

Mammen, Enno; van de Geer, Sara

Working Paper

Penalized quasi-likelihood estimation in partial linear models

SFB 373 Discussion Paper, No. 1996,6

Provided in Cooperation with:

Collaborative Research Center 373: Quantification and Simulation of Economic Processes, Humboldt University Berlin

Suggested Citation: Mammen, Enno; van de Geer, Sara (1996) : Penalized quasi-likelihood estimation in partial linear models, SFB 373 Discussion Paper, No. 1996,6, Humboldt University of Berlin, Interdisciplinary Research Project 373: Quantification and Simulation of Economic Processes, Berlin, <https://nbn-resolving.de/urn:nbn:de:kobv:11-10075403>

This Version is available at:

<https://hdl.handle.net/10419/66295>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

Penalized quasi-likelihood estimation in partial linear models *

By

ENNO MAMMEN
Institut für Stochastik
Fachbereich Mathematik
Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin
Germany

and

SARA VAN DE GEER
Mathematical Institute
University of Leiden
P.O. Box 9512
2300 RA Leiden
The Netherlands

Abstract. Consider a partial linear model, where the expectation of a random variable Y depends on covariates (x, z) through $F(\theta_0 x + m_0(z))$, with θ_0 an unknown parameter, and m_0 an unknown function. We apply the theory of empirical processes to derive the asymptotic properties of the penalized quasi-likelihood estimator.

AMS 1991 subject classifications. Primary 62G05, secondary 62G20.

Key words and phrases. Asymptotic normality, penalized quasi-likelihood, rates of convergence.

1. Introduction. Let $(Y_1, T_1), (Y_2, T_2), \dots$ be independent copies of (Y, T) , where Y is a real-valued random variable and $T \in \mathbf{R}^d$. Denote the distribution of (Y, T) by P_0 , and write

$$\mu_0(t) = E_0(Y|T = t),$$

for the conditional expectation of Y given $T = t$. In this paper, we shall study the partial linear model, where $T = (X, Z)$, $X \in \mathbf{R}^{d_1}$, $Z \in \mathbf{R}^{d_2}$, $d_1 + d_2 = d$, and

$$(1.1) \quad \mu_0(x, z) = F(\theta_0' x + m_0(z)),$$

* This research was supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 373 “Quantifikation und Simulation Ökonomischer Prozesse”, Humboldt-Universität zu Berlin and the European Union Human Capital and Mobility Programme ERB CHRX-CT 940693.

with $F : \mathbf{R} \rightarrow \mathbf{R}$ a given function, $\theta_0 \in \mathbf{R}^{d_1}$ an unknown parameter, θ'_0 the transpose of θ_0 , and m_0 an unknown function in a given class of smooth functions. Model (1.1) offers a flexible approach. The inclusion of the linear component $\theta'_0 x$ allows discrete covariates. The link function F may be useful in case of a bounded variable Y (see for instance Example 2 below, where binary observations are considered).

For simplicity, we shall restrict ourselves to the case $d_1 = d_2 = 1$. We shall assume that $T = (X, Z)$ has bounded support, say $T \in [0, 1]^2$, and that m_0 is in the Sobolev class $\{m : J(m) < \infty\}$, where

$$(1.2) \quad J^2(m) = \int_0^1 (m^{(k)}(z))^2 dz.$$

Here, $k \geq 1$ is a fixed integer, and $m^{(k)}$ denotes the k -th derivative of the function m . In summary, the model is

$$\mu_0 = F(g_0),$$

with

$$g_0 \in \mathcal{G} = \{g(x, z) = \theta x + m(z) : \theta \in \mathbf{R}, J(m) < \infty\}.$$

For $g \in \mathcal{G}$, $g(x, z) = \theta x + m(z)$, we shall often write $J(g) = J(m)$.

Define the quasi-(log-)likelihood function

$$(1.3) \quad Q(y; \mu) = \int_y^\mu \frac{(y-s)}{V(s)} ds,$$

with a known function $V : (a, b) \rightarrow (0, \infty)$, $-\infty \leq a < b \leq \infty$, given. The quasi-likelihood function was first considered by Wedderburn (1974). Properties of quasi-likelihood functions are discussed in McCullagh (1983) and McCullagh and Nelder (1989). There, the function V has been chosen as the conditional variance of the response Y , and it has been assumed that V depends only on the conditional mean μ of Y , i.e. $V = V(\mu)$. The quasi-likelihood approach is a generalization of generalized linear models. The log-likelihood of an exponential family is replaced by a quasi-likelihood, in which only the relation between the conditional mean and the conditional variance has to be specified. To see the relations of the quasi-likelihood functions with generalized linear models note for instance that the maximum likelihood estimate $\hat{\vartheta}$ based on an i.i.d. sample Y_1, \dots, Y_n from an exponential family with mean ϑ and variance $V(\vartheta)$ is given by

$$\sum_{i=1}^n \frac{d}{d\vartheta} Q(Y_i; \vartheta)|_{\vartheta=\hat{\vartheta}} = 0.$$

In this paper we do not assume that $V(\mu_0)$ is the conditional variance of Y . The only assumptions on the distribution of Y we use in this paper concern the form of the conditional mean (see (1.1)) and subexponential tails (see (A0), below). In particular, our results may be used in case of model misspecification.

Let us now describe the estimation procedure. Let $\lambda_n > 0$ be a smoothing parameter. The penalized quasi-likelihood estimator is defined by

$$(1.4) \quad \hat{g}_n \in \arg \max_{g \in \mathcal{G}} [\bar{Q}_n(F(g)) - \lambda_n^2 J^2(g)],$$

where

$$\bar{Q}_n(\mu) = \frac{1}{n} \sum_{i=1}^n Q(Y_i; \mu(T_i)).$$

Throughout, we assume that indeed a solution $\hat{g}_n \in \mathcal{G}$ of the maximization problem (1.4) exists. Then $\hat{g}_n(x, z) = \hat{\theta}_n x + \hat{m}_n(z)$, where $\hat{\theta}_n \in \mathbf{R}$, and $J(\hat{m}_n) < \infty$. The estimated conditional expectation is $\hat{\mu}_n = F(\hat{g}_n)$.

Generalized linear models of the form (1.1) have first been considered by Green and Yandell (1985) and Green (1987). The generalization to quasi-likelihood models has also been studied in e.g. Chen (1988), Speckmann (1988) and Severini and Staniswalis (1994). These papers however use different estimation procedures, such as polynomial approximation or kernel smoothing. Local polynomial smoothing based on quasi-likelihood functions is discussed in Fan, Heckman and Wand (1995).

Our main aim is to obtain asymptotic normality of the penalized quasi-likelihood estimator $\hat{\theta}_n$ of θ_0 , but first we derive a rate of convergence for \hat{g}_n . The asymptotic properties of the estimators depend of course on the behaviour of the smoothing parameter λ_n as $n \rightarrow \infty$. It may be random (e.g. determined through cross-validation). We assume $\lambda_n = o_{\mathbf{P}}(n^{-1/4})$, and $(1/\lambda_n) = O_{\mathbf{P}}(n^{k/(2k+1)})$.

The following example is an important special case.

Example 1. Let F be the identity, and $V \equiv 1$. Then $Q(y; \mu) = -(y - \mu)^2/2$, so that \hat{g}_n is the penalized least squares estimator. It is called a partial smoothing spline. If λ_n is non-random, $\hat{\theta}_n$ and \hat{m}_n are linear in Y_1, \dots, Y_n . See e.g. Wahba (1984), Silverman (1985).

Denote the conditional expectation of X given $Z = z$ by $h_1(z)$, $z \in [0, 1]$. If $J(h_1) < \infty$ and $\{\lambda_n\}$ is of the order given above and non-random, then the bias of $\hat{\theta}_n$ is $O(\lambda_n^2) = o(n^{-1/2})$, whereas its variance is $O(1/n)$. This is a result of Rice (1986). It indicates that the smoothness imposed on \hat{m}_n (in terms of the number of derivatives k) should not exceed the smoothness of h_1 . In Theorem 4.1, we shall prove \sqrt{n} -consistency and asymptotic normality of $\hat{\theta}_n$ under the condition $J(h_1) < \infty$. In Remark 4.1, we show that in case of rough functions h_1 , \sqrt{n} -consistency of $\hat{\theta}_n$ can be guaranteed by undersmoothing. More precisely, there we allow that h_1 depends on n and that $J(h_1) \rightarrow \infty$. We show that $\hat{\theta}_n$ is \sqrt{n} -consistent and asymptotically normal, as long as λ_n is chosen small enough. Even for the optimal choice $\lambda_n \sim n^{-k/(2k+1)}$, $J(h_1)$ may tend to infinity. This shows that much less smoothness is needed for h_1 than for m_0 .

The theory for general penalized quasi-likelihood estimators essentially boils down to that for Example 1, provided one can properly linearize in a neighbourhood of the true parameters. For this purpose, we first need to prove consistency, which is not too difficult if $V(s)$ stays away from zero. Unfortunately, this is frequently not the case, as we see in Examples 2 and 3 below. In Section 7, we shall employ an ad hoc method to handle

Example 2. In general, one can say that given consistency, the further conditions for asymptotic normality are relatively innocent, but proving consistency can be somewhat involved.

Example 2. Let $Y \in \{0, 1\}$, $P_0(Y = 1|T = t) = F(g_0(t))$, and let $V(s) = s(1 - s)$, $s \in (0, 1)$. In this case, the quasi-likelihood is the exact likelihood, so that \hat{g}_n is the penalized maximum likelihood estimator.

Example 3. Let $Y \in (0, \infty)$, and $V(s) = 1/s^2$, $s > 0$. Then $Q(y; \mu)$ is the log-likelihood corresponding to the exponential distribution with parameter $1/\mu$.

This paper can be seen as a statistical application of empirical process theory as considered in Dudley (1984), Giné and Zinn (1984), Pollard (1984, 1990), Ossiander (1987), and others. Some concepts and results in this field are presented in Section 2. In Section 3, rates of convergence are obtained, and Section 4 uses the rates to establish asymptotic normality. In Section 5, we discuss bootstrapping the distribution of $\hat{\theta}_n$. Examples 1-3 are studied in Section 6, and Section 7 revisits Example 2.

2. Main assumptions, notation and technical tools.

2.1. Main assumptions. We recall the assumption $(X, Z) \in [0, 1]^2$, and

$$(2.1) \quad \lambda_n = o_{\mathbf{P}}(n^{-1/4}), \quad 1/\lambda_n = O_{\mathbf{P}}(n^{k/(2k+1)}).$$

We also suppose throughout that $f(\xi) = dF(\xi)/d\xi$ exists for all $\xi \in \mathbf{R}$.

Write $W = Y - \mu_0(T)$ ($W_i = Y_i - \mu_0(T_i)$, $i = 1, 2, \dots$). The following condition is essential in Section 3: for some constant $0 < C_0 < \infty$,

$$(A0) \quad E_0(e^{|W|/C_0}|T) \leq C_0, \text{ almost surely.}$$

Let $\phi_j(x, z) = z^{j-1}$, $j = 1, \dots, k$, and $\phi_{k+1}(x, z) = x$. We assume that the matrix

$$A = \int \phi\phi' dP_0$$

is non-singular. Here, ϕ' denotes the transpose of ϕ .

2.2. Notation. By the Sobolev-embedding Theorem, one can write

$$m(z) = m_1(z) + m_2(z),$$

with

$$m_1(z) = \sum_{j=1}^k \beta_j z^{j-1},$$

and $|m_2(z)| \leq J(m_2) = J(m)$ (see e.g. Oden and Reddy (1976)). So for $g(x, z) = \theta x + m(z)$,

$$g(x, z) = g_1(x, z) + g_2(x, z),$$

with

$$g_1(x, z) = \sum_{j=1}^{k+1} \beta_j \phi_j(x, z), \quad \theta = \beta_{k+1},$$

i.e. $g_1 = \beta' \phi$, and $g_2(x, z) = m_2(z)$, $|g_2(x, z)| \leq J(g_2) = J(m)$.

For a measurable function $a : \mathbf{R} \times [0, 1]^2 \rightarrow \mathbf{R}$, $E_0(a(Y, T)) = \int a dP_0$ denotes the expectation of $a(Y, T)$ (whenever it exists) and

$$\|a\|^2 = E_0 a^2(Y, T), \quad \|a\|_n^2 = \frac{1}{n} \sum_{i=1}^n a^2(Y_i, T_i).$$

With a slight abuse of notation, we also write for $a : [0, 1]^2 \rightarrow \mathbf{R}$ depending only on $t \in [0, 1]^2$,

$$\|a\|^2 = E_0 a^2(T), \quad \|a\|_n^2 = \frac{1}{n} \sum_{i=1}^n a^2(T_i).$$

Moreover,

$$|a|_\infty = \sup_{t \in [0, 1]^2} |a(t)|,$$

and for $\beta \in \mathbf{R}^{k+1}$, $\|\beta\|^2 = \beta' \beta$.

Let \mathcal{A} be a subset of a (pseudo-)metric space (\mathcal{L}, ρ) of real-valued functions.

Definition. The δ -covering number $N(\delta, \mathcal{A}, \rho)$ of \mathcal{A} is the smallest value of N for which there exist functions a_1, \dots, a_N in \mathcal{L} , such that for each $a \in \mathcal{A}$, $\rho(a, a_j) \leq \delta$ for some $j \in \{1, \dots, N\}$. The δ -covering number with bracketing $N_B(\delta, \mathcal{A}, \rho)$ is the smallest value of N for which there exist pairs of functions $\{[a_j^L, a_j^U]\}_{j=1}^N \subset \mathcal{L}$, with $\rho(a_j^L, a_j^U) \leq \delta$, $j = 1, \dots, N$, such that for each $a \in \mathcal{A}$ there is a $j \in \{1, \dots, N\}$ such that $a_j^L \leq a \leq a_j^U$. The δ -entropy (δ -entropy with bracketing) is defined as $H(\delta, \mathcal{A}, \rho) = \log N(\delta, \mathcal{A}, \rho)$ ($H_B(\delta, \mathcal{A}, \rho) = \log N_B(\delta, \mathcal{A}, \rho)$).

2.3. Technical tools.

Theorem 2.1. For each $0 < C < \infty$ we have

$$\sup_{\delta > 0} \delta^{1/k} H(\delta, \{g \in \mathcal{G} : |g|_\infty \leq C, J(g) \leq C\}, |\cdot|_\infty) < \infty.$$

Proof. See Birman and Solomjak (1967). □

Theorem 2.2. Write (AA0) for the assumption that given T , W is (uniformly) sub-Gaussian, i.e. for some constant $0 < C_0 < \infty$,

$$(AA0) \quad E_0(e^{[W^2/C_0]} | T) \leq C_0, \text{ almost surely.}$$

Let \mathcal{A} be a uniformly bounded class of functions $a : [0, 1]^2 \rightarrow \mathbf{R}$ depending only on $t \in [0, 1]^2$. Let $0 < \nu < 2$. Suppose that either (A0) holds and

$$(2.2) \quad \limsup_{n \rightarrow \infty} \sup_{\delta > 0} \delta^\nu H_B(\delta, \mathcal{A}, \|\cdot\|_n) < \infty, \text{ almost surely,}$$

or that (AA0) holds and

$$(2.3) \quad \limsup_{n \rightarrow \infty} \sup_{\delta > 0} \delta^\nu H(\delta, \mathcal{A}, \|\cdot\|_n) < \infty, \text{ almost surely.}$$

Then

$$(2.4) \quad \sup_{a \in \mathcal{A}} \frac{(1/n) \sum_{i=1}^n W_i a(T_i)}{(\|a\|_n \vee n^{-\frac{1}{2+\nu}})^{1-\nu/2}} = O_{\mathbf{P}}(n^{-1/2}).$$

Proof. It is shown in van de Geer (1990) that (AA0) and (2.3) imply (2.4). Similar arguments as there, combined with e.g. a result of Birgé and Massart (1991, Theorem 4), show that (AA0) can be relaxed to (A0), provided (2.3) is strengthened to (2.2) (see also van de Geer (1995)). \square

The following theorem will be used in Section 4, only to show that consistency in $\|\cdot\|_n$ -norm implies consistency in $\|\cdot\|$ -norm. Nevertheless, we present it in its full strength, so that one can verify that the rates for the $\|\cdot\|_n$ -norm and $\|\cdot\|$ -norm coincide.

Theorem 2.3. *Suppose \mathcal{A} is uniformly bounded and that for some $0 < \nu < 2$,*

$$(2.5) \quad \sup_{\delta > 0} \delta^\nu H_B(\delta, \mathcal{A}, \|\cdot\|) < \infty.$$

Then for all $\eta > 0$ there exists a $0 < C < \infty$ such that

$$(2.6) \quad \limsup_{n \rightarrow \infty} \mathbf{P} \left(\sup_{a \in \mathcal{A}, \|a\| > Cn^{-\frac{1}{2+\nu}}} \left| \frac{\|a\|_n}{\|a\|} - 1 \right| > \eta \right) = 0.$$

Proof. See van de Geer (1988, Lemma 6.3.4). \square

Theorem 2.4. *Suppose that for some $0 < \nu < 2$,*

$$(2.7) \quad \sup_{\delta > 0} \delta^\nu H_B(\delta, \mathcal{A}, \|\cdot\|) < \infty.$$

Then for all $\eta > 0$ there is a $\delta > 0$ such that

$$(2.8) \quad \limsup_{n \rightarrow \infty} \mathbf{P} \left(\sup_{a \in \mathcal{A}, \|a\| \leq \delta} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n (a(Y_i, T_i) - P_0(a)) \right| > \eta \right) < \eta.$$

Proof. Condition (2.7) ensures that \mathcal{A} is a Donsker class, and (2.8) is the implied asymptotic equicontinuity of the empirical process. See e.g. Pollard (1990) and the references there for general theory on Donsker classes. \square

3. Rates of convergence. Define

$$l(\xi) = \frac{f(\xi)}{V(F(\xi))}, \quad \xi \in \mathbf{R},$$

and consider the following assumptions: for some constants $0 < C_1, C_2 < \infty$,

$$(A1) \quad V(s) \geq 1/C_1, \text{ for all } s \in (a, b),$$

and

$$(A2) \quad \frac{1}{C_2} \leq |l(\xi)| \leq C_2, \text{ for all } \xi \in \mathbf{R}.$$

Clearly, (A1) and (A2) hold in Example 1, where $f = V(F) \equiv 1$. In fact, under (A1) and (A2), the rates problem essentially reduces to the one of Example 1. It is possible to avoid these assumptions, as we shall illustrate in Section 7.

Lemma 3.1. *Suppose (A1) and (A2) are met. Then*

$$(3.1) \quad \|\hat{g}_n - g_0\|_n = O_{\mathbf{P}}(\lambda_n),$$

and

$$(3.2) \quad J(\hat{g}_n) = O_{\mathbf{P}}(1).$$

Proof. For a fixed y_0 , we write

$$\gamma_g = \int_{y_0}^{F(g)} \frac{1}{V(s)} ds, \quad g \in \mathcal{G},$$

and $\hat{\gamma}_n = \gamma_{\hat{g}_n}$, $\gamma_0 = \gamma_{g_0}$. We have

$$\bar{Q}_n(\hat{\mu}_n) - \bar{Q}_n(\mu_0) = \frac{1}{n} \sum_{i=1}^n W_i (\hat{\gamma}_n(T_i) - \gamma_0(T_i)) - \frac{1}{n} \sum_{i=1}^n \int_{\mu_0(T_i)}^{\hat{\mu}_n(T_i)} \frac{(s - \mu_0(T_i))}{V(s)} ds.$$

So, by the Cauchy-Schwarz inequality and (A1),

$$(3.3) \quad \bar{Q}_n(\hat{\mu}_n) - \bar{Q}_n(\mu_0) \leq \left(\frac{1}{n} \sum_{i=1}^n W_i^2 \right)^{1/2} \|\hat{\gamma}_n - \gamma_0\|_n - \frac{1}{2C_1} \|\hat{\gamma}_n - \gamma_0\|_n^2.$$

Note that $(1/n) \sum_{i=1}^n W_i^2 = O(1)$ almost surely (by (A0)). On the other hand, because $\hat{\mu}_n = F(\hat{g}_n)$ maximizes $\bar{Q}_n(F(g)) - \lambda_n^2 J^2(g)$, we have

$$(3.4) \quad \bar{Q}_n(\hat{\mu}_n) - \bar{Q}_n(\mu_0) \geq \lambda_n^2 (J^2(\hat{g}_n) - J^2(g_0)) \geq o_{\mathbf{P}}(1).$$

The combination of (3.3) and (3.4) gives

$$\|\hat{\gamma}_n - \gamma_0\|_n^2 \leq \|\hat{\gamma}_n - \gamma_0\|_n O(1) + o_{\mathbf{P}}(1),$$

which implies $\|\hat{\gamma}_n - \gamma_0\|_n = O_{\mathbf{P}}(1)$.

In view of (A2),

$$(3.5) \quad \frac{1}{C_2} |g(t) - \tilde{g}(t)| \leq |\gamma_g(t) - \gamma_{\tilde{g}}(t)| \leq C_2 |g(t) - \tilde{g}(t)|,$$

for all $t \in [0, 1]^2$ and all $g, \tilde{g} \in \mathcal{G}$. So also $\|\hat{g}_n - g_0\|_n = O_{\mathbf{P}}(1)$, so that $\|\hat{g}_n\|_n = O_{\mathbf{P}}(1)$. We shall now show that $|\hat{g}_n|_{\infty}/(1 + J(\hat{g}_n)) = O_{\mathbf{P}}(1)$. As in Section 2.2, write

$$\hat{g}_n = \hat{g}_{1n} + \hat{g}_{2n},$$

with $\hat{g}_{1n} = \hat{\beta}'_n \phi$, and $|\hat{g}_{2n}|_{\infty} \leq J(\hat{g}_n)$. Then

$$(3.6) \quad \frac{\|\hat{g}_{1n}\|_n}{1 + J(\hat{g}_n)} \leq \frac{\|\hat{g}_n\|_n}{1 + J(\hat{g}_n)} + \frac{\|\hat{g}_{2n}\|_n}{1 + J(\hat{g}_n)} = O_{\mathbf{P}}(1).$$

Now, $A = \int \phi \phi' dP_0$ is assumed to be non-singular, and

$$\frac{1}{n} \sum_{i=1}^n \phi(T_i) \phi'(T_i) \rightarrow A, \text{ almost surely.}$$

Thus, (3.6) implies that $\|\hat{\beta}_n\|/(1 + J(\hat{g}_n)) = O_{\mathbf{P}}(1)$. Because T is in a bounded set, also $|\hat{g}_{1n}|_{\infty}/(1 + J(\hat{g}_n)) = O_{\mathbf{P}}(1)$. So $|\hat{g}_n|_{\infty}/(1 + J(\hat{g}_n)) = O_{\mathbf{P}}(1)$.

In view of (3.5), we now have $|\hat{\gamma}_n|_{\infty}/(1 + J(\hat{g}_n)) = O_{\mathbf{P}}(1)$. Moreover, $|\gamma_g - \gamma_{\tilde{g}}|_{\infty} \leq C_2 |g - \tilde{g}|_{\infty}$, $g, \tilde{g} \in \mathcal{G}$. So by Theorem 2.1,

$$\sup_{\delta > 0} \delta^{1/k} H(\delta, \{ \frac{\gamma_g - \gamma_{g_0}}{1 + J(g)} : g \in \mathcal{G}, \frac{|\gamma_g|_{\infty}}{1 + J(g)} \leq C \}, |\cdot|_{\infty}) < \infty.$$

Using Theorem 2.2, assumption (A0), and the fact that $\|\hat{\gamma}_n - \gamma_0\|_n \geq (1/C_2) \|\hat{g}_n - g_0\|_n$, we find

$$(3.7) \quad \frac{(1/n) \sum_{i=1}^n W_i (\hat{\gamma}_n(T_i) - \gamma_0(T_i))}{\|\hat{g}_n - g_0\|_n^{1-1/(2k)} (1 + J(\hat{g}_n))^{1/(2k)} \vee (1 + J(\hat{g}_n)) n^{-\frac{2k-1}{2(2k+1)}}} = O_{\mathbf{P}}(n^{-1/2}).$$

Invoke this in (3.4):

$$(3.8) \quad \begin{aligned} & \lambda_n^2 (J^2(\hat{g}_n) - J^2(g_0)) \leq \bar{Q}_n(\hat{\mu}_n) - \bar{Q}_n(\mu_0) \\ & \leq \frac{1}{n} \sum_{i=1}^n W_i (\hat{\gamma}_n(T_i) - \gamma_0(T_i)) - \frac{1}{2C_1} \|\hat{\gamma}_n - \gamma_0\|_n^2 \\ & \leq (\|\hat{g}_n - g_0\|_n^{1-\frac{1}{2k}} (1 + J(\hat{g}_n))^{\frac{1}{2k}} \vee (1 + J(\hat{g}_n)) n^{-\frac{2k-1}{2(2k+1)}}) O_{\mathbf{P}}(n^{-1/2}) \end{aligned}$$

$$-\frac{1}{2C_1C_2^2} \|\hat{g}_n - g_0\|_n^2.$$

Thus,

$$\begin{aligned} & \lambda_n^2 J^2(\hat{g}_n) \leq \lambda_n^2 J^2(g_0) \\ & + (\|\hat{g}_n - g_0\|_n^{1-\frac{1}{2k}} (1 + J(\hat{g}_n))^{\frac{1}{2k}} \vee (1 + J(\hat{g}_n)) n^{-\frac{2k-1}{2(2k+1)}}) O_{\mathbf{P}}(n^{-1/2}), \end{aligned}$$

as well as

$$\begin{aligned} & \|\hat{g}_n - g_0\|_n^2 \leq \lambda_n^2 J^2(g_0) \\ & + (\|\hat{g}_n - g_0\|_n^{1-\frac{1}{2k}} (1 + J(\hat{g}_n))^{\frac{1}{2k}} \vee (1 + J(\hat{g}_n)) n^{-\frac{2k-1}{2(2k+1)}}) O_{\mathbf{P}}(n^{-1/2}). \end{aligned}$$

Solve these two inequalities to find that

$$\|\hat{g}_n - g_0\|_n^2 + \lambda_n^2 J^2(\hat{g}_n) = O_{\mathbf{P}}(\lambda_n^2 + \lambda_n^{-1/k} n^{-1} + \lambda_n^{-2} n^{-\frac{4k}{2k+1}}).$$

Because we assumed $\lambda_n^{-1} = O_{\mathbf{P}}(n^{k/(2k+1)})$, this completes the proof. \square

Remark 3.1. The situation can be adjusted to the case of triangular arrays. Let $(Y_{1,n}, T_{1,n}), \dots, (Y_{n,n}, T_{n,n})$ be independent copies of $(Y_{0,n}, T_{0,n})$, and suppose that the conditional expectation of $Y_{0,n}$ given $T_{0,n}$ is equal to $F(g_{0,n}(T_{0,n}))$, with $g_{0,n} \in \mathcal{G}$, $n = 1, 2, \dots$. Assume that (A0) holds for $W_{0,n} = Y_{0,n} - F(g_{0,n}(T_{0,n}))$ and $T_{0,n}$, with constant C_0 not depending on n . Assume moreover that for $A_{0,n} = \int \phi \phi' dP_{0,n}$, $P_{0,n}$ being the distribution of $T_{0,n}$, we have

$$\beta' A_{0,n} \beta \geq c_0 \beta' \beta \text{ for all } \beta \in \mathbf{R}^{k+1},$$

where $c_0 > 0$ is independent of n . Then one finds under (A1) and (A2), for $1/\lambda_n = O_{\mathbf{P}}(n^{k/(2k+1)}(1 + J(g_{0,n}))^{2k/(2k+1)})$,

$$\|\hat{g}_n - g_{0,n}\|_n = O_{\mathbf{P}}(\lambda_n(1 + J(g_{0,n}))),$$

and

$$J(\hat{g}_n) = O_{\mathbf{P}}(1 + J(g_{0,n})).$$

4. Asymptotic normality. We shall use the assumptions: for some constants $0 < \eta_0, C_3, C_4 < \infty$, and for all $t \in [0, 1]^2$, we have for $\xi_0 = g_0(t)$,

$$(A3) \quad |l(\xi_0)| \leq C_3, \text{ and } |l(\xi) - l(\xi_0)| \leq C_3 |\xi - \xi_0| \text{ for all } |\xi - \xi_0| \leq \eta_0,$$

and

$$(A4) \quad |f(\xi_0)| \leq C_4, \text{ and } |f(\xi) - f(\xi_0)| \leq C_4 |\xi - \xi_0| \text{ for all } |\xi - \xi_0| \leq \eta_0.$$

Write $l_0 = l(g_0)$ and $f_0 = f(g_0)$, and take

$$h_1(z) = \frac{E_0(X f_0(T) l_0(T) | Z = z)}{E_0(f_0(T) l_0(T) | Z = z)},$$

and

$$h_2(x, z) = x - h_1(z).$$

Also define

$$\tilde{h}_1(z) = E_0(X|Z = z),$$

and

$$\tilde{h}_2(x, z) = x - \tilde{h}_1(z).$$

Theorem 4.1 below gives conditions for asymptotic normality of $\hat{\theta}_n$. If the conditional distribution of Y belongs to an exponential family with mean μ and variance $V(\mu)$, then $\hat{\theta}_n$ is asymptotically efficient. The conditional variance of Y given T is in that case

$$\text{var}_0(Y|T) = E_0(W^2|T) = V(\mu_0(T)),$$

so that

$$E_0(W^2 l_0^2(T) h_2^2(T)) = \| (f_0 l_0)^{1/2} h_2 \|^2 .$$

According to Theorem 4.1, the asymptotic variance of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is then $\| (f_0 l_0)^{1/2} h_2 \|^2$.

Theorem 4.1. *Suppose (A3) and (A4) are met. Assume moreover that*

$$(4.1) \quad \|\hat{g}_n - g_0\|_n = o_{\mathbf{P}}(n^{-1/4}),$$

$$(4.2) \quad J(\hat{g}_n) = O_{\mathbf{P}}(1),$$

$$(4.3) \quad \|\tilde{h}_2\| > 0,$$

$$(4.4) \quad Z \text{ has density bounded away from 0 on its support,}$$

$$(4.5) \quad J(h_1) < \infty,$$

and

$$(4.6) \quad \|(f_0 l_0)^{1/2} h_2\| > 0.$$

Then,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n W_i l_0(T_i) h_2(T_i)}{\|(f_0 l_0)^{1/2} h_2\|^2} + o_{\mathbf{P}}(1).$$

Proof. We shall apply Theorem 2.3, to conclude from (4.1) that $\|\hat{g}_n - g_0\| = o_{\mathbf{P}}(1)$. Because Theorem 2.3 is on uniformly bounded classes, we first verify that $|\hat{g}_n|_{\infty} = O_{\mathbf{P}}(1)$. This follows by the same arguments as used in Lemma 3.1. Because (4.2) holds, also

$|\hat{g}_{2n}|_\infty = O_{\mathbf{P}}(1)$. So $\|\hat{g}_{1n}\|_n = O_{\mathbf{P}}(1)$. Again, because of the assumed non-singularity of A , this implies $|\hat{g}_{1n}|_\infty = O_{\mathbf{P}}(1)$, so $|\hat{g}_n|_\infty = O_{\mathbf{P}}(1)$.

Now, $\|\tilde{h}_2\| > 0$, so $\|\hat{g}_n - g_0\| = o_{\mathbf{P}}(1)$ implies $|\hat{\theta}_n - \theta_0| = o_{\mathbf{P}}(1)$. Hence, also $\|\hat{m}_n - m_0\| = o_{\mathbf{P}}(1)$. Assumption (4.4) ensures that

$$\sup_{z \in \text{support}(Z)} |\hat{m}_n(z) - m_0(z)| = o_{\mathbf{P}}(1).$$

Therefore, we may without loss of generality assume that

$$(4.7) \quad |\hat{g}_n - g_0|_\infty \leq \eta_0,$$

so that we can use (A3) and (A4).

Because of (4.5), we have that

$$\begin{aligned} \hat{g}_{ns}(x, z) &= \hat{g}_n(x, z) + sh_2(x, z) \\ &= (\hat{\theta}_n + s)x + (\hat{m}_n(z) - sh_1(z)) \in \mathcal{G}, \end{aligned}$$

for all $s \in \mathbf{R}$. Thus,

$$(4.8) \quad \frac{d}{ds} [\bar{Q}_n(F(\hat{g}_{ns})) - \lambda_n^2 J^2(\hat{g}_{ns})]_{s=0} = 0.$$

Clearly, for $\hat{l}_n = l(\hat{g}_n)$, $\hat{f}_n = f(\hat{g}_n)$,

$$\frac{d}{ds} \bar{Q}_n(F(\hat{g}_{ns}))_{s=0} = \frac{1}{n} \sum_{i=1}^n W_i \hat{l}_n(T_i) h_2(T_i) - \frac{1}{n} \sum_{i=1}^n [\hat{\mu}_n(T_i) - \mu_0(T_i)] \hat{l}_n(T_i) h_2(T_i) = I - II.$$

Use (A3) and Theorem 2.1, to find that the class

$$\{[y - \mu_0(t)]l(g(t))h_2(t) : g \in \mathcal{G}, |g - g_0|_\infty \leq \eta_0, J(g) \leq C\}$$

satisfies (2.7) of Theorem 2.4. Since, also by (A3), $\|\hat{g}_n - g_0\| = o_{\mathbf{P}}(1)$ implies $\|\hat{l}_n - l_0\| = o_{\mathbf{P}}(1)$, we obtain

$$(4.9) \quad I = \frac{1}{n} \sum_{i=1}^n W_i l_0(T_i) h_2(T_i) + o_{\mathbf{P}}(n^{-1/2}).$$

Let us write

$$\begin{aligned} II &= \frac{1}{n} \sum_{i=1}^n [(\hat{g}_n(T_i) - g_0(T_i))f_0(T_i)]l_0(T_i)h_2(T_i) \\ &+ \frac{1}{n} \sum_{i=1}^n [\hat{\mu}_n(T_i) - \mu_0(T_i) - (\hat{g}_n(T_i) - g_0(T_i))f_0(T_i)]l_0(T_i)h_2(T_i) \end{aligned}$$

$$\begin{aligned}
& + \frac{1}{n} \sum_{i=1}^n [\hat{\mu}_n(T_i) - \mu_0(T_i)] (\hat{l}_n(T_i) - l_0(T_i)) h_2(T_i) \\
& = III + IV + V.
\end{aligned}$$

Observe that

$$\hat{g}_n(x, z) - g_0(x, z) = (\hat{\theta}_n - \theta_0)x + \hat{m}_n(z) - m_0(z) = (\hat{\theta}_n - \theta_0)h_2(x, z) + \hat{a}_n(z) - a_0(z),$$

where $\hat{a}_n(z) - a_0(z) = (\hat{\theta}_n - \theta_0)h_1(z) + \hat{m}_n(z) - m_0(z)$. Hence,

$$(4.10) \quad III = (\hat{\theta}_n - \theta_0) \frac{1}{n} \sum_{i=1}^n f_0(T_i) l_0(T_i) h_2^2(T_i) + \frac{1}{n} \sum_{i=1}^n [\hat{a}_n(Z_i) - a_0(Z_i)] f_0(T_i) l_0(T_i) h_2(T_i).$$

Because $|\hat{\theta}_n - \theta_0| = o_{\mathbf{P}}(1)$ and $\|\hat{m}_n - m_0\| = o_{\mathbf{P}}(1)$, also $\|\hat{a}_n - a_0\| = o_{\mathbf{P}}(1)$. Moreover, for any measurable function $a : [0, 1] \rightarrow \mathbf{R}$, $E_0(a(Z)f_0(T)l_0(T)h_2(T)) = 0$. So, according to Theorem 2.1, combined with Theorem 2.4, the second term in (4.10) is $o_{\mathbf{P}}(n^{-1/2})$. This, and the law of large numbers, yields

$$III = (\hat{\theta}_n - \theta_0) (\| (f_0 l_0)^{1/2} h_2 \| + o(1)) + o_{\mathbf{P}}(n^{-1/2}).$$

Invoke (A3) and (A4) to conclude that under (4.7),

$$\begin{aligned}
|IV| & \leq C_4 \frac{1}{n} \sum_{i=1}^n (\hat{g}_n(T_i) - g_0(T_i))^2 |l_0(T_i) h_2(T_i)| \\
& \leq C_3 C_4 \|\hat{g}_n - g_0\|_n^2 = o_{\mathbf{P}}(n^{-1/2}),
\end{aligned}$$

and similarly,

$$|V| \leq C_3 C_4 \|\hat{g}_n - g_0\|_n^2 = o_{\mathbf{P}}(n^{-1/2}).$$

Thus,

$$(4.11) \quad II = (\hat{\theta}_n - \theta_0) (\| (f_0 l_0)^{1/2} h_2 \|^2 + o(1)) + o_{\mathbf{P}}(n^{-1/2}).$$

Finally, we note that (4.2), (4.5) and the condition $\lambda_n = o_{\mathbf{P}}(n^{-1/4})$ give

$$(4.12) \quad \frac{d}{ds} \lambda_n^2 J^2(\hat{g}_{ns})|_{s=0} \leq 2\lambda_n^2 J(\hat{g}_n) J(h_1) = o_{\mathbf{P}}(n^{-1/2}).$$

Combine (4.8), (4.9), (4.11) and (4.12) to obtain

$$0 = \frac{1}{n} \sum_{i=1}^n W_i l_0(T_i) h_2(T_i) + (\hat{\theta}_n - \theta_0) (\| (f_0 l_0)^{1/2} h_2 \|^2 + o(1)) + o_{\mathbf{P}}(n^{-1/2}).$$

Apply condition (4.6) to complete the proof. \square

Remark 4.1. Theorem 4.1 can be generalized to the situation as described in Remark 3.1. Let us suppose the assumptions given there are met, and that in addition (A3) and (A4) hold, with constants η_0 , C_3 and C_4 not depending on n . Suppose that also (4.3), (4.4) and (4.6) hold uniformly in n . Replace (4.1) and (4.2) by the condition

$$\lambda_n(1 + J(g_{0,n})) = o_{\mathbf{P}}(n^{-1/4}),$$

and replace (4.5) by

$$(4.13) \quad \lambda_n^2(1 + J(g_{0,n}))J(h_{1,n}) = o_{\mathbf{P}}(n^{-1/2}).$$

Then the conclusion of Theorem 4.1 is valid, provided that we can apply Theorem 2.3 to conclude that $\|\hat{g}_n - g_{0,n}\| = o_{\mathbf{P}}(1)$. For this purpose, we assume in addition to the above that

$$(1 + J(g_{0,n})) = O(n^{\frac{k}{2k+1}}).$$

For bounded $J(g_{0,n})$, condition (4.13) holds if $J(h_{1,n})$ is bounded. This follows from our assumption $\lambda_n = O_{\mathbf{P}}(n^{-1/4})$. For optimal choices $\lambda_n \sim n^{-k/(2k+1)}$, for (4.13) it suffices that $J(h_{1,n}) = o(n^{(2k-1)/(4k+2)})$, i.e. $J(h_{1,n})$ may converge to infinity. This means that weaker conditions on the smoothness of $h_{1,n}$ are needed than on $g_{0,n}$. Furthermore, if $J(h_{1,n}) \rightarrow \infty$, \sqrt{n} -consistency of $\hat{\theta}_n$ can always be guaranteed by choosing λ_n small (i.e. undersmoothing).

5. Estimating the distribution of the parametric component using Wild Bootstrap. Inference on the parametric component θ of the model could be based on our asymptotic result in Theorem 4.1. There it is stated that the distribution of $\hat{\theta}_n$ is not affected by the nonparametric nature of the other component of the model, at least asymptotically. This statement may be misleading for small sample sizes. An approach which reflects more carefully the influence of the nonparametric component is bootstrap. We discuss here three versions of bootstrap. The first version is Wild Bootstrap which is related to proposals of Wu (1986) (see also Beran (1986) and Mammen (1992)) and which was first proposed by Härdle and Mammen (1993) in nonparametric set ups. Note that in our model the conditional distribution of Y is not specified besides (1.1) and (A0).

The Wild Bootstrap procedure works as follows.

STEP 1. Calculate residuals $\hat{W}_i = Y_i - \hat{\mu}_n(T_i)$.

STEP 2. Generate n i.i.d. random variables $\varepsilon_1^*, \dots, \varepsilon_n^*$ with mean 0, variance 1 and which fulfill for a constant C that $|\varepsilon_i^*| \leq C$ (a.s.) for $i = 1, \dots, n$.

STEP 3. Put $Y_i^* = \hat{\mu}_n(T_i) + \hat{W}_i \varepsilon_i^*$ for $i = 1, \dots, n$.

STEP 4. Use the (pseudo) sample $((Y_1^*, T_1), \dots, (Y_n^*, T_n))$ for the calculation of the parametric estimate $\hat{\theta}_n^*$.

STEP 5. The distribution \mathcal{L}_n of $\hat{\theta}_n - \theta_0$ is estimated by the (conditional) distribution \mathcal{L}_n^* (given $(Y_1, T_1), \dots, (Y_n, T_n)$), of $\hat{\theta}_n^* - \hat{\theta}_n$.

Under the additional model assumption

$$\text{var}_0(Y|T=t) = V(\mu_0(t))$$

we propose the following modification of the resampling. In Step 3 put $Y_i^* = \hat{\mu}_n(T_i) + V(\hat{\mu}_n(T_i))\varepsilon_i^*$ for $i = 1, \dots, n$. In this case the condition that $|\varepsilon_i^*|$ is bounded can be weakened to the assumption that ε_i^* has subexponential tails, i.e. for a constant C it holds that $E(e^{|\varepsilon_i^*|/C}) \leq C$ for $i = 1, \dots, n$ (compare (A0)).

In the special situation that $Q(y; \mu)$ is the log-likelihood (a semiparametric generalized linear model), the conditional distribution of Y_i is specified by $\mu(T_i)$. Then we recommend to generate n independent Y_1, \dots, Y_n with distributions defined by $\hat{\mu}_n(T_1), \dots, \hat{\mu}_n(T_n)$, respectively. This is a version of parametric bootstrap. The following theorem states that these three bootstrap procedures work (for their corresponding models).

Theorem 5.1. *Assume that conditions (A0)-(A4) are met. In case of application of the second or third version of bootstrap assume that the just mentioned additional model assumptions hold. Then*

$$d_K(\mathcal{L}_n, \mathcal{L}_n^*) \rightarrow 0$$

in probability. Here d_K denotes the Kolmogorov distance (i.e. the sup norm of the corresponding distribution functions).

Proof. We will give only a sketch of the proof for the first version of resampling (Wild Bootstrap). The proof for the other versions is more simple and follows similarly.

We have to go again through the proofs of Lemma 3.1 and Theorem 4.1. We start with proving

$$(5.1) \quad \|\hat{g}_n^* - \hat{g}_n\|_n = O_{\mathbf{P}}(\lambda_n)$$

and

$$(5.2) \quad J(\hat{g}_n^*) = O_{\mathbf{P}}(1).$$

We write first for $W_i^* = Y_i^* - \hat{\mu}_n(T_i)$

$$\begin{aligned} W_i^* &= W_i \varepsilon_i^* + (\mu_0(T_i) - \hat{\mu}_n(T_i))\varepsilon_i^* \\ &= W_{i,1}^* + W_{i,2}^*. \end{aligned}$$

In the proof of Lemma 3.1 the main ingredient from empirical process theory was formula (2.4) (see (3.7)). We argue now that the following analogue formulas hold for $j = 1$ and $j = 2$

$$(5.3) \quad \sup_{a \in \mathcal{A}} \frac{(1/n) \sum_{i=1}^n W_{i,j}^* a(T_i)}{\|a\|_n^{1-\nu/2}} = O_{\mathbf{P}}(n^{-1/2}).$$

For $j = 1$ equation (5.3) follows from the fact that because of the boundedness of ε_i^* for $i = 1, \dots, n$, we have that there exists a constant C' with

$$E_0(e^{||W_{i,1}^*|/C'} | T_1, \dots, T_n) \leq C',$$

almost surely.

For $j = 2$ we have for every constant C'' that on the event $A_n = \{|\mu_0(T_i) - \hat{\mu}_n(T_i)| \leq C'' : i = 1, \dots, n\}$ the following holds

$$E_0(e^{[|W_{i,2}^*|/(CC'')]})|T_1, \dots, T_n) \leq e,$$

almost surely. Because the probability of A_n tends to one, we arrive at (5.3).

We would like to make here the following remark for two random variables U_n and V_n . If U_n fulfils $U_n = O_{\mathbf{P}}(c_n)$ for a sequence c_n then this implies that for every $0 < \delta < 1$ there exists a set B_n and a constant C with

$$\mathbf{P}(V_n \in B_n) > 1 - \delta,$$

$$\mathbf{P}(|U_n| \leq Cc_n | V_n = v) > 1 - \delta$$

for $v \in B_n$. This remark may help to understand why we can continue as in the proof of Lemma 3.1 to show (5.1) and (5.2).

The next step is to show that

$$(5.4) \quad \sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n) = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n W_i^* l_0(T_i) h_2(T_i)}{\|(f_0 l_0)^{1/2} h_2\|^2} + o_{\mathbf{P}}(1).$$

For seeing (5.4) we proceed similarly as in the proof of Theorem 4.1. In particular, we replace $\hat{g}_{n,s}$ by $\hat{g}_{n,s}^* = \hat{g}_n^* + sh_2$.

Now one applies (5.4) for the proof of

$$d_K(N(0, n\hat{\sigma}_n^2), \mathcal{L}_n^*) \rightarrow 0,$$

(in probability), where

$$\hat{\sigma}_n^2 = \frac{\frac{1}{n} \sum_{i=1}^n W_i^2 l_0(T_i)^2 h_2(T_i)^2}{\|(f_0 l_0)^{1/2} h_2\|^4}.$$

Because of $\hat{\sigma}_n^2 \rightarrow \|(f_0 l_0)^{1/2} h_2\|^{-4} E_0(W_i^2 l_0(T_i)^2 h_2(T_i)^2)$ (in probability) we get the statement of the theorem. \square

6. Examples.

Example 1. Recall that in this case,

$$Y = \theta_0 X + m_0(Z) + W,$$

where $E_0(W|X, T) = 0$, and that $\hat{g}_n(x, z) = \hat{\theta}_n x + \hat{m}_n(z)$ is the penalized least squares estimator. In van de Geer (1990), Lemma 3.1 has been proved under the condition (AA0) that the error W in the regression model is sub-Gaussian, using the same approach as in the proof of Lemma 3.1. Condition (AA0) can be relaxed to (A0), as a consequence of Theorem 2.2. This is in accordance with earlier results on rates of convergence (see e.g. Rice and Rosenblatt (1981) and Silverman (1985)).

Conditions (A1)-(A4) are clearly satisfied, because $V \equiv 1$, $f \equiv 1$ and $l \equiv 1$. Note further that $h_1 = \tilde{h}_1$ and $h_2 = \tilde{h}_2$. If W is normally distributed, then according to Theorem 4.1, the partial smoothing spline estimator $\hat{\theta}_n$ is an asymptotically efficient estimator of θ_0 .

Example 2. In this case, we have

$$P_0(Y = 1|X, Z) = 1 - P_0(Y = 0|X, Z) = F(\theta_0 X + m_0(Z)),$$

and $V(s) = s(1 - s)$, $s \in (0, 1)$. Let us consider the common choice

$$F(\xi) = \frac{e^\xi}{1 + e^\xi}, \quad \xi \in \mathbf{R}.$$

Then

$$f(\xi) = \frac{e^\xi}{(1 + e^\xi)^2} = V(F(\xi)), \quad \xi \in \mathbf{R},$$

so that $l \equiv 1$. We cannot use Lemma 3.1, because (A1) is not satisfied. Therefore, we present a separate proof of (3.1) and (3.2) in the next section. Since conditions (A3) and (A4) are met, Theorem 4.1 can then be applied.

Example 3. Let us assume that the conditional density of Y given $T = t$ is

$$p_0(y|t) = \lambda_0(t)e^{-\lambda_0(t)y}, \quad y > 0,$$

with $\lambda_0(t) = 1/\mu_0(t)$, $\mu_0(x, z) = F(\theta_0 x + m_0(z))$, and with

$$F(\xi) = e^\xi, \quad \xi \in \mathbf{R}.$$

Take $V(s) = 1/s^2$, $s > 0$. Then

$$f(\xi) = e^\xi,$$

$$V(F(\xi)) = e^{-2\xi},$$

and

$$l(\xi) = e^{-\xi},$$

$\xi \in \mathbf{R}$. Observe that (A0) is met. Again, we cannot apply Lemma 3.1, because (A1) and (A2) only hold on a bounded set. So if we show by separate means that the parameters are in a bounded set, then the result of Lemma 3.1 follows immediately. Conditions (A3) and (A4) hold, so asymptotic normality would also be implied by this. Note that $fl \equiv 1$, so as in Example 1, $h_1 = \tilde{h}_1$ and $h_2 = \tilde{h}_2$.

7. Rates of convergence for Example 2. Consider the model

$$P_0(Y = 1|X, Z) = 1 - P_0(Y = 0|X, Z) = F(\theta_0 X + m_0(Z)) = F(g_0(T)),$$

with

$$g_0 \in \mathcal{G} = \{g(x, z) = \theta x + m(z), \theta \in \mathbf{R}, J(m) < \infty\},$$

and $F : \mathbf{R} \rightarrow (0, 1)$ given. Furthermore, take $V(s) = s(1 - s)$, $s \in (0, 1)$. Assume that for some $0 < C_5 < \infty$,

$$(A5) \quad |f(\xi)| \leq C_5, \text{ for all } \xi \in \mathbf{R}.$$

Lemma 7.1. *Under condition (A5), we have*

$$\sup_n \sup_{\delta > 0} \delta^{1/k} H(\delta, \left\{ \frac{F(g)}{1 + J(g)} : g \in \mathcal{G} \right\}, \|\cdot\|_n) < \infty.$$

Proof. We can write for $g \in \mathcal{G}$,

$$g = \beta' \phi + g_2,$$

with $\beta \in \mathbf{R}^{k+1}$, and $|g_2|_\infty \leq J(g_2) = J(g)$ (see Section 2.2). Now, let \tilde{g} be a fixed function and consider the class

$$\{F(\beta' \phi + \tilde{g}) : \beta \in \mathbf{R}^{k+1}\}.$$

Since F is of bounded variation, the collection of graphs

$$\{(s, t) : 0 \leq s \leq F(\beta' \phi(t) + \tilde{g}(t))\} : \beta \in \mathbf{R}^{k+1}\}$$

is a Vapnik-Chervonenkis class, i.e. it forms a polynomial class of sets (see Pollard (1984, Chapter II) for definitions). Therefore (Pollard (1984, Lemma II.25)),

$$(7.1) \quad N(\delta, \{F(\beta' \phi + \tilde{g}) : \beta \in \mathbf{R}^{k+1}\}, \|\cdot\|_n) \leq A\delta^{-w}, \text{ for all } \delta > 0,$$

where the constants A and w depend on F and k , but not on \tilde{g} and n . (Here, we use the fact that the class is uniformly bounded by 1.)

Define for $g = \beta' \phi + g_2$,

$$\nu(g) = \left[\frac{1}{(1 + J(g))\delta} \right] \delta,$$

where $[s]$ denotes the integer part of $s \geq 0$. Then

$$\{\nu(g)g_2\} \subset \{h : |h|_\infty \leq 1, J(h) \leq 1\},$$

so by Theorem 2.1,

$$(7.2) \quad \sup_{\delta > 0} \delta^{1/k} H(\delta, \{\nu(g)g_2\}, |\cdot|_\infty) < \infty.$$

Of course, if we replace here the $|\cdot|_\infty$ -norm by the $\|\cdot\|_n$ -norm, the result remains true and holds uniformly in n .

Together, (7.1) and (7.2) give the required result. To see this, let $g \in \mathcal{G}$, $g = \beta' \phi + g_2$, and let $\nu_j = \nu(g)$. Suppose that h_j is such that

$$\| \nu(g)g_2 - h_j \|_n \leq \delta,$$

and that β_j is such that

$$\| F(\beta' \phi + \frac{h_j}{\nu_j}) - F(\beta'_j \phi + \frac{h_j}{\nu_j}) \|_n \leq \delta.$$

Then

$$\begin{aligned} & \left\| \frac{F(\beta' \phi + g_2)}{1 + J(g)} - F(\beta'_j \phi + \frac{h_j}{\nu_j}) \nu_j \right\|_n \\ & \leq \nu_j \| F(\beta' \phi + g_2) - F(\beta' \phi + \frac{h_j}{\nu_j}) \|_n + \left| \frac{1}{1 + J(g)} - \nu_j \right| \| F(\beta' \phi + \frac{h_j}{\nu_j}) - F(\beta'_j \phi + \frac{h_j}{\nu_j}) \|_n \\ & \leq C_5 \delta + \delta + \delta, \end{aligned}$$

since $|F(\xi) - F(\tilde{\xi})| \leq C_5 |\xi - \tilde{\xi}|$, by condition (A5). \square

The entropy result of Lemma 7.1 can be applied to establish a rate of convergence in the same way as in Lemma 3.1. For this purpose, we need the assumption: for some constant $0 < C_6 < \infty$,

$$(A6) \quad \frac{1}{C_6} \leq F(g_0(t)) \leq 1 - \frac{1}{C_6}, \text{ for all } t \in [0, 1]^2.$$

Lemma 7.2. *Suppose (A5) and (A6) hold true. Then*

$$(7.3) \quad \| F(\hat{g}_n) - F(g_0) \|_n = O_{\mathbf{P}}(\lambda_n),$$

and

$$(7.4) \quad J(\hat{g}_n) = O_{\mathbf{P}}(1).$$

Proof. Define

$$\bar{F}(g) = (F(g) + F(g_0))/2, \quad g \in \mathcal{G}.$$

By the concavity of the log-function, and the definition of \hat{g}_n ,

$$\begin{aligned} \bar{Q}_n(\bar{F}(\hat{g}_n)) - \bar{Q}_n(F(g_0)) &= \frac{1}{n} \sum_{i=1}^n Y_i \log\left(\frac{\bar{F}(\hat{g}_n(T_i))}{F(g_0(T_i))}\right) + \frac{1}{n} \sum_{i=1}^n (1 - Y_i) \log\left(\frac{1 - \bar{F}(\hat{g}_n(T_i))}{1 - F(g_0(T_i))}\right) \\ (7.5) \quad &\geq \frac{1}{2} \bar{Q}_n(F(\hat{g}_n)) - \frac{1}{2} \bar{Q}_n(F(g_0)) \geq \frac{1}{2} \lambda_n^2 (J^2(\hat{g}_n) - J^2(g_0)). \end{aligned}$$

On the other hand, since $\log(s) = 2\log(\sqrt{s}) \leq 2(\sqrt{s} - 1)$,

$$\begin{aligned}
\bar{Q}_n(\bar{F}(\hat{g}_n)) - \bar{Q}_n(F(g_0)) &\leq \frac{2}{n} \sum_{i=1}^n Y_i \left(\sqrt{\frac{\bar{F}(\hat{g}_n(T_i))}{F(g_0(T_i))}} - 1 \right) + \frac{2}{n} \sum_{i=1}^n (1 - Y_i) \left(\sqrt{\frac{1 - \bar{F}(\hat{g}_n(T_i))}{1 - F(g_0(T_i))}} - 1 \right) \\
(7.6) \quad &= \frac{2}{n} \sum_{i=1}^n \frac{W_i}{\sqrt{F(g_0(T_i))}} \left(\sqrt{\bar{F}(\hat{g}_n(T_i))} - \sqrt{F(g_0(T_i))} \right) \\
&\quad + \frac{2}{n} \sum_{i=1}^n \frac{W_i}{\sqrt{1 - F(g_0(T_i))}} \left(\sqrt{1 - \bar{F}(\hat{g}_n(T_i))} - \sqrt{1 - F(g_0(T_i))} \right) \\
&\quad - \left\| \sqrt{\bar{F}(\hat{g}_n)} - \sqrt{F(g_0)} \right\|_n^2 - \left\| \sqrt{1 - \bar{F}(\hat{g}_n)} - \sqrt{1 - F(g_0)} \right\|_n^2.
\end{aligned}$$

The combination of (7.5) and (7.6) gives an inequality of the same form as inequality (3.8) in the proof of Lemma 3.1. Moreover, we can invoke Lemma 7.1 in Theorem 2.2. First of all, condition (AA0) holds for W . Furthermore, for each $g, \tilde{g} \in \mathcal{G}$ we have

$$\frac{|\sqrt{\bar{F}(g)} - \sqrt{\bar{F}(\tilde{g})}|}{\sqrt{F(g_0)}} \leq \frac{|F(g) - F(\tilde{g})|}{2\sqrt{2}F(g_0)} \leq \frac{C_6}{2\sqrt{2}} |F(g) - F(\tilde{g})|,$$

by (A6). So the entropy condition (2.3) with $\nu = 1/k$ holds for the class

$$\left\{ \frac{\sqrt{\bar{F}(g)} - \sqrt{\bar{F}(g_0)}}{\sqrt{F(g_0)}(1 + J(g))} : g \in \mathcal{G} \right\}.$$

Thus,

$$\frac{\frac{1}{n} \sum_{i=1}^n W_i (\sqrt{\bar{F}(\hat{g}_n(T_i))} - \sqrt{F(g_0(T_i))}) / \sqrt{F(g_0(T_i))}}{\left\| \sqrt{\bar{F}(\hat{g}_n)} - \sqrt{F(g_0)} \right\|_n^{1-1/(2k)} (1 + J(\hat{g}_n))^{1/(2k)}} = O_{\mathbf{P}}(n^{-1/2}).$$

Similar results can be derived for $(\sqrt{1 - \bar{F}(\hat{g}_n)} - \sqrt{1 - F(g_0)})$. So, proceeding as in the proof of Lemma 3.1, we find $J(\hat{g}_n) = O_{\mathbf{P}}(1)$, and

$$(7.7) \quad \left\| \sqrt{\bar{F}(\hat{g}_n)} - \sqrt{F(g_0)} \right\|_n = O_{\mathbf{P}}(\lambda_n),$$

as well as

$$(7.8) \quad \left\| \sqrt{1 - \bar{F}(\hat{g}_n)} - \sqrt{1 - F(g_0)} \right\|_n = O_{\mathbf{P}}(\lambda_n).$$

Clearly, (7.7) and (7.8) yield (7.3). □

It remains to show that the rate of convergence also holds for $\|\hat{g}_n - g_0\|_n$. We then need an identifiability condition. Assume that for some constants $0 < \eta_0, C_7 < \infty$ and for all $t \in [0, 1]^2$, we have for $\xi_0 = g_0(t)$,

$$(A7) \quad |f(\xi)| \geq \frac{1}{C_7} \text{ for all } |\xi - \xi_0| \leq \eta_0.$$

Lemma 7.3. *Suppose that*

$$(7.9) \quad \inf_{\|g - g_0\| > \eta} \|F(g) - F(g_0)\| > 0 \text{ for all } \eta > 0.$$

Then, under conditions (A5),(A6),(A7),(4.3) and (4.4), we have

$$\|\hat{g}_n - g_0\|_n = O_{\mathbf{P}}(\lambda_n).$$

Proof. Due to Lemma 7.1 and a result of e.g. Pollard (1984, Theorem II.24) on uniform laws of large numbers, we have for all $0 < C < \infty$,

$$\sup_{J(g) \leq C} | \|F(g) - F(g_0)\|_n - \|F(g) - F(g_0)\| | = o(1), \text{ almost surely.}$$

So $\|F(\hat{g}_n) - F(g_0)\| = o_{\mathbf{P}}(1)$. By (7.9), this implies

$$(7.10) \quad \|\hat{g}_n - g_0\| = o_{\mathbf{P}}(1).$$

As in the proof of Theorem 4.1, we see that (4.3) and (4.4), together with (7.10), yield $|\hat{g}_n - g_0|_{\infty} = o_{\mathbf{P}}(1)$. Application of (A7) and Lemma 7.2 completes the proof. \square

8. References.

- [1] BERAN, R. (1986). Comment on "Jackknife, bootstrap, and other resampling methods in regression analysis" by C. F. J. Wu. *Ann. Statist.* **14**, 1295-1298.
- [2] BIRGÉ, L. and MASSART, P. (1991). Rates of convergence for minimum contrast estimators. *Techn. Report 140*, Université Paris 6.
- [3] BIRMAN, M. Š. and SOLOMJAK, M.J. (1967). Piece-wise polynomial approximations of functions of the classes W_p^{α} . *Mat. Sbornik* **73**, 295-317.
- [4] CHEN, H. (1988). Convergence rates for parametric components in a partly linear model. *Ann. Statist.* **16**, 136-146.
- [5] DUDLEY, R.M. (1984). A course on empirical processes. *Ecole d'Eté de Probabilités de St. Flour, 1982, Lecture Notes in Math.* 1-122. Springer, Berlin.
- [6] FAN, J. , HECKMAN, N. E. and WAND, M. P. (1995). Local polynomial regression for generalized linear models and quasi-likelihood functions. *J. Amer. Statist. Assoc.* **90**, 141-150.
- [7] GINÉ, E. and ZINN, J. (1984). On the central limit theorem for empirical processes. *Ann. Probab.* **12**, 929-989.

- [8] GREEN, P. J. and YANDELL, B.(1985). Semi-parametric generalized linear models. in *Proceedings 2nd international GLIM conference. Lecture Notes in Statistics* **32**, Springer Verlag, New York, 44-55.
- [9] GREEN, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *Int. Statist. Rev.* **55**, 245-60.
- [10] HÄRDLE, W. and MAMMEN, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist.* **21**, 1926-1947.
- [11] MAMMEN, E. (1992). When does bootstrap work: asymptotic results and simulations. *Lecture Notes in Statist.* **77**, Springer, Berlin.
- [12] McCULLAGH, P. and NELDER, J.A. (1989). *Generalized Linear Models*. (2nd edit.) Chapman and Hall, London.
- [13] ODEN, J.T. and REDDY, J.N. (1976). *An Introduction to the Mathematical Theory of Finite Elements*. Wiley, New York.
- [14] OSSIANDER, M. (1987). A central limit theorem under metric entropy with L_2 bracketing. *Ann. Probab.* **15**, 897-919.
- [15] POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.
- [16] POLLARD, D. (1990). *Empirical Processes: Theory and Applications*. NSF-CBMS Regional Conference Series in Probability and Statistics 2.
- [17] RICE, J. and ROSENBLATT, M. (1981). Integrated mean square error of a smoothing spline. *J. Approx. Theory* **33**, 353-369.
- [18] RICE, J. (1986). Convergence rates for partially splined models. *Statist. Probab. Letters* **4**, 203-208.
- [19] SEVERINI, T.A. and STANISWALIS, J.G. (1994). Quasi-likelihood estimation in semi-parametric models. *J. Amer. Statist. Ass.* **89**, 501-511.
- [20] SILVERMAN, B.W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *J. Roy. Statist. Soc. Ser. B* **47**, 1-52.
- [21] SPECKMAN, P. (1988). Kernel smoothing in partial linear models. *J. Roy. Statist. Soc. Ser. B* **50**, 413-436.
- [22] VAN DE GEER, S. (1988). *Regression Analysis and Empirical Processes*. CWI tract **45**, Centre for Mathematics and Computer Science, Amsterdam.
- [23] VAN DE GEER, S. (1990). Estimating a regression function. *Ann. Statist.* **18**, 907-924.
- [24] VAN DE GEER, S. (1995). A maximal inequality for the empirical process. *Techn. Report TW 95-05*, University of Leiden.
- [25] WAHBA, G. (1984). Partial spline models for the semi-parametric estimation of functions of several variables. In *Statistical Analysis of Time Series* 319-329. Institute of Statistical Mathematics, Tokyo.
- [26] WEDDERBURN, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss - Newton method. *Biometrika* **61**, 439-447.
- [27] WU, C F. J. (1986). Jackknife, bootstrap, and other resampling methods in regression analysis. *Ann. Statist.* **14**, 1261-1350.