\iint (\overline{\varphi}_n(u) - \overline{\varphi}(u+vb_n))^2 K(v)\,du dv,
\end{aligned}$$

where we interpret $\overline{\varphi}_n$ and $\overline{\varphi}$ as zero off the interval $[0,1]$. The right-hand side converges to zero because $\overline{\varphi}_n$ converges to $\overline{\varphi}$ in L_2, and since the map $w \mapsto \int (\overline{\varphi}(u+w) - \overline{\varphi}(u))^2\,du$ is bounded and uniformly continuous; see e.g. Rudin (1974, Theorem 9.5).

Since we already know that $\overline{\varphi}_{n2}$ converges to τ^2 uniformly, and hence in $L_2(M)$, we are left to show that $\overline{\varphi}_{n1}$ converges to ρ in $L_2(M)$. Let $\overline{\rho}_n(x) = \int (z-\overline{\Delta}_n(x)) h(z-\overline{\Delta}_n(x)) f(x,z)\,dz$.

Since $\{z : zh(z) > c_n\} \subset \{z : |z| > c_n^*\}$ for some $c_n^* \to \infty$, we obtain from an application of the Cauchy–Schwarz inequality that

$$|\overline{\varphi}_{n1}(x) - \overline{\rho}_n(x)|^2 \le \int_{|z-\overline{\Delta}_n(x)|>c_n^*} (z - \overline{\Delta}_n(x))^2 f(x,z)\,dz \int |h(z - \overline{\Delta}_n(x))|^2 f(x,z)\,dz.$$

In view of (5.9) and Assumptions A2 and (H1), it follows that $\overline{\varphi}_{n1} - \overline{\rho}_n$ converges to zero in $L_2(M)$. By an application of the Cauchy–Schwarz inequality,

$$|\overline{\rho}_n(x) - \rho(x)|^2 \le 2\tau^2(x) \int |h(z - \overline{\Delta}_n(x) - h(z)|^2 f(x,z)\,dz + 2\overline{\Delta}_n^2(x) \int |h(z - \overline{\Delta}_n(x)|^2 f(x,z)\,dz.$$

Thus (5.9) and Assumptions A1 and (H1) imply that $\overline{\rho}_n - \rho$ converges to zero in $L_2(M)$. This establishes that $\overline{\varphi}_{n1}$ converges to ρ in $L_2(M)$ and completes the proof of (5.8). Similarly one verifies $\frac{1}{n}\sum_{i=q+1}^n (\hat{\gamma}_i - \gamma_i)^2 = o_p(1)$. This completes the proof of (5.7). □

References

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

W. Hoeffding, Probability inequalities for sums of bounded random variables, *J. Amer. Statist. Assoc.* **58**, (1963), 13–30.

P. J. Bickel, C. A. J. Klaassen, Y. Ritov and J. A. Wellner, *Efficient and Adaptive Estimation for Semiparametric Models*, Springer, New York, 1998.

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

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

J. Hájek, A characterization of limiting distributions of regular estimates, *Z. Wahrsch. Verw. Gebiete* **14**, (1970), 323–330.

H. L. Koul, Asymptotic behavior of Wilcoxon type confidence regions in multiple linear regression, *Ann. Math. Statist.* **40**, (1969), 1950–1979.

H. L. Koul, Some convergence theorems for ranks and weighted empirical cumulatives, *Ann. Math. Statist.* **41**, (1970) 1768–1773.

H. L. Koul, *Weighted Empirical Processes in Dynamic Nonlinear Models, 2nd ed.*, Lecture Notes in Statistics 166, Springer, New York, 2002.

H. L. Koul and A. Schick, Testing for superiority among two regression curves, to appear in: *J. Statist. Plann. Inference* (2003).

L. Le Cam, On the asymptotic theory of estimation and testing hypotheses, in: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Vol. I*, University of California Press, Berkeley, 1956, pp. 129–156.

B. Li, Nonparametric estimating equations based on a penalized information criterion, *Canad. J. Statist.* **28**, (2000), 621–639.

B. Li, On quasi likelihood equations with non-parametric weights, *Scand. J. Statist.* **28**, (2001), 577–602.

R. M. Loynes, The empirical distribution function of residuals from generalised regression, *Ann. Statist.* **8**, (1980), 285–299.

E. Mammen, Empirical process of residuals for high-dimensional linear models, *Ann. Statist.* **24**, (1996), 307–335.

U. U. Müller, A. Schick and W. Wefelmeyer, Estimating linear functionals of the error distribution in nonparametric regression, to appear in: *J. Statist. Plann. Inference* (2003).

H.-G. Müller and U. Stadtmüller, Estimation of heteroscedasticity in regression analysis, *Ann. Statist.* **15**, (1987), 610–625.

M. Ossiander, A central limit theorem under metric entropy with L_2 bracketing, *Ann. Probab.* **15**, (1987), 897–919.

S. Portnoy, Asymptotic behavior of the empiric distribution of M-estimated residuals from a regression model with many parameters, *Ann. Statist.* **14**, (1986), 1152–1170.

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

W. Rudin, *Real and Complex Analysis, 2nd ed.*, McGraw-Hill, New York, 1974.

A. Schick, A note on the construction of asymptotically linear estimators, *J. Statist. Plann. Inference* **16**, (1987), 89–105.

A. Schick, On efficient estimation in regression models, *Ann. Statist.* **21**, (1993), 1486–1521. Correction and addendum: **23**, (1995), 1862–1863.

A. Schick, Improving weighted least-squares estimates in heteroscedastic linear regression when the variance is a function of the mean response, *J. Statist. Plann. Inference* **76**, (1999), 127–144.

A. Schick and W. Wefelmeyer, Estimating the innovation distribution in nonlinear autoregressive models, *Ann. Inst. Statist. Math.* **54**, (2002), 245–260.

A. W. van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes, With Applications to Statistics*, Springer, New York, 1996.

W. Wefelmeyer, Quasi-likelihood models and optimal inference, *Ann. Statist.* **24**, (1996), 405–422.