Moment Estimation in a Semiparametric Generalized Linear Model

Xueqin Wang [a], Hanxiang Peng [b,*]

[a] *School of Mathematics & Computational Science, Zhongshan School of Medicine, Sun Yat-Sen University, P.R. China*

[b] *Department of Mathematics, University of Mississippi, University, MS 38677-1848, USA*

Abstract

In this article, we propose to estimate the regression parameters in a semiparametric generalized linear model by moment estimating equations. These estimators are shown to be consistent and asymptotically normal. We present two estimators of the nonparametric part, provide conditions for the existence and uniform consistency, and obtain faster rates of convergence under weaker assumptions.

Key words: Exponential family, Generalized linear model, Kernel estimate, Moment estimate, Nonparamtric part, Semiparametric regression.

1 Introduction

In a semiparametric generalized linear model(SGLM), the response $Y \in \mathbb{R}$ and the covariate vector $(X, Z) \in [c, d]^m \times [0, 1]$ satisfy the structural relation:

$$\mathbb{E}(Y|X, Z) = h(X^\top \theta + \rho(Z)), \quad \theta \in \Theta \subset \mathbb{R}^m, \tag{1}$$

where Θ is a nonempty and open subset, ρ is an unknown nonparametric smooth function and h is a link. The density f of Y w.r.t. σ-finite measure ν belongs to an exponential family of the form

$$f(y|\phi) = \exp(\phi y - b(\phi)), \quad y \in \mathcal{Y} \subset \mathbb{R}, \quad \phi \in \Phi,$$

* Corresponding author.
Email address: mmpeng@olemiss.edu (Hanxiang Peng).

where Φ is the range of the composite $a = \mu^{-1} \circ h$ of the inverse μ^{-1} of the expectation μ of Y and the link h, assuming that μ is invertible. The covariate (X, Z) is random and has a joint distribution G. It is worth noting that the above form of the exponential family is *canonical*; other forms can be transformed to this by reparametrization.

Suppose now that we have independent observations $(Y_i, X_i, Z_i), i = 1, ..., n$ from (Y, X, Z). We are interested in estimating the regression parameter θ in the presence of the nuisance nonparametric parameter ρ. Our method is based on moment estimating equations(MEEs). Moment estimators in certain situations have simple structures such as explicit formulas, but they are not efficient in general. However, they can be improved to attain efficiency by the method of scoring. See e.g. Bickel, *et al.*, 1993, page 44. For estimating the Euclidean parameter θ, we have to estimate the nonparametric part ρ. We shall give two estimators of ρ based on the methods of moment and maximum likelihood. Severini and Wong(1992) proposed the profile likelihood procedure for semiparametric models. By this procedure, estimators of the nonparametric part can be obtained. These estimators and the partial derivatives converge at certain rates (see (18) below) under smoothness assumptions, involving high order partial derivatives of the densities. Forrester, *et al.*(2003) presented their estimator of the nonparametric part in partially linear model under weaker smoothness assumptions. Sparked from their work, we propose our estimators and obtain faster rates under weaker assumptions. See (18) and (23) below.

In the sequel, we shall denote θ the true unknown parameter and $\vartheta \in \Theta$ a generic parameter. We shall write $\mathbb{E}_\vartheta$ the expectation calculated under the probability measure $P_{\vartheta,\rho}$ for $\vartheta \in \Theta$, and $P = P_{\theta,\rho}$ and $\mathbb{E} = \mathbb{E}_\theta$. Throughout we reserve $\phi = a(X^\top\theta + \rho(Z))$. The rest of the article is organized as follows. In Section 2, we introduce the estimators. Examples are given. Section 3 contains the main theorems. The technical proofs are given in Section 4.

2 Moment Estimating Equations

In this section, we introduce the estimators and give examples. We assume throughout that Φ is a convex subset of the interior of the *natural parameter space*, consisting of all φ having the finite normalizing function

$$\exp(b(\varphi)) = \int \exp(\varphi y)\, d\nu(y).$$

Hence in Φ, all the derivatives of $b(\varphi)$ exist and all moments of Y are finite and can be computed (see e.g. Brown, 1986, page 34) by the equations:

$$\int y^k \exp(\varphi y))\, d\nu(y) = \partial^k/\partial\varphi^k \exp(b(\varphi)), \quad \varphi \in \Phi, \quad k = 1, 2, \cdots .$$

In particular, the expectation and the variance of Y are $\mu(\varphi) = \mathbb{E}_\varphi Y = b'(\varphi)$ and $\Sigma(\varphi) = \mathrm{Cov}_\varphi(Y) = b''(\varphi)$ respectively. In terms of conditional expectation, these equations can be rewritten as

$$\mathbb{E}(Y^k|X,Z) = \exp(-b(\phi))\partial^k/\partial\phi^k \exp(b(\phi)), \quad k = 1, 2, \cdots . \tag{2}$$

In particular, for $k = 1$, the conditional expectation of Y given Z is

$$\mathbb{E}(Y|Z) = \mathbb{E}\left(\exp(-b(\phi))\partial/\partial\phi \exp(b(\phi))|Z\right) = \mathbb{E}(h(X^\top\theta + \rho(Z))|Z).$$

Replacing the above conditional expectation with the ordinary kernel estimator, we estimate the nonparametric part $\rho(z)$ by a moment-type(M-type) kernel estimator $r = \hat{\rho}_{\theta,M}(z)$, the solution to the MEE:

$$\sum_{i=1}^n Y_i K((Z_i - z)/h_n) = \sum_{i=1}^n h(X_i^\top\theta + r)K((Z_i - z)/h_n), \ z \in [0,1] \tag{3}$$

where $h_n > 0$ is a bandwidth and K is a kernel satisfying the usual assumption.

Assumption 1 *The kernel K is a bounded Lipschitz probability density having support $[-1,1]$.*

Taking expectation across (2) gives

$$\mathbb{E}(Y^k) = \mathbb{E}\left(\exp(-b(\phi))\partial^k/\partial\phi^k \exp(b(\phi))\right), \quad k = 1, 2, \cdots .$$

In order to obtain the moment equations for estimating θ, we shall replace the $\rho(Z)$ in $\phi = X^\top\theta + \rho(Z)$ with an estimator $\hat{\rho}_\theta(Z)$, and replace the moments with their sample moments in the above equations. To this end, set $\phi_{n,i}(\theta) = a(X_i^\top\theta + \hat{\rho}_\theta(Z_i))$, where $\hat{\rho}_\theta$ is an estimator of ρ (it could be $\hat{\rho}_{\theta,M}$, or $\hat{\rho}_{\theta,ML}$ below, or any other estimator). Our proposed estimator $\hat{\theta}_n$ of θ is then the solution to the m MEEs:

$$(1/n)\sum_{i=1}^n Y_i = (1/n)\sum_{i=1}^n b'(\phi_{n,i}(\theta)), \tag{4}$$

$$(1/n)\sum_{i=1}^n Y_i^2 = (1/n)\sum_{i=1}^n \left(b'^2(\phi_{n,i}(\theta) + b''(\phi_{n,i}(\theta))\right), \tag{5}$$

$$\cdots\cdots\cdots\cdots\cdots\cdots$$

$$(1/n)\sum_{i=1}^n Y_i^m = (1/n)\sum_{i=1}^n \exp\left(-b(\phi_{n,i})\right)\partial^m/\partial\phi^m \exp\left(b(\phi_{n,i})\right). \tag{6}$$

We note that (3) and (4) may be linearly dependent as is the case when h is a canonical link (so that a is the identity map). In this case, we may simply replace (4) with an additional MEE:

$$(1/n)\sum_{i=1}^n Y_i^{m+1} = (1/n)\sum_{i=1}^n \exp\left(-b(\phi_{n,i})\right)\partial^{m+1}/\partial\phi^{m+1} \exp\left(b(\phi_{n,i})\right). \tag{7}$$

Another approach in coping with the linear dependence is to estimate $\rho(z)$ by a maximum likelihood-type(ML-type) estimator $\hat{\rho}_{\vartheta,ML}(z)$. The log likelihood is $a(\phi)Y - b(\phi)$, so that $r = \hat{\rho}_{\vartheta,ML}(z)$ is the solution to the equation:

$$\frac{1}{nh_n}\sum_{i=1}^{n} a'(X_i^\top\theta + r)\Big(Y_i - h(X_i^\top\theta + r)\Big)K((Z_i - z)/h_n) = 0, \quad z \in [0,1]. \tag{8}$$

Clearly, this last approach only works provided that a is not an identity map (the canonical link), otherwise it simplifies to (3) and the same additional equation (7) may have to be used. Thus, if a is an identity map, we shall resort to MEEs (3), (5)-(7); otherwise (a is not identity map, noncanonical link), we may use equations (8), (4)-(6). Simplifications such as closed formulas or convenient equations are possible when the MEEs are judiciously chosen. For instance, when a is noncanonical and θ is real, one may use equations (3) and (8). More specifically, it is advantageous to employ (8) as an estimating equation for ρ from the perspective that such an estimator is a "least favorable curve". For more details on this, see e.g. Severini and Wong(1992). It is interesting to note that one has options to choose different combinations of MEEs. Such an example is the aforementioned two systems of equations. Another two examples are MEEs (3), (5)-(7) and equations (8), (5)-(7).

Example 1 Consider the canonical link $h(\zeta) = \zeta$ and $b(\varphi) = \varphi^2/2$, so Y has the normal distribution $\mathcal{N}(\varphi, 1)$. The M-type kernel estimator $\hat{\rho}_{\vartheta,M}(z)$ of the nonparametric part is given by

$$\hat{\rho}_{\vartheta,M}(z) = \sum_{i=1}^{n} K_{n,i}(z)Y_i \Big/ \sum_{i=1}^{n} K_{n,i}(z) - \left(\sum_{i=1}^{n} K_{n,i}(z)X_i \Big/ \sum_{i=1}^{n} K_{n,i}(z)\right)^\top \vartheta,$$

where $K_{n,i}(\cdot) = K((Z_i - \cdot)/h_n)$. The ML-type estimator $\hat{\rho}_{\vartheta,ML}(z;\vartheta)$ has the same formula. Since the link is canonical, we have to use the equations (5)-(7) to cope with the linear dependence. The moment estimator $\hat{\theta}_n$ does not have an explicit formula but is the solution to the following MEEs:

$$(1/n)\sum_{i=1}^{n} Y_i^k = (1/n)\sum_{i=1}^{n} \exp\Big(-\phi_{n,i}^2(\theta)/2\Big)\partial^k/\partial\phi^k \exp(\phi_{n,i}^2(\theta)/2))$$

for $k = 2, \cdots, m+1$, where $\phi_{n,i}(\theta) = X_i^\top\theta + \hat{\rho}_\theta(Z_i)$.

Example 2 For the Gamma distribution with mean μ and unit shape parameter, the k^{th} moment is $k!\mu^k$. For link $h(\eta) = e^\eta$, the M-type kernel estimator of the nonparametric part has an explicit formula given by

$$\hat{\rho}_{\vartheta,M}(z) = \log\Big[\sum_j K_{n,j}(z)Y_j \Big/ \sum_j K_{n,j}(z)e^{X_j^\top\vartheta}\Big].$$

The moment estimator $\hat{\theta}_n$ is the solution to the m MEEs:

$$(1/n)\sum_{i=1}^{n} Y_i^k = (1/n)\sum_{i=1}^{n} k!\exp(X_i^\top\theta + \hat{\rho}_\theta(Z_i)), \quad k = 2, \cdots, m+1.$$

Example 3 For the linear Poisson regression with $b(\varphi) = \exp(\varphi)$ and the noncanonical link $h(\zeta) = \exp(\zeta)$, the M-type kernel estimator $\hat{\rho}_{\vartheta,M}(z)$ of the nonparametric part has an explicit formula given by

$$\hat{\rho}_{\vartheta,M}(z) = \log\Big[\sum_i K_{n,i}(z)Y_i \Big/ \sum_i K_{n,i}(z)\exp(X_i^\top\vartheta)\Big], \quad z \in [0,1].$$

The ML-type kernel estimator $\hat{\rho}_{\vartheta,ML}(z)$ does not have an explicit formula but is the solution to the nonlinear equations (w.r.t. r):

$$\sum_i K_{n,i}(z)Y_i/(X_i^\top\vartheta + r) = \sum_i K_{n,i}(z), \quad z \in [0,1].$$

Since we may use $\hat{\rho}_{\vartheta,ML}(z)$ as an estimator $\hat{\rho}_\vartheta(z)$ of ρ, the moment estimator $\hat{\theta}_n$ may be taken as the solution to the MEEs (4)-(6). Specifically, it is the solution to the following MEEs, with $\phi_{n,i}(\theta) = X_i^\top\theta + \hat{\rho}_\theta(Z_i)$,

$$(1/n)\sum_{i=1}^{n} Y_i^k = (1/n)\sum_{i=1}^{n}\exp(\phi_{n,i}(\theta))\sum_{j=0}^{k-2}\binom{k-1}{i}\mu_i, \quad k = 1, \cdots, m.$$

where $\mu_0 = 1$, $\mu_1 = \exp(\phi_{n,i}(\theta))$, $\mu_2 = \exp(\phi_{n,i}(\theta)) + \exp(2\phi_{n,i}(\theta))$, and so forth. Clearly, we may also use $\hat{\rho}_{\vartheta,M}$ as an estimator of ρ. In this case we should use the equations (5)-(7) to avoid the linear dependence.

For convenience, let us focus on the MEEs (5)-(7), while $\hat{\rho}_\theta$ is an estimator of ρ. Other combinations of MEEs may be analogously considered. Let $A(y) = (y^2, \cdots, y^{m+1})^\top$ and $B(\phi)$ be a m-dimensional vector with components $B_1(\phi) = b'^2(\phi) + b''(\phi)$, $B_2(\phi) = b'^3(\phi) + 3b'(\phi)b''(\phi) + b'''(\phi)$, $\cdots$, and $B_m(\phi) = b'^{m+1}(\phi) + \cdots + b^{(m+1)}(\phi)$. Then (5)-(7) can be written as:

$$\Lambda_n(\theta) \equiv (1/n)\sum_{i=1}^{n}\Big(A(Y_i) - B\Big(a(X_i^\top\theta + \hat{\rho}_\theta(Z_i))\Big)\Big) = 0. \tag{9}$$

Equation (9) may have many solutions as in usual MEEs. We can only prove asymptotic existence, consistency and normality for a sequence of solutions. We focus on a compact neighborhood $N(\theta)$ of θ in which $\Lambda_n(\vartheta)$ has at most one zero point for $\vartheta \in N(\theta)$. Define $\hat{\theta}_n$ to be such a zero point if it exists otherwise define it to be an arbitrary number inside $N(\theta)$.

3 Asymptotic Behaviors

In this section, we introduce the conditions for the asymptotic behavior of the estimators, followed by the main theorem. We then study the asymptotic behavior of the estimators of the nonparametric part. Write $\|v\|_E$ for the Euclidean norm of vector $v \in \mathbb{R}^m$; denote $\sup_\vartheta$ for $\sup_{\vartheta\in\Theta}$, $\sup_z$ for $\sup_{z\in[0,1]}$ and so on if the definitions are clear from the context. For a function k from a metric space $\mathbb{M}$ into the reals $\mathbb{R}$, denote the supremum $\|k\|_{\mathbb{M}} = \sup_{m\in\mathbb{M}} |k(m)|$ and write $\|k\|$ for $\|k\|_{\mathbb{M}}$ if there is no ambiguity from the context.

Asymptotic Normality of $\hat{\theta}_n$. In our mind, $\hat{\rho}_n$ is either $\hat{\rho}_{\vartheta,M}$ or $\hat{\rho}_{\vartheta,ML}$. But in what follows, we shall keep $\hat{\rho}_n$ in general and introduce conditions that ensure the asymptotic behavior of the estimator. For $\vartheta \in N(\theta)$ and $z \in [0,1]$, let $\rho_\vartheta(z)$ be the limit of $\hat{\rho}_\vartheta(z)$ in probability as n tends to infinity, fulfilling the following assumption.

Assumption 2 *For $\vartheta \in N(\theta)$ and $z \in [0,1]$, $\rho_\vartheta(z)$ is bounded and the partial derivatives $\rho'_\vartheta(z) = (\partial/\partial\vartheta)\rho_\vartheta(z)$ and $\hat{\rho}'_\vartheta(z)$ (bounded) exist, are continuous and G-square integrable, and satisfy*

$$\sup_{\vartheta\in N(\theta)} \|\hat{\rho}_\vartheta - \rho_\vartheta\| = o_{\mathrm{p}}(1), \quad n \to \infty, \quad \textit{and} \tag{10}$$

$$\sup_{\vartheta\in N(\theta)} \|\hat{\rho}'_\vartheta - \rho'_\vartheta\| = o_{\mathrm{p}}(1), \quad n \to \infty. \tag{11}$$

The next assumption gives the interchange of differentiation and integration.

Assumption 3 *a is twice continuously differentiable satisfying*

$$\mathbb{E} \sup_{\|r\|\le r_0} \sup_{\vartheta\in N(\theta)} |a'(X^\top\vartheta + \rho_\vartheta(Z) + r)| < \infty, \quad \textit{and} \tag{12}$$

$$\mathbb{E} \sup_{\|r\|\le r_0} \sup_{\vartheta\in N(\theta)} |a''(X^\top\vartheta + \rho_\vartheta(Z) + r)| < \infty. \tag{13}$$

Set $D(\vartheta) = \mathbb{E}[(\partial B/\partial\vartheta^\top)(a(X^\top\vartheta + \rho_\vartheta(Z)))], \vartheta \in \Theta$. Then by the chain rule,

$$D(\vartheta) = \mathbb{E}[B'(a(X^\top\vartheta + \rho_\vartheta(Z))a'(X^\top\vartheta + \rho_\vartheta(Z))(X^\top + \rho'^\top_\vartheta(Z))].$$

Assumption 4 *For $\vartheta \in N(\theta)$, the total derivative $(\partial B/\partial\vartheta^\top)(a(X^\top\vartheta+\rho_\vartheta(Z)))$ exists and is continuous, square-integrable, and nonsingular; $D(\vartheta)$ is bounded from below on $N(\theta)$: $\inf_{\vartheta\in N(\theta)} \|D(\vartheta)\|_E > 0$. Further,*

$$\mathbb{E} \sup_{|r|\le r_0} \sup_{\vartheta\in N(\theta)} \|B'\Big(a(X^\top\vartheta + \rho_\vartheta(Z) + r)\Big)\|_E < \infty, \tag{14}$$

$$\mathbb{E} \sup_{|r|\le r_0} \sup_{\vartheta\in N(\theta)} \|B''\Big(a(X^\top\vartheta+\rho_\vartheta(Z)+r)\Big)\|_E < \infty. \tag{15}$$

Theorem 1 *Suppose that Assumptions 2 – 4 hold. Then $\hat{\theta}_n$ is a consistent estimator of θ, i.e., $\hat{\theta}_n \xrightarrow{P} \theta$. Further, it is asymptotically normal,*

$$\sqrt{n}(\hat{\theta}_n - \theta) \Rightarrow \mathcal{N}_m(0, V(\theta)),$$

where $\mathcal{N}_m(v, V)$ denotes the m-dimensional normal distribution with mean $v \in \mathbb{R}^m$ and $m \times m$ covariance matrix V. Here $V(\theta) = D^{-1}(\theta)\,\mathbb{E}[A(Y) - B(a(X^\top\theta + \rho_\vartheta(Z))]^{\otimes 2} D^{-\top}(\theta)$.

Note that Assumptions 3-4 restrict the distribution of the covariates X, Z such that the (partial) derivatives a', a'', B', B'' as functions of the semiparametric part $X^\top\vartheta + \rho_\vartheta(Z)$ are uniformly integrable.

Estimating Nonparametric Part ρ. We now study conditions that ensure the existence and uniform weak consistency of the estimator of the nonparametric part, and give the rates of convergence along the line of Severini and Wong(1992). We assume henceforth that the parameter space Θ is a compact set of $\mathbb{R}^m$ and H is a compact set of $\mathbb{R}$.

For $\vartheta \in \Theta, z \in [0,1]$, let $r = \rho_{\vartheta,M}(z)$ be the unique solution to the equation

$$\mathbb{E}(Y - h(X^\top\vartheta + r)|Z = z) = 0. \tag{16}$$

Let $r = \rho_{\vartheta,ML}(z)$ be the unique solution to the equation

$$\mathbb{E}(a'(X^\top\vartheta + r))(Y - h(X^\top\vartheta + r))|Z = z) = 0, \tag{17}$$

if the solution exists; otherwise define it to be an arbitrary number.

Severini and Wong (1992) proposed the profile likelihood procedure, by which estimators of the nonparametric part can be constructed. They gave the rates of uniform convergence of the estimators. Specifically, they showed

$$\sup_\vartheta \|\hat{\rho}_{\vartheta,ML} - \rho_{\vartheta,ML}\| = o_{\mathrm{p}}(n^{\gamma - q/(2q+4)} h_n^{-(q+4)/(q+2)}) \tag{18}$$

for any $\gamma > 0$. Here $q \ge 2$ is an integer such that the q-th moments of certain statistic exist. They also obtained the rates of uniform convergence for the derivatives of the estimators. Assumption 5 next was employed in their Lemma 5 to assert preliminary uniform consistency. Let $\psi(\vartheta, r, z) = \mathbb{E}\left(a(X^\top\vartheta + r)Y - b(a(X^\top\vartheta + r))|Z = z\right)$, and formally write

$$\psi'(\vartheta, r, z) = (\partial\psi/\partial r)(\vartheta, r, z), \quad \psi''(\vartheta, r, z) = (\partial^2\psi/\partial r^2)(\vartheta, r, z).$$

Their assumption may be stated as follows.

Assumption 5 *For any $\epsilon > 0$, there exists $\delta > 0$ such that*

$$\sup_{\vartheta} \sup_{z} |\psi'(\vartheta, \bar{\rho}_{\vartheta}(z), z)| \leq \delta \Longrightarrow \sup_{\vartheta} \sup_{z} |\bar{\rho}_{\vartheta}(z) - \rho_{\vartheta}(z)| \leq \epsilon.$$

This assumption together with the uniform convergence (20) in Theorem 4 below is essentially a condition of uniform (weak) consistency of the estimator of the nonparametric part. The weak consistency is the basis upon which rates of convergence and asymptotic normality can further be established. Forrester, *et al.*(2003) proposed their estimator of the nonparametric part in partially linear models. We shall follow the idea of the latter to construct our estimator and give conditions which guarantee the existence and uniform consistence. We generalize their result in that their result follows from ours when h assumes the identity link ($h(\eta) = \eta$). We now introduce these conditions. Let

$$S_{\vartheta}(r, y, x) = (y - h(x^{\top}\vartheta + r))a'(x^{\top}\vartheta + r), \quad x \in [c, d]^m, y \in \mathbb{R}.$$

$$D_{\vartheta}(r, z) = \mathbb{E}\left[(\partial S_{\vartheta}/\partial r)(\rho_{\vartheta}(z) + r, Y, X)|Z = z\right], \quad \vartheta \in \Theta, r \in H.$$

Assumption 6 *(i) Monotonicity: $r \mapsto S_{\vartheta}(r, y, x)$ is monotone (incr) for every $\vartheta \in \Theta, x \in [c, d]^m, y \in \mathbb{R}$. Moreover, $\rho_{\vartheta}(z)$ satisfies*

$$\mathbb{E}[S_{\vartheta}(\rho_{\vartheta}(z), Y, X)|Z = z] = 0, \quad \vartheta \in \Theta, z \in [0, 1].$$

(ii) Bounded away from zero: $\inf_{\vartheta \in \Theta} \inf_{|r| \leq r_0} \inf_{z \in [0,1]} |D_{\vartheta}(r, z)| = D_0 > 0$.

To derive the consistency, we need the following smoothness assumptions.

Assumption 7 *(i) The joint density g (w.r.t. Lebesgue measure) of the covariate (X, Z) is Lipschitz in z:*

$$|g(x, z_1) - g(x, z_2)| \leq L_g|z_1 - z_2|, \quad x \in [c, d]^m,\ z_1,\ z_2 \in [0, 1]$$

for some Lipschitz constant $L_g > 0$ independent of x.
(ii) The marginal density q of Z (w.r.t. Lebesgue measure) is bounded away from zero: $q(z) > b_0 > 0, z \in [0, 1]$ for some constant b_0,

Assumption 8 *$z \mapsto \rho_{\vartheta}(z)$ is Lipschitz:*

$$|\rho_{\vartheta}(z_1) - \rho_{\vartheta}(z_2)| \leq L_{\rho}|z_1 - z_2|, \quad z_1, z_2 \in [0, 1], \vartheta \in \Theta$$

for some Lipschitz constant $L_{\rho} > 0$ independent of ϑ.

Assumption 7 (i) and (ii) appeared in Forrester, *et al.*(2003) and are Condition A2 (a) and (b) of Bhattacharya and Zhao (1997). Severini and Wong(1992) imposed, in Lemma 8, assumptions on high order differentiability of the marginal density of Z and the conditional density, the uniform boundedness, and high

order derivatives, see their (c) and (d), page 1787. These assumptions obviously imply (i) and (ii).

Remark 1 *Assumption 7 (i) implies that the marginal density q of Z is Lipschitz and therefore is bounded.*

Assumption 9 next is an identifiability condition under which the parameter θ and the nonparametric ρ are discriminative.

Assumption 9 $\inf_{\vartheta\in\Theta}\inf_{r\in H}\inf_{z\in[0,1]}|\psi''(\vartheta,r,z)|>0.$

Recall $S_\vartheta(r,Y,X)=(Y-h(X^\top\vartheta+r))a'(X^\top\vartheta+r)$. Formally, we write

$$S_\vartheta^{(i,j)}(r,Y,X)=(\partial S_\vartheta^{i+j}/\partial\vartheta^i\partial r^j)(r,Y,X),\quad i=0,\cdots,3,j=0,\cdots,2.$$

The following assumption guarantees the differentiability and the passage to under integrals and uniform integrability.

Assumption 10 *The partial derivatives $S_\vartheta^{(i,j)}(r,y,x),y\in R,x\in[c,d]^m$ exist for $i=0,\cdots,3,j=0,\cdots,2$ and fulfill the following conditions.*

(i) $\sup_\vartheta\sup_{r,s\in H}\int S_\vartheta^{(i,j)}(r,y,x)^2 f(y|a(x^\top\vartheta+s))\,d\nu(y)dG(x,z)<\infty,\ i,j=0,1.$

(ii) *There exists $p\geq 2$ such that* $\sup_{r,\vartheta}\mathbb{E}\,\|S_\vartheta^{(i,j)}(r,Y,X)\|_E^p<\infty,\ i,j=0,1.$

(iii) $\mathbb{E}\sup_{\vartheta,r}\|S_\vartheta^{(i,j)}(r,Y,X)\|_E<\infty,\quad i=0,\cdots,3,j=0,\cdots,2.$

It follows from the Lebesgue dominated convergence theorem that Assumption 10 (i) and (iii) guarantee the passage of differentiation to under the integral signs. We will use this fact without referring to it. Since the above (i) is only used to claim (27) below, we have the following weaker assumption. For more details, see the proof of Theorem 4.

Remark 2 *If $z\mapsto\rho(z)$ is differentiable, then Assumption 10 (i) can be replaced with the following assumption,*

$$\sup_\vartheta\sup_{r,z}\mathbb{E}(S_\vartheta^{(i,j)}(r,Y,X)^2|Z=z)<\infty,\quad i,j=0,1. \tag{19}$$

The proof of uniform convergence in Theorem 2 and Theorem 3 below is similar to Forrester, *et al.*(2003), their proof is essentially in the spirit of Härdle and Luckhaus (1984) and Härdle, *et al.*(1988). We omit the details of the proof.

Theorem 2 *Suppose that Assumptions 6, 7, 8, 10 with $i=j=0$ hold. Then, with $h_n=n^{-p/(4p+12)}$, we have* $\sup_{\vartheta\in\Theta}\|\hat\rho_{\vartheta,ML}-\rho_\vartheta\|=O_P(h_n)$.

Theorem 3 *Suppose that Assumptions 6, 7, 8, and 10 with $i = j = 0$ hold with a being an identity map. Then* $\sup_{\vartheta\in\Theta} \|\hat{\rho}_{\vartheta,M} - \rho_\vartheta\| = O_P(n^{-p/(4p+12)})$.

Theorem 4 next is an analog of Lemma 8 of Severini and Wong(1992). They imposed on the assumptions on the densities and the high order partial derivatives; in their notations, on $f_\theta(y|x)$ and $f(x)$ and the high order partial derivatives, where $f_\theta(y|x)$ is the conditional density of the statistic $T_\theta(Y)$ given X and $f(x)$ is the density of X, which is independent of the parameter θ. What were used in their proof are the assumptions on the densities $f_\theta(x|y)$ and $f_\theta(y)$ and the high order partial derivatives, where $f_\theta(x|y)$ is the conditional density of X given the statistic $T_\theta(Y)$ and $f_\theta(y)$ is the density of $T_\theta(Y)$, which is dependent of the parameter θ. Though the equality $f_\theta(y|x)f(x) = f_\theta(x|y)f_\theta(y)$ may be used, additional assumptions seem required. We give Theorem 4 below, with the proof in the last section.

Theorem 4 *Suppose that Assumptions 7 and 10 hold. Assume Assumption 8 holds with $\vartheta = \theta$. Then, with $\tau_n = n^{-p/(2p+4)}h_n^{-(p+4)/(p+2)}$, we have*

$$\sup_{\vartheta,r,z} |\hat{\psi}'_n(\vartheta, r, z) - \psi'(\vartheta, r, z)| = O_P(h_n + \tau_n). \tag{20}$$

$$\sup_{\vartheta,r,z} \|(\partial/\partial\vartheta)\hat{\psi}'_n(\vartheta, r, z) - (\partial/\partial\vartheta)\psi'(\vartheta, r, z)\|_E = O_P(h_n + \tau_n). \tag{21}$$

$$\sup_{\vartheta,r,z} |\hat{\psi}''_n(\vartheta, r, z) - \psi''(\vartheta, r, z)| = O_P(h_n + \tau_n). \tag{22}$$

We now give the convergence of the ML-type estimator and omit the proof.

Theorem 5 *Suppose that Assumptions 5, 7, 9, and 10 hold. Assume Assumption 8 holds with $\vartheta = \theta$. Then, with τ_n given in Theorem 4, we have*

$$\sup_{\vartheta\in\Theta} \|\hat{\rho}_{\vartheta,ML} - \rho_{\vartheta,ML}\| = O_P(h_n + \tau_n), \ \sup_{\vartheta\in\Theta} \|\hat{\rho}'_{\vartheta,ML} - \rho'_{\vartheta,ML}\| = O_P(h_n + \tau_n). \tag{23}$$

Theorem 4 and Theorem 5 are analogs of Theorems 8, 5 of Severini and Wong (1992). Ours are given in vector parameters. Also we relax the smoothness assumptions of their uniformly boundedness of high order partial derivatives of the density to be Lipschitz continuity, which thus results in the reduction of the convergence rate from their $O_P(h_n^2)$ to our $O_P(h_n)$. We have obtained a sharper rate $O_P(h_n + \tau_n)$, which is their rate when $\gamma = 0$ in their $\gamma > 0$, see the rate on the right hand side of (18). We give the following convergence for the M-type estimator of the nonparametric part with the proof omitted.

Theorem 6 *Suppose that Assumptions 5, 7, 9, and 10 hold; Assumption 8 holds with $\vartheta = \theta$; and with a being replaced with the identity map. Then*

$$\sup_{\vartheta\in\Theta} \|\hat{\rho}_{\vartheta,M} - \rho_{\vartheta,M}\| = O_P(h_n + \tau_n), \quad \sup_{\vartheta\in\Theta} \|\hat{\rho}'_{\vartheta,M} - \hat{\rho}'_{\vartheta,M}\| = O_P(h_n + \tau_n).$$

4 Proofs of Theorem 1 and Theorem 4.

Proof of Theorem 1: By definition, $\hat{\theta}_n$ is the only element in $N(\theta)$ such that $\Lambda_n(\hat{\theta}_n) = 0$. We expand $\Lambda_n(\hat{\theta}_n)$ at the true parameter value θ, so that

$$\hat{\theta}_n - \theta = -\Big(\Lambda_n'(\theta_n^*)\Big)^{-1}\Lambda_n(\theta) \tag{24}$$

for some θ_n^* in between $\hat{\theta}_n$ and θ on an event that $\Lambda_n'(\theta_n^*)$ is invertible. For $\vartheta \in \Theta$, write $\Lambda_n(\vartheta) = \Lambda_{n,1}(\vartheta) + \Lambda_{n,2}(\vartheta)$, where

$$\Lambda_{n,1}(\vartheta) = (1/n)\sum_{i=1}^{n}\Big(A(Y_i) - B\Big(a(X_i^\top\vartheta + \rho_\vartheta(Z_i))\Big)\Big),$$

$$\Lambda_{n,2}(\vartheta) = (1/n)\sum_{i=1}^{n}\Big(B\Big(a(X_i^\top\vartheta + \rho_\vartheta(Z_i))\Big) - B\Big(a(X_i^\top\vartheta + \hat{\rho}_\vartheta(Z_i))\Big)\Big).$$

By the weak law of large numbers and the central limit theorem, we have

$$\Lambda_{n,1}(\theta) = o_{\mathrm{p}}(1), \quad \sqrt{n}\Lambda_{n,1}(\theta) \Longrightarrow \mathcal{N}\Big(0, \mathbb{E}\left(A(Y) - B\Big(a(X^\top\theta + \rho_\theta(Z))\Big)^{\otimes 2}\right).$$

An application of the mean value theorem yields

$$\Lambda_{n,2}(\theta) = (1/n)\sum_{i=1}^{n} B'\Big(a(\eta_{n,i}^*)\Big)a'(\eta_{n,i}^*)(\hat{\rho}_\theta(Z_i) - \rho_\theta(Z_i)),$$

where $\eta_{n,i}^* = X_i^\top\theta + \rho_\theta(Z_i) + u(\hat{\rho}_\theta(Z_i) - \rho_\theta(Z_i))$ for some $u \in [0,1]$. It follows from (10),(12) and (14) that $\Lambda_{n,2}(\theta) = o_{\mathrm{p}}(n^{-1/2})$ and hence

$$\Lambda_n(\theta) = o_{\mathrm{p}}(1), \quad \sqrt{n}\Lambda_n(\theta) \Longrightarrow \mathcal{N}\Big(0, \mathbb{E}\left(A(Y) - B\Big(a(X^\top\theta + \rho_\theta(Z))\Big)^{\otimes 2}\right) \tag{25}$$

By the chain rule of vector functions, we have $\Lambda_n'(\vartheta) = \Lambda_{n,1}'(\vartheta) + \Lambda_{n,2}'(\vartheta)$ with

$$\begin{aligned}
\Lambda_{n,1}'(\vartheta) &= -(1/n)\sum_{i=1}^{n} B'\Big(a(X_i^\top\vartheta + \rho_\vartheta(Z_i))\Big)a'(X_i^\top\vartheta + \rho_\vartheta(Z_i))(X_i^\top + \rho_\vartheta'(Z_i)),\\
\Lambda_{n,2}'(\vartheta) &= (1/n)\sum_{i=1}^{n}\Big(B'\Big(a(X_i^\top\vartheta + \rho_\vartheta(Z_i)\Big)a'(X_i^\top\vartheta + \rho_\vartheta(Z_i))(X_i^\top + \rho_\vartheta'(Z_i))\\
&\qquad -B'\Big(a(X_i^\top\vartheta + \hat{\rho}_\vartheta(Z_i)\Big)a'(X_i^\top\vartheta + \hat{\rho}_\vartheta(Z_i))(X_i^\top + \hat{\rho}_\vartheta'(Z_i)\Big).
\end{aligned}$$

By the dominance assumptions in (14), Assumption 2, (12), and the usual uniform strong law of large numbers(e.g. Ferguson, 1996, Page 108), we have

$$\lim_{n\to\infty}\sup_{\vartheta\in N(\theta)}\|\Lambda_{n,1}'(\vartheta) - D(\vartheta)\| = 0, \quad a.s. \tag{26}$$

Analogously by (10)-(11), (12)-(13) and (14)-(15), one could show $\Lambda'_{n,2}(\theta_n^*) = o_{\rm p}(n^{-1/2})$. This, (26) and the fact that $D(\vartheta)$ is bounded away from below show that $\Lambda'_n(\theta_n^*) = D(\theta) + o_{\rm p}(n^{-1/2})$. Thus in view of the first equality of (25), the expansion (24) gives $\hat{\theta}_n \xrightarrow{P} \theta$. This, (26), and the second equality of (25) yield the desired asymptotic normality. □

For convenience we write $\hat{\rho}_{\vartheta,ML} = \hat{\rho}_\vartheta$. Recall $S_\vartheta(r,y,x) = (y - h(x^\top\vartheta + r))a'(x^\top\vartheta + r)$. The proof of Theorem 4 is in the spirit of the proof of Lemma 8 of Severini and Wong(1992), though their proof used erroneous assumptions.

Proof of Theorem 4. Let $\Psi_n(\vartheta,r,z) = (1/nh_n)\sum_{j=1}^n S_\vartheta(r,Y_j,X_j)K_{nj}(z)$. Then $\hat{\psi}'_n(\vartheta,r,z) = \Psi_n(\vartheta,r,z)/\hat{q}(z)$, where $\hat{q}(z) = (1/nh_n)\sum_{i=1}^n K_{n,i}(z)$ is the kernel estimator of $q(z)$ for $z \in [0,1]$. With simple manipulation,

$$\mathbb{E}\,\Psi_n(\vartheta,r,z) - q(z)h'(\vartheta,r,z) = \int S_\vartheta(r,y,x)\Delta(y,x,u)K(u)\,\nu(dy)dxdu$$

where $\Delta(y,x,u) = f(y|\phi(z-uh_n))g(x,z-uh_n) - f(y|\phi(z))g(x,z)$, with $\phi(t) = a(x^\top\vartheta + \rho(t))$. By the Lipschitz continuity in Assumption 8,

$$|\Delta(y,x,u)| \le cuh_n\left[f(y|\phi(z)) + |(y-h(\eta^*))a'(\eta^*)|f(y|a(\eta^*))g(x,z-uh_n)\right],$$

where $\eta^* = x^\top\vartheta + \rho^*$ with ρ^* lying in between $\rho(z)$ and $\rho(z-uh_n)$ (clearly if $z \mapsto \rho(z)$ is differentiable, then $\rho^* = \rho(z-u^*h_n)$ for some $0 \le u^* \le 1$, and thus Assumption 10 (i) can be replaced with the weaker Assumption 10(ii)), and c is a constant depending only on the Lipschitz constant of ρ. It follows from Cauchy inequality, Assumption 10(i) with $(i,j) = (0,0)$, and the Lebesgue dominated convergence theorem that

$$\sup_{\vartheta,r,z} |\,\mathbb{E}\,\Psi_n(\vartheta,r,z) - q(z)\psi'(\vartheta,r,z)| = O(h_n). \tag{27}$$

Let $\bar{\Psi}_n(\vartheta,r,z) = \Psi_n(\vartheta,r,z) - \mathbb{E}\,\Psi_n(\vartheta,r,z)$. As in Severini and Wong(page 1800) and by Assumption 10(ii) with $(i,j) = (0,0)$, one has for any $\epsilon > 0$,

$$P(|\bar{\Psi}_n(\vartheta,r,z)| > \epsilon) \le c/(n^{p/2}(\epsilon h_n)^p). \tag{28}$$

By Assumption 1 and Assumption 10(iii) with $(i,j) = (0,0)$,

$$|\bar{\Psi}_n(\vartheta,r,z_1) - \bar{\Psi}_n(\vartheta,r,z_2)| \le c_1\frac{|z_1-z_2|}{h_n^2}\frac{1}{n}\sum_{j=1}^n A_{1j}$$

for an integrable sequence $\{A_{1j}\}$ independent of z_1, z_2. By Assumption 10(iii) with $(i,j) = (1,0)$,

$$|\bar{\Psi}_n(\vartheta_1,r,z) - \bar{\Psi}_n(\vartheta_2,r,z)| \le c_2\frac{\|\vartheta_1-\vartheta_2\|}{h_n}\frac{1}{n}\sum_{j=1}^n A_{2j}$$

for an integrable sequence $\{A_{2j}\}$ independent of ϑ_1, ϑ_2. Again by Assumption 10(iii) with $(i,j) = (0,1)$,

$$|\bar{\Psi}_n(\vartheta, r_1, z) - \bar{\Psi}_n(\vartheta, r_2, z)| \leq c_3 \frac{|r_1 - r_2|}{h_n} \frac{1}{n} \sum_{j=1}^{n} A_{3j}$$

for an integrable sequence $\{A_{3j}\}$ independent of r_1, r_2. Consequently there exists of a sequence $\{A_j\}$ independent of $w_i = (\vartheta_i, r_i, z_i), i = 1, 2$ such that

$$\sup_{\|w_1 - w_2\| \leq \delta} \left| \bar{\Psi}_n(w_1) - \bar{\Psi}_n(w_2) \right| \leq c_0 \frac{\delta}{h_n^2} \frac{1}{n} \sum_{j=1}^{n} A_j.$$

Let $\{\delta_n\}$ be a sequence tending to zero and $\Theta_n, \mathbb{H}_n, \mathbb{Z}_n$ be δ_n-nets of $\Theta, H, [0,1]$ respectively. Then, for some constant c,

$$\begin{aligned} P(\sup_{\vartheta, r, z} |\bar{\Psi}_n(\vartheta, r, z)| > \epsilon) &\leq P(\max_{\vartheta \in \Theta_n, r \in \mathbb{H}_n, z \in \mathbb{Z}_n} |\bar{\Psi}_n(\vartheta, r, z)| > \epsilon/2) \\ &+ P(\sup_{\|w_1 - w_2\| \leq \delta} |\bar{\Psi}_n(w_1) - \bar{\Psi}_n(w_2)| > \epsilon/2) \leq \frac{c}{\epsilon^p} \frac{1}{\delta_n^2 n^{p/2} h_n^p} + \frac{c}{\epsilon} \frac{\delta_n}{h_n^2} \to 0 \end{aligned}$$

if $\epsilon = M_n \tau_n = M_n n^{-p/(2p+4)} h_n^{(p+4)/(p+2)}$ with $M_n \to \infty$ and $\delta_n = O(\tau_n h_n^2)$. Likewise, one can establish $\|\hat{q}_n - q\| = O_P(h_n + n^{-p/(2p+4)} h_n^{-(p+4)/(p+2)})$. Combining the above and in view of Assumption 7 (ii) yields the desired (20). Analogously one can show (21) and (22). This completes the proof. □

References

Bhattacharya, P.K. and Zhao, P., 1993. Semiparametric inference in a partially linear model. *Ann. Statist.* **25**, 244–262.

Bickel, P.J., Klaassen, C.A.J., Ritov, Y., and Wellner, J.A., 1993. *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins Univerisity Press, Baltimore.

Brown, L.D., 1986. *Fundamentals of statistical exponential families: with applications in statistical decision theory.* Institute of Mathematical Statistics.

Ferguson, T.S., 1996. *A Course in Large Sample Theory.* Chapman & Hall.

Forrester, J., Hooper, W., Peng, H. & Schick, A., 2003. On construction of efficient estimators in semiparametric models. *Statist.&Decision* **21**, 109 - 137.

Härdle, W., Janssen, P. and Serfling, R., 1988. Strong uniform consistency rates for estimators of conditional functionals. *Ann.Statist.* **16**, 1428 – 1449.

Severini, T. A. and Wong, W. H., 1992. Profile likelihood and conditionally parametric models. *Ann. Statist.* **20**, 1768 - 1802.

Severini, T. A. and Staniswalis, J. G., 1994. Quasi-likelihood Estimation in Semimparametric Models. *J. Amer. Statist. Assoc.* **89**, 502 - 511.