Generalized Partially Linear Single-Index Models

R. J. Carroll, Jianqing Fan, Irène Gijbels and M. P. Wand*

April 20, 1995

The typical generalized linear model for a regression of a response Y on predictors $(\mathbf{X}, \mathbf{Z})$ has conditional mean function based upon a linear combination of $(\mathbf{X}, \mathbf{Z})$. We generalize these models to have a nonparametric component, replacing the linear combination $\boldsymbol{\alpha}_0^T\mathbf{X} + \boldsymbol{\beta}_0^T\mathbf{Z}$ by $\eta_0(\boldsymbol{\alpha}_0^T\mathbf{X}) + \boldsymbol{\beta}_0^T\mathbf{Z}$, where $\eta_0(\cdot)$ is an unknown function. These are the *generalized partially linear single-index models.* The models also generalize the "single-index" models, which have $\boldsymbol{\beta}_0 = 0$. Using local linear methods, estimates of the unknown parameters $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ and the unknown function $\eta_0(\cdot)$ are proposed, and their asymptotic distributions obtained. An example illustrates the algorithms and the models.

KEY WORDS: Asymptotic Theory; Generalized Linear Models; Kernel Regression; Local Estimation; Local Polynomial Regression; Nonparametric Regression; Single-Index Models.

* R. J. Carroll is Professor, Department of Statistics, Texas A&M University, College Station, TX 77843-3143. Jianqing Fan is Associate Professor, Department of Statistics, University of North Carolina, Chapel Hill, NC 27599–3260. Irène Gijbels is Assistant Professor, Institut de Statistique, Université Catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium. M. P. Wand is Senior Lecturer, Australian Graduate School of Management, University of New South Wales, Sydney 2052, Australia. Carroll's research was supported by a grant from the National Cancer Institute (CA-57030). Fan's research was supported by a grant from the National Science Foundation (DMS-9203135) and an NSF Postdoctoral Fellowship. Gijbels' research was supported by the 'Programme d'Action de Recherche Concertée' (No. 93-98–164).

1 INTRODUCTION

1.1 Motivation

The Framingham Heart Study (Kannel, et al., 1986) consists of a series of exams taken two years apart. For purposes of illustration, we use Exam #3 as the baseline. There are 1,615 men aged 31–65 in this data set, with the outcome indicating the occurrence of coronary heart disease (CHD) within an eight-year period following Exam #3; there were 128 such cases of CHD. Predictors employed in this example are the patient's age, smoking status and serum cholesterol, in addition to systolic blood pressure (SBP) at Exam #3, the latter being the average of two measurements taken by different examiners during the same visit.

For these data, let the response Y be the incidence of CHD, and let Z be the indicator of smoking status. The other covariates used are a vector, denoted by X, consisting of the three variables

$$\begin{aligned} X_1 &= \text{age of patient;} \\ X_2 &= \log(\text{SPB} - 25); \\ X_3 &= \log(\text{cholesterol level}). \end{aligned}$$

An ordinary logistic regression model says that the logit of CHD probabilities satisfies

$$\text{logit}\,\{P(\text{CHD}|X,Z)\} = \gamma_0 + \alpha_0^T X + \beta_0 Z. \tag{1}$$

The advantage of the linear-logistic models lies not only in its computational convenience, but more importantly in the ease of interpretation of the model parameters, and our ability to make inference about them.

As discussed in Section 7, this linear-logistic model fits the data reasonably well, but there is some curvature not captured by it. This article is concerned with simple semiparametric alternatives to the fully parametric model (1) which allow for such curvature, but yet retain the ease of interpretation of parameters such as α_0 and β_0. In this particular example, our generalization consists of two parts: (a) the linear combination $\alpha_0^T X$ enters the model via a nonparametric link function; and (b) smoking status $\beta_0 Z$ enters the model as a logistic offset. Combining (a) and (b) suggests the simple model

$$\text{logit}\,\{P(\text{CHD}|X,Z)\} = \eta_0(\alpha_0^T X) + \beta_0 Z, \tag{2}$$

for some completely unknown function η_0. Model (2) retains much of the ease of interpretation of model (1), in the sense that nonzero components of α_0 or β_0 indicate a "significant" predictor of CHD, but model (2) allows for curvature in the logit.

The purpose of this article is to introduce versions of (2) for generalized linear and quasilikelihood models, describe one way to fit such models, and derive an asymptotic theory which allows inference about the parameters (α_0, β_0). In the rest of this section, we will describe the general class of models of interest to us here, which we call *generalized partially linear single-index models*, or GPLSIM. We show that these models are a natural combination and generalization of simpler models already in the literature, namely single-index models and partially linear models. Further sections deal with fitting and making inference about GPLSIM.

1.2 The Models

We consider semiparametric versions of generalized linear models where a response Y is to be predicted by covariates $(\mathbf{X}, \mathbf{Z})$, where $\mathbf{X}$ and $\mathbf{Z}$ are possibly vector-valued predictors of lengths p and q, respectively. Generalized linear models are derived as follows. The conditional density of Y given $(\mathbf{X}, \mathbf{Z}) = (\mathbf{x}, \mathbf{z})$ belongs to a canonical exponential family

$$f_{Y|\mathbf{X},\mathbf{Z}}(y|\mathbf{x}, \mathbf{z}) = \exp\left[y\theta(\mathbf{x}, \mathbf{z}) - \mathcal{B}\left\{\theta(\mathbf{x}, \mathbf{z})\right\} + \mathcal{C}(y)\right] \tag{3}$$

for known functions $\mathcal{B}$ and $\mathcal{C}$. In parametric generalized linear models, the unknown regression function

$$\mu(\mathbf{x}, \mathbf{z}) = E(Y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) = \mathcal{B}'\left\{\theta(\mathbf{x}, \mathbf{z})\right\}$$

is modeled linearly via a link function g by

$$g\left\{\mu(\mathbf{x}, \mathbf{z})\right\} = \gamma_0 + \boldsymbol{\alpha}_0^T\mathbf{x} + \boldsymbol{\beta}_0^T\mathbf{z}. \tag{4}$$

If $g = (\mathcal{B}')^{-1}$ (the inverse function of $\mathcal{B}'$), then g is the canonical link function. See Nelder & Wedderburn (1972) and McCullagh & Nelder (1989) for more details.

In many practical situations, however, the linear model (4) is not complex enough to capture the underlying relationship between the response variable and its associated covariates. Indeed some components can be highly nonlinear. A natural generalization of (4) is to allow only some of the predictors to be modeled linearly, with others being modeled nonlinearly. This leads us to consider the class of *generalized partially linear single-index models*

$$g\left\{\mu(\mathbf{x}, \mathbf{z})\right\} = \eta_0(\boldsymbol{\alpha}_0^T\mathbf{x}) + \boldsymbol{\beta}_0^T\mathbf{z}, \text{ with } ||\boldsymbol{\alpha}_0|| = 1. \tag{5}$$

The restriction that $||\boldsymbol{\alpha}_0|| = 1$ is only a matter of identifiability.

The model (5) is flexible enough to cover a variety of situations. When $\boldsymbol{\beta}_0 = 0$, or equivalently there are no predictors $\mathbf{Z}$, (5) is simply a generalized linear model with an *unknown* link function. The problem of the "missing link" function in generalized linear models has been considered previously by Weisberg &

Welsh (1994). In other contexts, when only the mean function is specified, this problem is known as the nonparametric *single-index model*, and has been considered by Härdle, Hall & Ichimura (1993). The appeal of these models is that by focusing on an index $\boldsymbol{\alpha}_0^T\mathbf{X}$, the so-called "curse of dimensionality" in fitting multivariate nonparametric regression functions is avoided; at the cost of some loss in flexibility.

Severini & Staniswalis (1994) and Hunsberger (1994) consider model (5), but with $\mathbf{X}$ scalar, so that $p = 1$ and $\boldsymbol{\alpha}_0 = 1$. In this case, model (5) becomes

$$g\{\mu(\mathbf{x},\mathbf{z})\} = \eta_0(\mathbf{x}) + \boldsymbol{\beta}_0^T\mathbf{z}. \tag{6}$$

Model (6) is particularly popular in the spline literature. See, for example, Wahba (1984), Heckman (1986), Chen (1988), Speckman (1988) and Cuzick (1992), where it is called the *partial spline* model or the *partially linear* model.

In the context of the unknown link function, the single-index model, or the model with $\boldsymbol{\alpha}_0 = 1$, our method differs from those previously cited in that we use local linear rather than simple kernel regression methods.

Our aim then is to estimate the unknown parameters $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$ and the unknown function $\eta_0(\cdot)$ in the full model (5), thus generalizing both the single-index or missing link model and the partially linear model.

Our work also applies to quasilikelihood models, where only the relationship between the mean and the variance is specified. In this situation, estimation of the mean can be achieved by replacing the conditional loglikelihood $\ln f_{Y|\mathbf{X},\mathbf{Z}}(y|\mathbf{x},\mathbf{z})$ by a *quasilikelihood function* $Q\{\mu(\mathbf{x},\mathbf{z}),y\}$. If the conditional variance is modeled as $\mathrm{var}(Y|\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z}) = \sigma^2 V\{\mu(\mathbf{x},\mathbf{z})\}$ for some known positive function V, then the corresponding quasilikelihood function $Q(w,y)$ satisfies

$$\frac{\partial}{\partial w}Q(w,y) = (y-w)/V(w). \tag{7}$$

(Wedderburn 1974, McCullagh & Nelder 1989, Chapter 9). The quasiscore (7) possesses properties similar to those of the usual likelihood score function.

1.3 Outline

In Section 2, we discuss the algorithms used to estimate the generalized partially linear single-index model. Section 3 describes distribution theory for estimating $\eta_0(\cdot)$, while Section 4 gives the theory for the estimates of $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$. Section 5 describes methods for estimating the standard errors of the parametric and nonparametric parts of the model. Implementation details are given in Section 6 and a numerical example is presented in Section 7. In Section 8 we briefly discuss the issue of incorporating interactions in the model and comment on different modeling approaches. All proofs are given in the appendix (Section 9).

2 MAXIMUM QUASILIKELIHOOD

2.1 Local Quasilikelihood

Under model (5), the primary interest is to estimate $\boldsymbol{\alpha}_0$, $\boldsymbol{\beta}_0$ and $\eta_0(\cdot)$. Since $\eta_0(\cdot)$ is modeled nonparametrically, it is natural to consider local quasilikelihood. However, efficient estimation of the global parameters $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$ requires using all data points. In local quasilikelihood, we will approximate $\eta_0(\cdot)$ locally by a linear function

$$\eta_0(v) \approx \eta_0(u) + \eta_0'(u)(v-u) \equiv a + b(v-u),$$

for v in a neighborhood of u, where $a = \eta_0(u)$ and $b = \eta_0'(u)$.

Let K be a symmetric probability density function, and $K_h(t) = K(t/h)/h$ be a rescaling of K. The function K is usually called a kernel function and the parameter h is called the bandwidth. For $i = 1, ..., n$, a sample $(Y_i, \mathbf{X}_i, \mathbf{Z}_i)$ is observed.

With any given $\boldsymbol{\alpha}$, we first find $\widehat{a}$, $\widehat{b}$ and $\widehat{\boldsymbol{\beta}}$ to maximize the local quasilikelihood

$$\sum_{i=1}^{n} Q\left[g^{-1}\left\{a + b(\boldsymbol{\alpha}^T\mathbf{X}_i - u) + \boldsymbol{\beta}^T\mathbf{Z}_i\right\}, Y_i\right] K_h(\boldsymbol{\alpha}^T\mathbf{X}_i - u). \tag{8}$$

Note that the local quasilikelihood is really a weighted quasilikelihood, with weights $K_h(\boldsymbol{\alpha}^T\mathbf{X}_i - u)$.

Denote $\widehat{\eta}(u) \equiv \widehat{\eta}(u; h, \boldsymbol{\alpha}) = \widehat{a}$ and $\widehat{\boldsymbol{\beta}} \equiv \widehat{\boldsymbol{\beta}}(u; h, \boldsymbol{\alpha})$. Having constructed the function $\widehat{\eta}(u; h, \boldsymbol{\alpha})$, estimate the global parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ by maximizing the global quasilikelihood

$$\sum_{i=1}^{n} Q\left[g^{-1}\left\{\widehat{\eta}(\boldsymbol{\alpha}^T\mathbf{X}_i; h, \boldsymbol{\alpha}) + \boldsymbol{\beta}^T\mathbf{Z}_i\right\}, Y_i\right]. \tag{9}$$

Note that maximizing (9) simultaneously yields estimators for $\boldsymbol{\alpha}_0$, $\boldsymbol{\beta}_0$ and η_0.

Thus, the basic idea is simple: estimate $\eta_0(\cdot)$ locally via (8), then use all the data and (9) to estimate $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$, with $\widehat{\eta}(\cdot)$ replacing $\eta_0(\cdot)$. To implement the idea, one can either use a one-step approach, which is simple computationally, or a fully iterated approach, which is preferable from the point of view of asymptotic efficiency.

2.2 Details of the One-Step Algorithm

The following simple one-step algorithm makes (8)–(9) operational. When $\boldsymbol{\alpha}_0$ is known, or $\mathbf{X}$ is scalar (so that $\boldsymbol{\alpha}_0 = 1$), (8)–(9) give a fully specified, noniterative algorithm. When $\boldsymbol{\alpha}_0$ is unknown, we proceed as follows.

STEP 0: Fit a parametric generalized linear model to obtain the initial estimates $(\widehat{\boldsymbol{\alpha}}_1, \widehat{\boldsymbol{\beta}})$ and set $\widehat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\alpha}}_1/\|\widehat{\boldsymbol{\alpha}}_1\|$.

STEP 1: With the estimated $\widehat{\boldsymbol{\alpha}}$, find $\widehat{\eta}(u; h, \widehat{\boldsymbol{\alpha}})$ by maximizing the local quasilikelihood (8). We take h to be an estimate of the bandwidth that is optimal for estimation of $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$.

STEP 2: With the estimated $\widehat{\eta}(u; h, \widehat{\boldsymbol{\alpha}})$, update $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$ by maximizing the global quasilikelihood (9).

STEP 3: Continue Steps 1 and 2 until convergence.

STEP 4: Fix $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ at its estimated value from Step 3. The final estimate of $\eta_0(\cdot)$ is $\widehat{\eta}(u; h, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}) = \widehat{a}$ where $(\widehat{a}, \widehat{b})$ maximizes

$$\sum_{i=1}^{n} Q\left[g^{-1}\left\{a + b(\widehat{\boldsymbol{\alpha}}^T \mathbf{X}_i - u) + \widehat{\boldsymbol{\beta}}^T \mathbf{Z}_i\right\}, Y_i\right] K_h(\widehat{\boldsymbol{\alpha}}^T \mathbf{X}_i - u). \tag{10}$$

At this final step we take h to be an estimate of the bandwidth that is optimal for estimation of $\eta_0(\cdot)$ when $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$ are known.

We recommend calculating $\widehat{\eta}(\cdot; h, \widehat{\boldsymbol{\alpha}})$ at a *fixed*, but fine grid of points and using linear interpolation to calculate the other values of $\widehat{\eta}(\cdot; h, \widehat{\boldsymbol{\alpha}})$ when needed. The values $\widehat{\boldsymbol{\beta}}$ in Steps 0 and 1 can be used as the initial values in Steps 1 and 2 respectively. Estimation of the bandwidth is discussed in Section 3.3.

The term "one-step algorithm" is a bit of a misnomer, because it is noniterative only when $\boldsymbol{\alpha}_0$ is known, i.e., the partially linear model. The algorithm can be made noniterative by two changes. First, maximize (8) in all of $(a, b, \boldsymbol{\alpha}, \boldsymbol{\beta})$. Second, maximize (10) in $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ only. We conjecture that this noniterative algorithm is less efficient than our defined one-step algorithm; this would be interesting to verify.

2.3 Details of the Fully Iterated Algorithm

The fully iterated algorithm can be defined as follows.

STEP 0: Obtain the initial estimates $(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})$ either through Step 0 in the one-step algorithm or through the final results of one-step algorithm itself.

STEP 1: For any given $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, find $\widehat{\eta}(u; h, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \widehat{a}$ by solving (10), with $(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})$ in (10) replaced by $(\boldsymbol{\alpha}, \boldsymbol{\beta})$. We take h to be an estimate of the bandwidth that is optimal for estimation of $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$.

STEP 2: Update $(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})$ by maximizing

$$\sum_{i=1}^{n} Q\left[g^{-1}\left\{\widehat{\eta}(\boldsymbol{\alpha}^T \mathbf{X}_i; h, \boldsymbol{\alpha}, \boldsymbol{\beta}) + \boldsymbol{\beta}^T \mathbf{Z}_i\right\}, Y_i\right]. \tag{11}$$

STEP 3: Fix $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ at its estimated value from Step 2. The final estimate of $\eta_0(\cdot)$ is $\widehat{\eta}(u; h, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}) = \widehat{a}$ where $(\widehat{a}, \widehat{b})$ is obtained by solving (10). At this final step we take h to be an estimate of the bandwidth that is optimal for estimation of $\eta_0(\cdot)$ when $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$ are known.

Instead of maximizing (11) as stated, an alternative approach is to update the current $(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})$ by maximizing

$$\sum_{i=1}^{n} Q\left[g^{-1}\left\{\widehat{\eta}(\boldsymbol{\alpha}^T \mathbf{X}_i; h, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}) + \boldsymbol{\beta}^T \mathbf{Z}_i\right\}, Y_i\right]. \tag{12}$$

For completely arbitrary functions Q, the estimators resulting from this type of updating are not equivalent to those obtained by maximizing (11). However, our results are stated for a special form of Q, namely (7), in which case the estimates from the two algorithms can be shown to be asymptotically equivalent. We use (12) in our numerical calculations. Details of the theoretical calculations will appear elsewhere.

2.4 Alternatives to Local Kernel Quasilikelihood Estimation

The algorithms suggested here use local linear weighted fits based upon kernel weights with a fixed global bandwidth. Restriction to this class enables us to provide asymptotic distributions for the nonparametric and parametric parts of the model. In particular, we will conclude that

(a) Estimation of the nonparametric part of the model, $\eta_0(\cdot)$, can be done just as well as if the parametric parts, $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$, were known;

(b) The parametric parts can be estimated at the usual parametric rate of convergence;

(c) Estimation of the nonparametric part of the model has an effect on the distribution of the estimates of the parametric parts; and

(d) Best estimation of the parametric parts, in particular $\boldsymbol{\beta}$ requires that one undersmooth the nonparametric part.

There is no need to restrict oneself to the local linear case, and indeed local quadratic fits are known to have advantages in ordinary nonparametric regression. All of our results generalize easily to local polynomial fitting (see Fan, Heckman & Wand, 1995). One is not restricted to local kernel quasilikelihood estimation with a global bandwidth to determine smoothing neighborhoods. For example, one might use a locally determined bandwidth or even a nearest neighbor scheme to determine the smoothing neighborhoods, similar to what is done in lowess (Cleveland, 1979). In principle, there is no need to restrict to kernel estimation of any type, with splines being a natural alternative. It would be interesting to investigate all of these possibilities. From a theoretical viewpoint, however, it seems to us that conclusions (a)–(d) will hold even for these other schemes, with only the details differing from case-to-case.

3 DISTRIBUTION THEORY: NONPARAMETRIC PART

3.1 Introduction

The true parameters in model (5) are denoted by $\boldsymbol{\alpha}_0$, $\boldsymbol{\beta}_0$ and $\eta_0(\cdot)$. In this section, we investigate properties of the estimators of the nonparametric part $\eta_0(\cdot)$ of (5) when $\boldsymbol{\alpha}_0$ is either known or estimated to the order $O_P(n^{-1/2})$. The distribution theory depends on two cases: (a) the one-step approach when $\boldsymbol{\beta}_0$ is estimated locally as in (8); and (b) the fully iterated approach (10). In case (a), we are using only $O(nh)$ data points to estimate both $\eta_0(\cdot)$ and $\boldsymbol{\beta}_0$. In case (b), $\boldsymbol{\beta}_0$ is estimated at parametric rates, and thus $\eta_0(\cdot)$ can be estimated asymptotically as well as if $\boldsymbol{\beta}_0$ were known.

3.2 One-Step Estimate of the Nonparametric Part

We consider case (a) first. Let $\rho_\ell(t) = \left\{\frac{dg^{-1}(t)}{dt}\right\}^\ell / \left[\sigma^2 V\left\{g^{-1}(t)\right\}\right]$, $\ell = 1, 2$, and denote by $f(\cdot)$ the marginal density of $U = \boldsymbol{\alpha}_0^T \mathbf{X}$. For the model (3) with the canonical link function $g = (\mathcal{B}')^{-1}$, we have $\rho_2\{g(\mu)\} = \sigma^2 V(\mu)$. Define $\kappa_j = \int t^j K(t)dt$, $\nu_j = \int t^j K^2(t)dt$, and

$$\begin{aligned}
\boldsymbol{\Sigma}(u) &= E\left[\rho_2\left\{\eta_0(U) + \boldsymbol{\beta}_0^T \mathbf{Z}\right\}\left(\begin{array}{cc} 1 & \mathbf{Z}^T \\ \mathbf{Z} & \mathbf{Z}\mathbf{Z}^T \end{array}\right)\middle| U = u\right]; \qquad (13) \\
q_1(x, y) &= \left\{y - g^{-1}(x)\right\}\rho_1(x); \\
m_i &= m_i(U_i) = \eta_0(U_i) + \boldsymbol{\beta}_0^T \mathbf{Z}_i; \\
W_i &= \text{first element of the vector } q_1(m_i, Y_i)\boldsymbol{\Sigma}^{-1}(u)(1, \mathbf{Z}_i^T)^T; \\
d(u) &= \text{first diagonal element of the matrix } \boldsymbol{\Sigma}^{-1}(u).
\end{aligned}$$

Theorem 1: Consider the maximizer of the local quasilikelihood (8). Then, as $n \to \infty$, $h \to 0$ and $nh \to \infty$, under Condition 1 as stated in the appendix (Section 9),

$$\begin{aligned}
&(nh)^{1/2}\left(\left[\begin{array}{c} \widehat{\eta}(u) - \eta_0(u) \\ \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \end{array}\right] - \frac{\kappa_2}{2}\eta_0''(u)h^2\boldsymbol{\Sigma}^{-1}(u)E\left[\rho_2\left\{\eta_0(U) + \boldsymbol{\beta}_0^T\mathbf{Z}\right\}\left(\begin{array}{c} 1 \\ \mathbf{Z} \end{array}\right)\middle| U = u\right]\right) \\
&\xrightarrow{D} \text{Normal}\left[\mathbf{0}, \frac{\nu_0}{f(u)}\boldsymbol{\Sigma}^{-1}(u)\right]. \qquad (14)
\end{aligned}$$

In fact, we have the asymptotic expansion

$$\widehat{\eta}(u) - \eta_0(u) = (\kappa_2/2)\eta_0''(u)h^2 + \frac{1}{nf(u)}\sum_{i=1}^n W_i K_h(\boldsymbol{\alpha}_0^T\mathbf{X}_i - u) + o_P\left\{(nh)^{-1/2} + h^2\right\}, \qquad (15)$$

and hence

$$(nh)^{1/2}\left\{\widehat{\eta}(u) - \eta_0(u) - \frac{\kappa_2}{2}\eta_0''(u)h^2\right\} \xrightarrow{D} \text{Normal}\left[\mathbf{0}, \frac{\nu_0}{f(u)}d(u)\right]. \qquad (16)$$

Remark 1: Consider the situation that $\sigma^2 V(\mu) \equiv \sigma^2$ and $E(\mathbf{Z}|\mathbf{X}) = \mathbf{0}$. For this normal model with the identity link, the quasilikelihood estimates are the ordinary least squares estimates. It is easily seen that $d(u) = \sigma^2$. Hence, in this particular case, even though $\boldsymbol{\beta}_0$ is estimated locally, the bias and variance of $\widehat{\eta}(u)$ are the same as if $\boldsymbol{\beta}_0$ were known. In other words, it admits the same bias and variance expression as the ordinary local linear regression smoother obtained by Tsybakov (1986) and Fan (1992).

Remark 2: When the variance function is misspecified, i.e. $\text{var}(Y|\mathbf{X}, \mathbf{Z}) \neq \sigma^2 V\{\mu(\mathbf{X}, \mathbf{Z})\}$, the rate results in Theorem 1 continue to hold. One has to change the matrix $\boldsymbol{\Sigma}(u)$ to reflect the misspecification of the variance function. See Theorem 1 of Fan, et al. (1995) for such a modification.

3.3 Fully Iterated Estimate of the Nonparametric Part

For reasons described more fully in Section 4 (see particularly Remark 3), one-step estimation of the parameters $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$ requires a smaller bandwidth than the one that is optimal for estimating the nonparametric component $\eta_0(\cdot)$. Thus, while the one-step algorithm of Section 2.2 results in an estimate of $\eta_0(\cdot)$, the estimate is undersmoothed, i.e., too variable. If interest focuses on obtaining a good estimate of $\eta_0(\cdot)$, a final step should be added to the algorithm, as follows.

Fix $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$, and solve (10), i.e., maximize it with respect to (a, b) only, announcing $\widehat{\eta}(u) = \widehat{a}$. In this case, the results are very easy: (16) continues to hold if we replace $d(u)$ by $d_*(u) = \left(E\left[\rho_2\left\{\eta_0(u) + \boldsymbol{\beta}_0^T\mathbf{Z}\right\}|U = u\right]\right)^{-1}$. This result coincides with the univariate result given in Fan, et al. (1995).

These results suggests bandwidth estimators in the spirit of that of Ruppert, Sheather & Wand (1995). For example, consider estimation of $\eta_0(\cdot)$ at the final step. For a given function $w(\cdot)$ with compact support, minimizing the asymptotic weighted mean squared error with weight $f(\cdot)w(\cdot)$ yields the optimal global bandwidth

$$h_{\text{opt}} = C(K)n^{-1/5}\left\{\frac{\int d_*(u)w(u)du}{\int \eta_0''(u)^2 f(u)w(u)du}\right\}^{1/5}, \tag{17}$$

where $C(K) = \left(\nu_0\kappa_2^{-2}\right)^{1/5}$.

The example in Section 7 treats the case where both Y and Z are $0-1$ variables, so we will briefly describe how one might use h_{opt} to motivate a rough rule for choosing the bandwidth in this context. Extension to other contexts is straightforward. For the Bernoulli likelihood with logit link,

$$d_*(u)^{-1} = \frac{e^{\eta_0(u)}\{1-\zeta_0(u)\}}{\{1+e^{\eta_0(u)}\}^2} + \frac{e^{\eta_0(u)+\beta_0}\zeta_0(u)}{\{1+e^{\eta_0(u)+\beta_0}\}^2},$$

where $\zeta_0(u) = P(Z = 1|U = u)$. Let $\hat{\eta}_Q(\cdot)$ be the quadratic and $\hat{\zeta}_L(\cdot)$ be the linear logistic regression estimates of $\eta_0(\cdot)$ and $\zeta_0(\cdot)$ respectively. Let $\widehat{\beta}$ be the estimate of $\boldsymbol{\beta}_0$ from the previous iteration. Then the integral on the numerator of (17) can be estimated by direct replacement of $\eta_0(\cdot)$, $\zeta_0(\cdot)$ and $\boldsymbol{\beta}_0$ by $\widehat{\eta}(\cdot)$, $\widehat{\zeta}_L(\cdot)$ and $\widehat{\boldsymbol{\beta}}$, respectively. An estimate for the integral on the denominator is $n^{-1}\sum_{i=1}^n \hat{\eta}_Q''(U_i)^2 w(U_i)$. A sensible choice for w is the indicator function on the range of the U_i; with about 10% clipped off either end to avoid boundary problems. This results in an estimated bandwidth, $\hat{h}_{\text{opt}}$, for use in Step 4 of the fully iterated algorithm. The rule will give close to optimal answers when the true logit$\{\eta_0(\cdot)\}$ and logit$\{\zeta_0(\cdot)\}$ are approximated reasonably well by a quadratic and straight line, respectively.

A sensible rule for choice of h in Step 1 of either algorithm is more difficult to derive. For the one-step algorithm, a relatively *ad hoc* possibility is

$$\hat{h}_{\text{opt}} \times n^{1/5} \times n^{-1/3} = \hat{h}_{\text{opt}} \times n^{-2/15}, \tag{18}$$

since this guarantees that the required bandwidth has correct order of magnitude for the conjectured optimal asymptotic performance (see Remark 3 for more details).

4 DISTRIBUTION THEORY: PARAMETRIC PARTS

We now study the global estimator for the parametric components $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$. We treat separately the one-dimensional case ($p = 1$), for which $\boldsymbol{\alpha}_0 = 1$ and $\boldsymbol{\alpha}_0^T \mathbf{X} = X$.

4.1 The Scalar Case

We now state the results for the one-step estimate and the fully iterated estimate.

Theorem 2: Let $\widehat{\boldsymbol{\beta}}$ be the one-step estimate which maximizes the quasilikelihood (9) with $\boldsymbol{\alpha}_0 = 1$. Since $U = \boldsymbol{\alpha}_0^T \mathbf{X} = X$, write $\boldsymbol{\Sigma}(U) = \boldsymbol{\Sigma}(X)$ in Theorem 1. Under Conditions 1 and 2 in the appendix, as $n \to \infty$, $nh^4 \to 0$ and $nh^2/\log(1/h) \to \infty$,

$$n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{D} N(\mathbf{0}, \mathbf{B}^{-1}\boldsymbol{\Sigma}_1\mathbf{B}^{-1}), \tag{19}$$

where

$$\mathbf{B} = E\left[\rho_2\left\{\eta_0(X) + \boldsymbol{\beta}_0^T\mathbf{Z}\right\}\mathbf{Z}\mathbf{Z}^T\right],$$

$$\boldsymbol{\Sigma}_1 = \mathbf{B} + E\left\{\boldsymbol{\gamma}(X)\boldsymbol{\gamma}^T(X)\mathbf{e}_1^T\boldsymbol{\Sigma}^{-1}(X)\mathbf{e}_1\right\},$$

$\boldsymbol{\gamma}(u) = E[\rho_2\left\{\eta_0(u) + \boldsymbol{\beta}_0^T\mathbf{Z}\right\}\mathbf{Z}|X = u]$, $\mathbf{e}_1$ is the unit vector with 1 in the first position.

Thus, one iteration leads already to a root-n consistent estimator of $\boldsymbol{\beta}_0$.

Theorem 3: Under the conditions of Theorem 2, for the fully iterated estimator defined by (11) with $\boldsymbol{\alpha} = 1$,

$$n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{D} N(\mathbf{0}, \mathbf{B}_2^{-1}), \tag{20}$$

provided that $\boldsymbol{\beta}$ is maximized in a consistent neighborhood of $\boldsymbol{\beta}_0$. Here,

$$\mathbf{B}_2 = E\left\{\mathbf{Z}\mathbf{Z}^T\rho_2(\cdot)\right\} - E\left[\frac{E\left\{\mathbf{Z}\rho_2(\cdot)|X\right\}E\left\{\mathbf{Z}^T\rho_2(\cdot)|X\right\}}{E\left\{\rho_2(\cdot)|X\right\}}\right] \quad \text{where} \quad \rho_2(\cdot) = \rho_2\left\{\eta_0(X) + \boldsymbol{\beta}_0^T\mathbf{Z}\right\}$$

Remark 3: Theorem 2, which concerns the one-step estimator, has an important restriction on the bandwidth h, which precludes the nearly universally familiar optimal bandwidth rates for nonparametric regression, in which h is proportional to $n^{-1/5}$. Basically, our conditions require that in order to estimate $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ *at the rate* $n^{-1/2}$, one must undersmooth the nonparametric part $\eta_0(\cdot)$. One can show that for the one-step estimator, if one uses any ordinary bandwidth selector, the rate of convergence of the estimates is slower than the familiar one. The need to undersmooth to obtain usual rates of convergence is standard in the kernel literature and has analogues in the spline literature (Hastie & Tibshirani, 1990, pp. 154–155). On the other hand, for the fully iterated estimator defined by (11) and described by Theorem 3, in the linear regression single index model with no $\mathbf{Z}$, ordinary bandwidth rates are permissible. as shown by Härdle, Hall & Ichimura (1993), who suggest that one maximize (11) simultaneously in the bandwidth and the parameters. Severini & Staniswallis (1994) and Hunsberger (1994) (see also Severini & Wong, 1992) show the same thing for the partially linear model. Since ordinary bandwidths "work" for single-index models and also for partially linear models, it is reasonable to suppose that they work for the combination, namely our GPLSIM's. We provide a very brief sketch of an argument in the appendix (section 9.4) verifying that ordinary bandwidth rates are possible for full GPLSIM when (11) is maximized.

Remark 4: In the normal model with identity link function, an interesting simplification occurs. We set $E(\mathbf{Z}) = 0$ without loss of generality and define $q(\mathbf{X}) = E(\mathbf{Z}|\mathbf{X})$. Then $\mathbf{B}_2 = \sigma^{-2} E\{\mathrm{var}(\mathbf{Z}|\mathbf{X})\}$, while the asymptotic variance (19) for the one-step estimator is

$$\sigma^2[\{E\mathbf{Z}\mathbf{Z}^T\}^{-1} + Eq(\mathbf{X})q(\mathbf{X})^T\{1 - q(\mathbf{X})^T(E\mathbf{Z}\mathbf{Z}^T)^{-1}q(\mathbf{X})\}^{-1}].$$

Since $\mathbf{B}_2^{-1} = \sigma^2\{E\mathbf{Z}\mathbf{Z}^T - q(\mathbf{X})q(\mathbf{X})^T\}^{-1}$, one can easily see that the fully iterated estimator is uniformly as or more efficient than the one-step estimator. However, when $\mathbf{X}$ and $\mathbf{Z}$ are independent, the one-step estimator is as efficient as the fully iterated estimator. Therefore, iteration is not necessary when $\mathbf{X}$ and $\mathbf{Z}$ are weakly correlated.

4.2 The Multivariate Case

We now turn to the question of the simultaneous estimation of $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$. Again, results are obtained for the one-step and fully iterated estimates. First consider the one-step estimate, so that for a given $\widehat{\eta}$, let $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$ maximize the quasilikelihood (9). Thus, $\widehat{\boldsymbol{\alpha}}$, $\widehat{\boldsymbol{\beta}}$ are solutions to the quasilikelihood equations

$$\begin{aligned} \frac{\partial}{\partial\boldsymbol{\alpha}} \textstyle\sum_{i=1}^n Q\left[g^{-1}\left\{\widehat{\eta}(\boldsymbol{\alpha}^T\mathbf{X}_i) + \boldsymbol{\beta}^T\mathbf{Z}_i\right\}, Y_i\right] &= \mathbf{0} \\ \frac{\partial}{\partial\boldsymbol{\beta}} \textstyle\sum_{i=1}^n Q\left[g^{-1}\left\{\widehat{\eta}(\boldsymbol{\alpha}^T\mathbf{X}_i) + \boldsymbol{\beta}^T\mathbf{Z}_i\right\}, Y_i\right] &= \mathbf{0}, \end{aligned} \tag{21}$$

in a consistent neighborhood of respectively $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$. Further, suppose that (44) and its version for the first derivative functions hold and that $\widehat{\eta}''(\cdot)$ is continuous and stochastically bounded.

Define

$$\mathbf{P}_{\boldsymbol{\alpha}} = \begin{bmatrix} \mathbf{I} - \boldsymbol{\alpha}_0\boldsymbol{\alpha}_0^T & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$$

and let $\mathbf{A}^-$ denote a generalized inverse of a square matrix $\mathbf{A}$.

Theorem 4: Define

$$\mathcal{G}(u) = e_1^T\Sigma^{-1}(u)E\left[\rho_2\left\{\eta_0(u) + \boldsymbol{\beta}_0^T\mathbf{Z}\right\}\mathbf{X}(1, \mathbf{Z}^T)\middle| U = u\right],$$

where $e_1^T = (1, 0, ..., 0)$. Under Conditions 1 and 2, the assumptions above and the restrictions on the bandwidths as stated in Theorem 2, the one-step estimate

$$n^{1/2}\begin{pmatrix} \widehat{\boldsymbol\alpha} - \boldsymbol\alpha_0 \\ \widehat{\boldsymbol\beta} - \boldsymbol\beta_0 \end{pmatrix} \xrightarrow{D} \text{Normal}\left[\mathbf{0}, (\mathbf{P}_{\boldsymbol\alpha}\mathbf{C})^{-}\mathbf{P}_{\boldsymbol\alpha}\boldsymbol\Sigma_2\mathbf{P}_{\boldsymbol\alpha}\left\{(\mathbf{P}_{\boldsymbol\alpha}\mathbf{C})^{-}\right\}^T\right], \tag{22}$$

where

$$\mathbf{C} = E\left(\rho_2\left\{\eta_0(\boldsymbol\alpha_0^T\mathbf{X}) + \boldsymbol\beta_0^T\mathbf{Z}\right\}\begin{bmatrix}\eta_0'(\boldsymbol\alpha_0^T\mathbf{X})\mathbf{X} \\ \mathbf{Z}\end{bmatrix}\begin{bmatrix}\eta_0'(\boldsymbol\alpha_0^T\mathbf{X})\left\{\mathbf{X} - \mathcal{G}(\boldsymbol\alpha_0^T\mathbf{X})\right\} \\ \mathbf{Z}\end{bmatrix}^T\right);$$

$$\boldsymbol\Sigma_2 = E\left[\rho_2\left\{\eta_0(\boldsymbol\alpha_0^T\mathbf{X}) + \boldsymbol\beta_0^T\mathbf{Z}\right\}\boldsymbol\delta\boldsymbol\delta^T\right].$$

Further,

$$\boldsymbol\delta = \begin{bmatrix}\eta_0'(\boldsymbol\alpha_0^T\mathbf{X})\mathbf{X} - \boldsymbol\gamma_1(\boldsymbol\alpha_0^T\mathbf{X})\mathbf{e}_1^T\boldsymbol\Sigma^{-1}(\boldsymbol\alpha_0^T\mathbf{X})(1, \mathbf{Z}^T)^T \\ \mathbf{Z} - \mathbf{e}_1^T\boldsymbol\Sigma^{-1}(\boldsymbol\alpha_0^T\mathbf{X})(1, \mathbf{Z}^T)^T\boldsymbol\gamma(\boldsymbol\alpha_0^T\mathbf{X})\end{bmatrix},$$

$\boldsymbol\gamma_1(u) = E\left[\rho_2\left\{\eta_0(u) + \boldsymbol\beta_0^T\mathbf{Z}\right\}\eta_0'(u)\mathbf{X}|\boldsymbol\alpha_0^T\mathbf{X} = u\right]$, $\boldsymbol\gamma(u)$ is as in Theorem 2, and $\boldsymbol\Sigma(u)$ is as in Theorem 1.

Theorem 5: Under the conditions of Theorem 4, for the fully iterated estimator,

$$n^{1/2}\begin{pmatrix} \widehat{\boldsymbol\alpha} - \boldsymbol\alpha_0 \\ \widehat{\boldsymbol\beta} - \boldsymbol\beta_0 \end{pmatrix} \xrightarrow{D} \text{Normal}\left(\mathbf{0}, \mathbf{Q}^{-}\right), \tag{23}$$

where, if $\rho_2(\cdot) = \rho_2\left\{\eta_0(\boldsymbol\alpha_0^T\mathbf{X}) + \boldsymbol\beta_0^T\mathbf{Z}\right\}$,

$$\begin{aligned}\mathbf{Q} &= E\left[\rho_2(\cdot)\begin{Bmatrix}\mathbf{X}\eta_0'(U) \\ \mathbf{Z}\end{Bmatrix}\begin{Bmatrix}\mathbf{X}\eta_0'(U) \\ \mathbf{Z}\end{Bmatrix}^T\right] \\ &\quad - E\left(\rho_2(\cdot)\begin{Bmatrix}\mathbf{X}\eta_0'(U) \\ \mathbf{Z}\end{Bmatrix}\begin{bmatrix}E\left\{\mathbf{X}\eta_0'(U)\rho_2(\cdot)|U\right\}/E\left\{\rho_2(\cdot)|U\right\} \\ E\left\{\mathbf{Z}\rho_2(\cdot)|U\right\}/E\left\{\rho_2(\cdot)|U\right\}\end{bmatrix}^T\right).\end{aligned}$$

Remark 5: When $\sigma^2V(\mu) = \sigma^2$, with identity link, and when there is no $\boldsymbol\beta$-component, Theorems 4 and 5 reduce to the result of Härdle, Hall & Ichimura (1993) for the single-index model.

Remark 6: Consult Remark 3 after Theorem 3 for discussion of the bandwidth conditions.

5 INFERENCE AND STANDARD ERRORS

A consistent estimate of σ^2 is the weighted mean squared error of the residuals Y_i against their predicted mean, with weights $1/V\{\widehat\mu(\mathbf{X}_i, \mathbf{Z}_i)\}$; one can use $n - \ell_n - p - q$ degrees of freedom, where ℓ_n is the effective number of parameters used in estimating $\eta_0(\cdot)$. The rest of this section discusses estimating the other variance terms.

5.1 Estimation of the Nonparametric Part

We have already indicated that the distribution of the final estimate of $\eta_0(\cdot)$ is the same as if $\boldsymbol\alpha_0$ and $\boldsymbol\beta_0$ were known, namely (16) with $d(u)$ replaced by $d_*(u) = \left(E\left[\rho_2\left\{\eta_0(u) + \boldsymbol\beta_0^T\mathbf{Z}\right\}|U = u\right]\right)^{-1}$. Each of the terms in (16) can be estimated after replacing $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0, \eta_0)$ by their estimates: the numerator by a nonparametric regression, the denominator by a kernel density estimate.

An alternative method uses the local quasilikelihood structure of the problem. Since our algorithm is simply quasilikelihood with kernel weights, one can formally apply the usual expansions for quasilikelihood to get an estimated variance for $\widehat{\eta}(u)$. This argument is essentially the same as that of Hastie & Tibshirani (1990, pp. 156–157), and results in the following variance estimate:

$$\begin{aligned}
\widehat{\text{var}}\left\{\widehat{\eta}(u)\right\} &= \mathbf{e}_1^T \mathbf{G}(u)^{-1}\mathbf{H}(u)\mathbf{G}(u)^{-1}\mathbf{e}_1; \\
\mathbf{H}(u) &= \sum_{i=1}^{n} \begin{pmatrix} 1 \\ \widehat{\boldsymbol{\alpha}}^T\mathbf{X}_i - u \end{pmatrix} \begin{pmatrix} 1 \\ \widehat{\boldsymbol{\alpha}}^T\mathbf{X}_i - u \end{pmatrix}^T K_h^2(\widehat{\boldsymbol{\alpha}}^T\mathbf{X}_i - u)\rho_2\left\{\widehat{\eta}(\widehat{\boldsymbol{\alpha}}^T\mathbf{X}_i) + \widehat{\boldsymbol{\beta}}^T\mathbf{Z}_i\right\}; \\
\mathbf{G}(u) &= \sum_{i=1}^{n} \begin{pmatrix} 1 \\ \widehat{\boldsymbol{\alpha}}^T\mathbf{X}_i - u \end{pmatrix} \begin{pmatrix} 1 \\ \widehat{\boldsymbol{\alpha}}^T\mathbf{X}_i - u \end{pmatrix}^T K_h(\widehat{\boldsymbol{\alpha}}^T\mathbf{X}_i - u)\rho_2\left\{\widehat{\eta}(\widehat{\boldsymbol{\alpha}}^T\mathbf{X}_i) + \widehat{\boldsymbol{\beta}}^T\mathbf{Z}_i\right\},
\end{aligned}$$

where, as before, $\mathbf{e}_1$ denotes the unit column vector with 1 in the first position. When using local polynomials instead of local linear fits, simply add polynomial terms to the definitions of $\mathbf{H}$ and $\mathbf{G}$.

5.2 Estimation in Partially Linear Models

When X is scalar, so that $\boldsymbol{\alpha}_0 = 1$ is known, each of the terms in the limiting covariance matrices (19) and (20) can be estimated by nonparametric regression techniques. At least for the fully iterated estimator, this fairly tedious process to estimate (20) can be replaced by a simple consistent alternative based upon the usual expansions for quasilikelihood, which we explain after the introduction of further notation. The derivations are based upon the simple form (12), instead of taking derivatives in (11), since these are complex to compute.

Set $U_i = \boldsymbol{\alpha}_0^T\mathbf{X}_i = X_i$, $\widetilde{\mathbf{Z}} = (\mathbf{Z}_1, ..., \mathbf{Z}_n)^T$ and $\widetilde{\mathbf{A}}$ be diagonal with elements ρ_{2i}, where $\rho_{2i} \equiv \rho_2\left\{\eta(U_i) + \boldsymbol{\beta}^T\mathbf{Z}_i\right\}$. Further, set $\widetilde{\boldsymbol{\eta}} = \{\eta(U_1), \ldots, \eta(U_n)\}^T$ and $\widetilde{\boldsymbol{\epsilon}}$ to be the vector with ith element $\eta(U_i) + \boldsymbol{\beta}^T\mathbf{Z}_i$ $+(Y_i - \mu_i)/\left(\sigma^2 V_i \rho_{1i}\right)$, where $\mu_i = g^{-1}\left\{\eta(U_i) + \boldsymbol{\beta}^T\mathbf{Z}_i\right\}$ and $V_i = V(\mu_i)$. The smoothing matrix is the $n \times n$ matrix

$$\widetilde{\mathbf{S}} = \begin{bmatrix} \mathbf{e}_1^T\{\mathbf{L}(U_1)^T\widetilde{\mathbf{A}}\mathbf{K}(U_1)\mathbf{L}(U_1)\}^{-1}\mathbf{L}(U_1)^T\widetilde{\mathbf{A}}\mathbf{K}(U_1) \\ \vdots \\ \mathbf{e}_1^T\{\mathbf{L}(U_n)^T\widetilde{\mathbf{A}}\mathbf{K}(U_n)\mathbf{L}(U_n)\}^{-1}\mathbf{L}(U_n)^T\widetilde{\mathbf{A}}\mathbf{K}(U_n) \end{bmatrix}, \tag{24}$$

where $\mathbf{L}(u_0)$ is the $n \times 2$ matrix with first column all 1's and second column the terms $(U_i - u_0)/h$, and $\mathbf{K}(u_0)$ is diagonal with elements $K_h(U_i - u_0)$. It is convenient to use the terms $(U_i - u_0)/h$ rather than $(U_i - u_0)$ for theoretical purposes, although numerically they yield the same results.

Here is the motivation for $\widetilde{\mathbf{S}}$. For fixed $\boldsymbol{\beta}$ and u_0, note that the intercept $a(u_0)$ and h times the slope

$b(u_0)$ from the local quasilikelihood regression are the iterative solutions to the equations

$$\begin{aligned}\begin{bmatrix} a(u_0) \\ hb(u_0) \end{bmatrix} &= \left\{\sum_{k=1}^{n} \mathbf{L}_k(u_0)\mathbf{L}_k(u_0)^T K_h(U_k-u_0)A_k(u_0)\right\}^{-1} \\ &\quad \times \sum_{k=1}^{n} \mathbf{L}_k(u_0)K_h(U_k-u_0)A_k(u_0)\left\{a(u_0)+b(u_0)(U_i-u_0)+(Y_i-\mu_i)/\left(\sigma^2 V_i\rho_{1i}\right)\right\},\end{aligned} \tag{25}$$

where $\mathbf{L}_k(u_0)=\{1,(U_k-u_0)/h\}^T$ and $A_k(u_0)=\rho_2\left\{\eta(u_0)+\boldsymbol{\beta}^T\mathbf{Z}_k\right\}$. Setting $u_0=U_i$ for $i=1,...,n$ and multiplying both sides of (25) by $\mathbf{e}_1^T$ yields (24).

The following argument has similarities with equation (6.22) of Hastie & Tibshirani (1990, p. 154). Because of the local nature of the fit, the term $b(u_0)(U_i-u_0)$ in the last part of (25) can be ignored asymptotically. This means that the local quasilikelihood algorithm is asymptotically equivalent to solving in $\boldsymbol{\beta}$ and η the equations

$$\begin{aligned}\beta &= \left(\widetilde{\mathbf{Z}}^T\widetilde{\mathbf{A}}\widetilde{\mathbf{Z}}\right)^{-1}\widetilde{\mathbf{Z}}^T\widetilde{\mathbf{A}}\left(\widetilde{\boldsymbol{\epsilon}}-\widetilde{\boldsymbol{\eta}}\right); \\ \widetilde{\eta} &= \widetilde{\mathbf{S}}\left(\widetilde{\boldsymbol{\epsilon}}-\widetilde{\mathbf{Z}}\beta\right),\end{aligned}$$

This means that the estimate of $\boldsymbol{\beta}_0$ is asymptotically equivalent to solving $\boldsymbol{\beta}=\widetilde{\mathbf{H}}_1\widetilde{\boldsymbol{\epsilon}}$, where

$$\widetilde{\mathbf{H}}_1=\left\{\widetilde{\mathbf{Z}}^T\widetilde{\mathbf{A}}(\mathbf{I}-\widetilde{\mathbf{S}})\widetilde{\mathbf{Z}}\right\}^{-1}\widetilde{\mathbf{Z}}^T\widetilde{\mathbf{A}}(\mathbf{I}-\widetilde{\mathbf{S}}).$$

Since $\widetilde{\boldsymbol{\epsilon}}$ has covariance matrix $\widetilde{\mathbf{A}}^{-1}$, an approximate covariance matrix for $\widehat{\beta}$ is $\widetilde{\mathbf{H}}_1\widetilde{\mathbf{A}}^{-1}\widetilde{\mathbf{H}}_1^T$. One can show that this estimate yields asymptotically consistent standard errors for $\widehat{\boldsymbol{\beta}}$.

For local polynomial regression, the only changes necessary are the obvious modifications of $\mathbf{L}(u_0)$ and $\mathbf{L}_k(u_0)$.

5.3 Estimation in General Models

When $\boldsymbol{\alpha}_0$ is unknown, there are again two strategies. Nonparametric regression techniques can be used to estimate the terms in (22) and (23). We again develop a simpler consistent estimate of (23) for the fully iterated estimate. We build upon the notation in Section 5.2.

Let $\widetilde{\mathbf{Q}}$ be the $n\times p$ matrix with ith row given as $\eta'(U_i)\mathbf{X}_i^T$, and let $\widetilde{\mathbf{R}}=(\widetilde{\mathbf{Q}},\widetilde{\mathbf{Z}})$. Let

$$\mathbf{P}_{\boldsymbol{\alpha}}^*=\begin{bmatrix}\mathbf{I}-\boldsymbol{\alpha}\boldsymbol{\alpha}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\end{bmatrix}$$

and let $\widetilde{\boldsymbol{\epsilon}}$ to be the vector with ith element $\eta(U_i)+\eta'(U_i)(\boldsymbol{\alpha}^T\mathbf{X}_i)+(\boldsymbol{\beta}^T\mathbf{Z}_i)+(Y_i-\mu_i)/\left\{\sigma^2 V_i\rho_{1i}\right\}$. Remembering that we must have $||\boldsymbol{\alpha}||=1$ for identifiability, note that we find $(\boldsymbol{\alpha},\boldsymbol{\beta})$ by solving

$$\mathbf{0}=\widetilde{\mathbf{R}}^T\widetilde{\mathbf{A}}(\widetilde{\boldsymbol{\epsilon}}-\widetilde{\boldsymbol{\eta}})-\widetilde{\mathbf{R}}^T\widetilde{\mathbf{A}}\widetilde{\mathbf{R}}\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{pmatrix}+\begin{pmatrix}\theta\boldsymbol{\alpha}\\ \mathbf{0}\end{pmatrix}$$

where θ is a Lagrange multiplier associated with the constraint $\boldsymbol{\alpha}^T\boldsymbol{\alpha} = 1$. Of course, the same argument in deleting a term explained following (25) is used here.

Multiplying both sides by $P^*_{\boldsymbol{\alpha}}$ and solving, we find that

$$\begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} = \left(\mathbf{P}^*_{\boldsymbol{\alpha}}\widetilde{\mathbf{R}}^T\widetilde{\mathbf{A}}\widetilde{\mathbf{R}}\right)^- \mathbf{P}^*_{\boldsymbol{\alpha}}\widetilde{\mathbf{R}}^T\widetilde{\mathbf{A}}(\widetilde{\boldsymbol{\epsilon}} - \widetilde{\boldsymbol{\eta}}).$$

Remembering that $\widetilde{\boldsymbol{\eta}} = \widetilde{\mathbf{S}}(\widetilde{\boldsymbol{\epsilon}} - \widetilde{\mathbf{Q}}\boldsymbol{\alpha} - \widetilde{\mathbf{Z}}\boldsymbol{\beta})$, we find after some algebra that

$$\begin{aligned} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} &= \widetilde{\mathbf{H}}_2\widetilde{\boldsymbol{\epsilon}} \\ \widetilde{\mathbf{H}}_2 &= \left\{\mathbf{P}^*_{\boldsymbol{\alpha}}\widetilde{\mathbf{R}}^T\widetilde{\mathbf{A}}(\mathbf{I} - \widetilde{\mathbf{S}})\widetilde{\mathbf{R}}\right\}^- \mathbf{P}^*_{\boldsymbol{\alpha}}\widetilde{\mathbf{R}}^T\widetilde{\mathbf{A}}(\mathbf{I} - \widetilde{\mathbf{S}}). \end{aligned}$$

The estimated (and consistent) covariance matrix is $\widetilde{\mathbf{H}}_2\widetilde{\mathbf{A}}^{-1}\widetilde{\mathbf{H}}_2^T$.

6 IMPLEMENTATION

We will describe the implementation details in the case where the dimension of $\mathbf{Z}$ is $q = 1$. Extensions to $q > 1$ are straightforward.

For a given value of $\boldsymbol{\alpha}$ let $U_i = \boldsymbol{\alpha}^T\mathbf{X}_i$. In most cases, including all exponential family models, (8) and its derivatives with respect to a, b and β can be written as a linear combination of functions of the form

$$\hat{s}(u) = \sum_{i=1}^n \mathcal{S}(U_i - u, Z_i) \quad \text{and} \quad \hat{t}(u) = \sum_{i=1}^n \mathcal{T}(U_i - u, Z_i)Y_i$$

for bivariate functions $\mathcal{S}(\cdot,\cdot)$ and $\mathcal{T}(\cdot,\cdot)$ that depend upon Q, g and K_h. Exact computation of $\hat{s}(u)$ and $\hat{t}(u)$ over a fine grid of u values is expensive, so fast binned approximations (e.g. Härdle & Scott 1992, Fan & Marron 1994) are desirable.

Let $g_1^U < \cdots < g_{M_U}^U$ be an equally-spaced grid of width δ_U containing each of the U_i values. Similarly, let $g_1^Z < \ldots < g_{M_Z}^Z$ be a grid of width δ_Z containing the Z_i. With respect to the linear binning strategy (see e.g. Fan & Marron 1994) the grid counts at the grid point pair (g_ℓ^U, g_M^Z) are

$$\begin{aligned} c_{\ell m} &= \sum_{i=1}^n (1 - |(U_i - g_\ell^U)/\delta_U|)_+(1 - |(Z_i - g_m^Z)/\delta_Z|)_+ \\ \text{and} \quad d_{\ell m} &= \sum_{i=1}^n (1 - |(U_i - g_\ell^U)/\delta_U|)_+(1 - |(Z_i - g_m^Z)/\delta_Z|)_+Y_i \end{aligned} \tag{26}$$

where $x_+ = \max\{0, x\}$. The binned approximation to $\hat{s}(g_j^U)$ is

$$\widetilde{s}(g_j^U) = \sum_{\ell=1}^{M_U}\sum_{m=1}^{M_Z} \mathcal{S}(g_\ell^U - g_j^U, g_m^Z)c_{\ell m} = \sum_{\ell=1-j}^{M_U-j}\sum_{m=1}^{M_Z} \mathcal{S}(\delta_U\ell, g_m^Z)c_{\ell+j,m} \tag{27}$$

for $j = 1, \ldots, M_U$. The advantage of binning is that the first argument of $\mathcal{S}(\cdot,\cdot)$ takes no more than M_U distinct values. In fact, if a compactly supported kernel is used then this number will usually be much smaller than M_U. Furthermore, note that the computational burden is further reduced when Z is a discrete variable, as is often the case in applications. For example, if Z is a binary variable then we can take $M_Z = 2$ and $(g_1^Z, g_2^Z) = (0, 1)$. This results in a relatively low number of distinct pairs of arguments of $\mathcal{S}(\cdot, \cdot)$ for computing the $\widetilde{s}(g_j^U)$ values.

The binned approximation to $\hat{t}(g_j^U)$, denoted by $\widetilde{t}(g_j^U)$ is the same as (27) but with $c_{\ell m}$ replaced by $d_{\ell m}$.

Maximization of the resulting binned approximation to either (8) or (10) leads to a grid of approximations to $\hat{\eta}(g_j^U)$, $j = 1, \ldots, M_U$. Linear interpolation of this grid can be used to approximate values of $\hat{\eta}(\boldsymbol{\alpha}^T \mathbf{X}_i)$, and hence solve (9) or (11).

For grid sizes of about $M_U = 200$ the binning strategy described above leads to fast and accurate approximations of (8) and (10) and their derivatives for many situations. This permits use of the Newton-Raphson algorithm for fast maximization of each of these quantities. However, care needs to be taken with the choice of the starting values since the optimization may be with respect to several variables. To overcome this problem we suggest use of a more robust optimization strategy, such as that developed by Nelder & Mead (1965) (see e.g. Press, Flannery, Teukolsky & Vetterling 1992) to obtain a good starting value for the first grid point. We suggest that the first grid point be chosen near the center of the grid since boundary regions tend to be more problematic. Maximizing values for a particular grid point serve as very good starting values for an adjacent grid point.

In summary, our strategy for Steps 1 and 2 of the fully iterated algorithm (the one-step algorithm can be treated similarly) is:

- Set up fine equally-spaced grids g_ℓ^U, $\ell = 1, \ldots, M_U$ and g_m^Z, $m = 1, \ldots, M_Z$, covering the ranges of the U_i values and Z_i values respectively.
- Use (26) to obtain grid counts $c_{\ell m}$ and $d_{\ell m}$ for each grid point pair (g_ℓ^U, g_m^Z).
- Use 10–30 iterations of the Nelder-Mead algorithm to search for a maximizer of the binned approximation to (10) at grid point $g_{\lfloor M_U/2 \rfloor}^U$ (where $\lfloor \cdot \rfloor$ denotes integer part). Use the Newton-Raphson algorithm, with binned approximations to the derivatives of (10), to improve the maximizer.
- Use the Newton-Raphson algorithm to maximize (10) for each of the other grid points, with maximizers from previous adjacent grid points used as starting values, to give approximations $\hat{\eta}(g_\ell^U)$, $\ell = 1, \ldots, M_U$.
- Use 10–30 iterations of the Nelder-Mead algorithm followed by the Newton-Raphson algorithm to maximize (11) with linear interpolation of the $\hat{\eta}(g_\ell^U)$ used to approximate the $\hat{\eta}(\boldsymbol{\alpha}^T \mathbf{X}_i)$.

Binning methods can also be used for fast computation of the standard error estimates. Details of such calculations are given in Turlach & Wand (1995).

7 EXAMPLE

The Framingham data were described in Section 1.1. For these data, Y corresponds to incidence of CHD and Z corresponds to smoking status. In the discussion below we will use `disease` and `smoker` to denote these variables. As covariates we used

$$\begin{aligned} X_1 &= \texttt{age} = \text{age of patient;} \\ X_2 &= \texttt{trblood} = \log(\text{average of two SPB values} - 25); \\ X_3 &= \texttt{logchol} = \log(\text{cholesterol level}), \end{aligned}$$

with each variable scaled to lie between 0 and 1. To avoid problems with sparse data near the boundaries, after some experimentation only those data with a single-index value in range $[0.4, 1.2]$ were used for curve estimation. This excluded about 45 of the 1615 observations. We applied our methodology to the model:

$$\begin{aligned} &\text{logit}\{P(\texttt{disease} = 1 | \texttt{age}, \texttt{trblood}, \texttt{logchol}, \texttt{smoker})\} \\ &\qquad = \eta_0\{\alpha_{01}(\texttt{age}) + \alpha_{02}(\texttt{trblood}) + \alpha_{03}(\texttt{logchol})\} + \beta_0(\texttt{smoker}). \end{aligned}$$

We used the bandwidth h_{opt} defined in (17), obtaining nearly identical results with or without the modification suggested in the discussion centering on (18). The results of our analysis and that of an ordinary logistic regression are displayed in Table 1. The ordinary logistic regression coefficients for `age`, `trblood` and `logchol` have been made comparable to the single-index analysis by making their Euclidean norm equal to 1.0, and their standard error estimates have been adjusted accordingly. The parameter estimates derived from ordinary logistic regression and by our method are remarkably similar.

Figure 1 shows the estimates of (a) η_0 and (b) the conditional probability of heart disease for both smokers and non-smokers. The striking feature of this figure is the curvature of $\widehat{\eta}$ when the single-index becomes greater than 0.8. This curvature appears to be a real feature of the data, and has been checked in two ways. First, we used the ordinary logistic regression estimates to define a single-index, and then to this index and the smoking indicator fit a generalized additive model (Hastie & Tibshirani, 1990) to the data. The resulting estimate also showed curvature, of the same form as displayed in Figure 1. We also fit an ordinary generalized additive model with nonparametric components in `age`, `trblood` and `logchol`, and a nonlinear structure with the "flatness" of Figure 1 was found for `age` and to a lesser extent for the other variables.

8 DISCUSSION

8.1 Interactions

Model (5) does not explicitly deal with interactions between $\mathbf{X}$ and $\mathbf{Z}$, for example of the form

$$g\{\mu(\mathbf{x},\mathbf{z})\} = \eta_{z_1}(\boldsymbol{\alpha}_0^T\mathbf{x}) + \boldsymbol{\beta}_0^T\mathbf{z}_2, \tag{28}$$

where $\mathbf{z} = (z_1, \mathbf{z}_2)$ with z_1 binary. However, our methods can be modified to handle (28). The local quasilikelihood (8) should be replaced by

$$\begin{aligned} &\sum_{i=1}^{n} Q\left[g^{-1}\left\{a_0 + b_0(\boldsymbol{\alpha}^T\mathbf{X}_i - u) + \boldsymbol{\beta}^T\mathbf{Z}_i\right\}, Y_i\right] K_{h_0}(\boldsymbol{\alpha}^T\mathbf{X}_i - u) I(Z_{1,i} = 0) \\ + &\sum_{i=1}^{n} Q\left[g^{-1}\left\{a_1 + b_1(\boldsymbol{\alpha}^T\mathbf{X}_i - u) + \boldsymbol{\beta}^T\mathbf{Z}_i\right\}, Y_i\right] K_{h_1}(\boldsymbol{\alpha}^T\mathbf{X}_i - u) I(Z_{1,i} = 1), \end{aligned}$$

where h_0 and h_1 are bandwidths for η_0 and η_1, respectively. The estimators for η_0 and η_1 are respectively $\hat{\eta}_0(u) = \hat{a}_0$ and $\hat{\eta}_1(u) = \hat{a}_1$. One can modify (9) analogously.

Model (5) also allows modeling interactions of the form

$$g\{\mu(\mathbf{x},\mathbf{z})\} = \eta_0\{\boldsymbol{\alpha}_0^T\mathbf{x} + (\mathbf{x}^T, \mathbf{z}^T)\Lambda(\mathbf{x}^T, \mathbf{z}^T)^T\} + \boldsymbol{\beta}_0^T\mathbf{z}$$

where Λ is the parameter matrix for interactions. This model is included in (5) by forming a new and longer X-vector. The effective number of parameters is about $\ell_n + (p+q-1)(p+q+2)/2$, where ℓ_n is the effective numbers used for estimating $\eta_0(\cdot)$. One can also incorporate partial interaction terms in (5), which would reduce the number of effective parameters.

8.2 Other Models

A much different approach to modeling (and coping with the "curse of dimensionality") are the generalized additive models (GAM), see Hastie & Tibshirani (1990). These models replace the nonparametric component of (5) by a sum of nonparametric functions over the components of $\mathbf{X}$. When they adequately fit the data, the generalized partially linear single-index models (GPLSIM) (5) are attractive because of their model parsimony, and their ease of interpretation. For example, if on average each nonparametric component effectively uses ℓ_n (say) parameters, then the total number of parameters for GAM is about $p\ell_n + q$ for p covariates $\mathbf{X}$. In contrast, the GPLSIM use effectively $(\ell_n + p + q)$ parameters, which, depending on ℓ_n can be a much smaller number. For dedicated GAM users who wish to fit models of the form (5), the parameter $\boldsymbol{\alpha}_0$ can be estimated using our techniques, and then GAM can be applied to $\mathbf{Z}$ and $\widehat{\boldsymbol{\alpha}}^T\mathbf{X}$. In this context, one can think of our techniques as providing a preliminary dimension reduction. Clearly, an important issue for future work is to test for model misspecification of the GPLSIM against a richer class of models.

There are also various schools of thought about the need to employ parsimonious parametric models (see Royston & Altman, 1994, and the discussions therein). The GPLSIM falls somewhere between the fully parametric flexible models of Royston and Altman (1994) and the almost fully nonparametric models of Hastie and Tibshirani (1990).

REFERENCES

Chen, H. (1988). Convergence rates for parametric components in a partly linear model. *Annals of Statistics*, **16**, 136–146.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, **87**, 829–836.

Cuzick, J. (1992). Semiparametric additive regression. *Journal of the Royal Statistical Society, Series B*, **54**, 831–843.

Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, **87**, 998–1004.

Fan, J. , Heckman, N. E. & Wand, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasilikelihood functions. *Journal of the American Statistical Association*, to appear.

Fan, J. & Marron, J. S. (1994). Fast implementations of nonparametric curve estimators. *Journal of Computational and Graphical Statistics*, **3**, 35–56.

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

Härdle, W. & Scott, D. W. (1992). Smoothing by weighted averaging of rounded points. *Comput. Statist.*, **7**, 97–128.

Hastie, T. & Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association*, **82**, 371–386.

Hastie, T.J. & Tibshirani, R. (1990). *Generalized Additive Models.* Chapman & Hall, London.

Heckman, N. (1986). Spline smoothing in a partly linear model. *Journal of the Royal Statistical Society, Series A*, **48**, 244–248.

Hunsberger, S. (1994). Semiparametric regression in likelihood-based models. *Journal of the American Statistical Association*, **89**, 1354–1365.

Kannel, W. B., Neaton, J. D., Wentworth, D., Thomas, H. E., Stamler, J., Hulley, S. B. & Kjelsberg, M. O. (1986). Overall and coronary heart disease mortality rates in relation to major risk factors in 325,348 men screened for MRFIT. *American Heart Journal*, 112, 825–836.

Mack, Y. P. & Silverman, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, **61**, 405–415.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models.* Second Edition. Chapman & Hall, London.

Nelder, J. A. & Mead, R. (1965). A simplex method for function minimization. *Comput. Journal*, **7**, 308–313.

Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, **135**, 370–384.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory.* **7**, 186–199.

Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. (1988). *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, Cambridge.

Royston , P. & Altman, D.G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Appl. Statist.*, **43**, 429–467.

Ruppert, D., Sheather, S. J. & Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. Preprint.

Severini, T. A. & Staniswalis, J. G. (1994). Quasilikelihood estimation in semiparametric models. *Journal of the American Statistical Association*, **89**, 501–511.

Severini, T. A. & Wong, W. H. (1992). Profiule likelihood and conditionally parametric models. *Annals of Statistics*, 20, 1768–1802.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society, Series B*, **50**, 413–436.

Tsybakov, A. B. (1986). Robust reconstruction of functions by the local-approximation method. *Problems of Information Transmission*, **22**, 133–146.

Turlach, B. A. & Wand, M. P. (1995). Fast computation of auxiliary quantities in local polynomial regression. Unpublished manuscript.

Wahba, G. (1984). Partial spline models for semiparametric estimation of functions of several variables. In *Statistical Analysis of Time Series*, Proceedings of the Japan U.S. Joint Seminar, Tokyo, 319–329.

Wedderburn, R. W. M. (1974). Quasilikelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika*, **61**, 439–447.

Weisberg, S. & Welsh, A. H. (1994). Estimating the Missing Link Function. *Annals of Statistics*, to appear.

9 APPENDIX

9.1 Proof of Theorem 1

Here we prove the results for the one-step estimators. The methods of technique are similar for the fully iterated estimators.

For simplicity of notation, in this appendix we will absorb σ^2 into $V(\cdot)$, so that the variance of Y given $(\mathbf{Z}, \mathbf{X})$ is $V\{\mu(\mathbf{Z}, \mathbf{X})\}$.

Denote $q_\ell(x, y) = (\partial^\ell/\partial x^\ell) Q\{g^{-1}(x), y\}$, $\ell = 1, 2, 3$. Then

$$q_1(x, y) = \{y - g^{-1}(x)\}\rho_1(x) \qquad \text{and} \qquad q_2(x, y) = \{y - g^{-1}(x)\}\rho_1'(x) - \rho_2(x), \tag{29}$$

where $\rho_\ell(t) = \left\{\frac{dg^{-1}(t)}{dt}\right\}^\ell / V\{g^{-1}(t)\}$ is introduced in the beginning of Section 3.

The following conditions will be needed:

Condition 1:

(i) the function $q_2(x, y) < 0$ for $x \in \mathbb{R}$ and y in the range of the response variable,

(ii) the marginal density of $\boldsymbol{\alpha}_0^T \mathbf{X}$ is positive and continuous at the point u,

(iii) the function $\eta_0''(\cdot)$ is continuous at the point u,

(iv) $g''(\cdot)$ and $V(\cdot)$ are continuous functions,

(v) the conditional moments $E\left[q_1^2\left\{\eta_0(\boldsymbol{\alpha}_0^T\mathbf{X}) + \boldsymbol{\beta}_0^T\mathbf{Z}, Y\right\}|U = t\right]$, $E\left[q_1^2\left\{\eta_0(\boldsymbol{\alpha}_0^T\mathbf{X}) + \boldsymbol{\beta}_0^T\mathbf{Z}, Y\right\}\mathbf{Z}|U = t\right]$ and $E\left[q_1^2\left\{\eta_0(\boldsymbol{\alpha}_0^T\mathbf{X}) + \boldsymbol{\beta}_0^T\mathbf{Z}, Y\right\}\mathbf{Z}\mathbf{Z}^T|U = t\right]$ are continuous in t at the point u. Moreover, $E\left[q_2^2\left\{\eta_0(\boldsymbol{\alpha}_0^T\mathbf{X}) + \boldsymbol{\beta}_0^T\mathbf{Z}, Y\right\}\right] < \infty$ and $E\left[q_1^{2+\delta}\left\{\eta_0(\boldsymbol{\alpha}_0^T\mathbf{X}) + \boldsymbol{\beta}_0^T\mathbf{Z}, Y\right\}\right] < \infty$, for some $\delta > 2$.

(vi) the kernel K is a symmetric density function with bounded support,

(vii) the random vector $\mathbf{Z}$ is assumed to have a bounded support,

Condition 1 (i) is imposed so that (8) is concave in the parameters, which ensures the uniqueness of the solution. For the generalized linear model (3) with the canonical link function, this condition is satisfied. Conditions 1 (vi) and 1 (vii) are imposed just for simplicity of the proofs. They can be weakened significantly at the expense of lengthier proofs.

Condition 2:

(i) the function $\eta_0''(\cdot)$ is continuous in $u \in D$, a compact set, which is the support of the random variable $\mathbf{X}$,

(ii) the density function of $\mathbf{X}$ has a continuous second derivative on the set D,

(iii) the functions $V''(\cdot)$ and $g'''(\cdot)$ are continuous,

(iv) the conditional moments $E\left[q_1^2(\eta_0(\boldsymbol{\alpha}_0^T\mathbf{X}) + \boldsymbol{\beta}_0^T\mathbf{Z}, Y)|\mathbf{X} = u\right]$, $E\left[q_1^2(\eta_0(\boldsymbol{\alpha}_0^T\mathbf{X}) + \boldsymbol{\beta}_0^T\mathbf{Z}, Y)\mathbf{Z}|\mathbf{X} = u\right]$ and $E\left[q_1^2(\eta_0(\boldsymbol{\alpha}_0^T\mathbf{X}) + \boldsymbol{\beta}_0^T\mathbf{Z}, Y)\mathbf{Z}\mathbf{Z}^T|\mathbf{X} = u\right]$ are twice differentiable in $u \in D$,

Conditions 2 (ii) and 2 (iv) can be weakened to the assumption that the involved functions are Hölder continuous of order strictly larger than 1/2. The proofs remain valid under these less stringent conditions.

Proof of Theorem 1: We will prove something stronger than Theorem 1, namely that it holds for $\boldsymbol{\alpha}$ satisfying $||\boldsymbol{\alpha} - \boldsymbol{\alpha}_0|| = o\{h^2 + (nh)^{-1/2}\}$. To obtain such a result, we need to assume Condition 1 holds for each given $\boldsymbol{\alpha}$ and the corresponding quantities as a function of $\boldsymbol{\alpha}$ is continuous at the point $\boldsymbol{\alpha}_0$. Recall that for a fixed u and a given $\boldsymbol{\alpha}$, the vector $(\widehat{a}, \widehat{b}, \widehat{\boldsymbol{\beta}})^T$ maximizes (8). Let $c_n = (nh)^{-1/2}$, $U_i = \boldsymbol{\alpha}^T\mathbf{X}_i$,

$$\mathbf{X}_i^* = \begin{pmatrix} 1 \\ (U_i - u)/h \\ \mathbf{Z}_i \end{pmatrix}, \qquad \widehat{\boldsymbol{\beta}}^* = \begin{pmatrix} c_n^{-1}\{\widehat{a} - \eta_0(u)\} \\ c_n^{-1}h\left\{\widehat{b} - \eta_0'(u)\right\} \\ c_n^{-1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \end{pmatrix},$$

and let $f^*(\cdot)$ denote the density function of $U_i = \boldsymbol{\alpha}^T \mathbf{X}_i$. Denote further $\bar{\eta}_i = \bar{\eta}_i(u) = \eta_0(u) + \boldsymbol{\beta}_0^T \mathbf{Z}_i + \eta_0'(u)(U_i - u)$. If $(\widehat{a}, \widehat{b}, \widehat{\boldsymbol{\beta}})^T$ maximizes (8) then $\widehat{\boldsymbol{\beta}}^*$ maximizes

$$\sum_{i=1}^n Q\left\{g^{-1}(c_n \boldsymbol{\beta}^{*T} \mathbf{X}_i^* + \bar{\eta}_i), Y_i\right\} K_h(U_i - u)$$

with respect to $\boldsymbol{\beta}^*$. Consider the normalized function

$$\ell_n(\boldsymbol{\beta}^*) = h \sum_{i=1}^n \left[Q\left\{g^{-1}(c_n \boldsymbol{\beta}^{*T} \mathbf{X}_i^* + \bar{\eta}_i), Y_i\right\} - Q\left\{g^{-1}(\bar{\eta}_i), Y_i\right\}\right] K_h(U_i - u),$$

which is maximized by $\widehat{\boldsymbol{\beta}}^*$. The concavity of the function $\ell_n(\boldsymbol{\beta}^*)$ is ensured by Condition 1 (i). We aim at applying the Convexity Lemma (see Pollard, 1991) to this concave function. By a Taylor expansion of the function $Q(g^{-1}(\cdot), Y_i)$ we obtain that

$$\ell_n(\boldsymbol{\beta}^*) = \mathbf{W}_n^T \boldsymbol{\beta}^* + \tfrac{1}{2} \boldsymbol{\beta}^{*T} \mathbf{A}_n \boldsymbol{\beta}^* \{1 + o_P(1)\}, \tag{30}$$

where

$$\mathbf{W}_n = h c_n \sum_{i=1}^n q_1(\bar{\eta}_i, Y_i) \mathbf{X}_i^* K_h(U_i - u)$$

and

$$\mathbf{A}_n = h c_n^2 \sum_{i=1}^n q_2(\bar{\eta}_i, Y_i) \mathbf{X}_i^* \mathbf{X}_i^{*T} K_h(U_i - u).$$

It will be shown that

$$\mathbf{A}_n = -f^*(u) E\left[\rho_2(\eta_0(U) + \boldsymbol{\beta}_0^T \mathbf{Z}) \mathbf{A}(\mathbf{Z}) | U = u\right] + o_P(1) \equiv -A + o_P(1), \tag{31}$$

where $\mathbf{A}(\mathbf{Z})$ is given in (39). Therefore, by (30),

$$\ell_n(\boldsymbol{\beta}^*) = \mathbf{W}_n^T \boldsymbol{\beta}^* - \tfrac{1}{2} \boldsymbol{\beta}^{*T} \mathbf{A} \boldsymbol{\beta}^* + o_P(1). \tag{32}$$

By applying the Convexity Lemma, we obtain that

$$\widehat{\boldsymbol{\beta}}^* = \mathbf{A}^{-1} \mathbf{W}_n + o_P(1). \tag{33}$$

Hence the asymptotic normality of $\widehat{\boldsymbol{\beta}}^*$ will follow from that of $\mathbf{W}_n$, which we will establish below. Since $\mathbf{W}_n$ is a sum of i.i.d. random vectors, we basically only need to compute the first two moments and check Liapounov's condition. From the definition of $\mathbf{W}_n$ in (30), the expression for $q_1(x, y)$ in (29) and using Taylor's expansion we get for the first moment, with $\bar{\eta} = \eta_0(u) + \boldsymbol{\beta}_0^T \mathbf{Z} + \eta_0'(u)(U - u)$,

$$\begin{aligned}
E\mathbf{W}_n &= c_n^{-1} E\left[E\left\{q_1(\bar{\eta}, Y) \mathbf{X}^* K_h(U - u) | \mathbf{X}, \mathbf{Z}\right\}\right] \\
&= c_n^{-1} E\left(\left[g^{-1}\left\{\eta_0(\boldsymbol{\alpha}_0^T \mathbf{X}) + \boldsymbol{\beta}_0^T \mathbf{Z}\right\} - g^{-1}(\bar{\eta})\right] \rho_1(\bar{\eta}) \mathbf{X}^* K_h(U - u)\right) \\
&= c_n^{-1} E\left[(g^{-1})'\left\{\eta_0(\boldsymbol{\alpha}^T \mathbf{X}) + \boldsymbol{\beta}_0^T \mathbf{Z}\right\}\left\{\eta_0(\boldsymbol{\alpha}_0^T \mathbf{X}) - \eta_0(\boldsymbol{\alpha}^T \mathbf{X}) + o(\|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\|)\right.\right. \\
&\qquad \left.\left. + \tfrac{1}{2} \eta_0''(u)(U - u)^2 + o(h^2)\right\} \rho_1(\bar{\eta}) \mathbf{X}^* K_h(U - u)\right] \\
&= c_n^{-1} \tfrac{1}{2} \eta_0''(u) E\left[(g^{-1})'\left\{\eta_0(\boldsymbol{\alpha}^T \mathbf{X}) + \boldsymbol{\beta}_0^T \mathbf{Z}\right\}(U - u)^2 \rho_1(\bar{\eta}) \mathbf{X}^* K_h(U - u)\right] \\
&\qquad + O(c_n^{-1} \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\|) + o(c_n^{-1} h^2) \\
&= c_n^{-1} \tfrac{1}{2} \eta_0''(u) h^2 f^*(u) E\left[\rho_2\left\{\eta_0(U) + \boldsymbol{\beta}_0^T \mathbf{Z}\right\}(\kappa_2, 0, \kappa_2 \mathbf{Z}^T)^T | U = u\right] \\
&\qquad + o(c_n^{-1} h^2) + O(c_n^{-1} \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\|).
\end{aligned} \tag{34}$$

For the variance of $\mathbf{W}_n$ we find

$$\begin{aligned}
\mathrm{var}(\mathbf{W}_n) &= h\ \mathrm{var}\left\{q_1(\bar{\eta}, Y)\mathbf{X}^* K_h(U-u)\right\} \\
&= hE\left\{q_1^2(\bar{\eta}, Y)\mathbf{X}^*\mathbf{X}^{*T} K_h^2(U-u)\right\} + O\left(\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_0\|^2 + h^4\right) \\
&= f^*(u)E\left[q_1^2(\eta_0(U) + \boldsymbol{\beta}_0^T\mathbf{Z}, Y)B(\mathbf{Z})|U=u\right] + o(1) \\
&= f^*(u)E\left[\rho_2\left\{\eta_0(U) + \boldsymbol{\beta}_0^T\mathbf{Z}\right\} B(\mathbf{Z})|U=u\right] + o(1) \\
&\equiv \mathbf{B} + o(1),
\end{aligned} \tag{35}$$

where

$$\mathbf{B}(\mathbf{Z}) = \begin{pmatrix} \nu_0 & 0 & \nu_0\mathbf{Z}^T \\ 0 & \nu_2 & 0 \\ \nu_0\mathbf{Z} & 0 & \nu_0\mathbf{Z}\mathbf{Z}^T \end{pmatrix}. \tag{36}$$

Hence expressions for the first two moments of $\mathbf{W}_n$ are derived. Using similar arguments as before and Condition 1 (v), it can be shown that Liapounov's condition is satisfied and hence

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}^* - c_n^{-1}\tfrac{1}{2}\mathbf{A}^{-1}\eta_0''(u)h^2 f^*(u)E\left[\rho_2\left\{\eta_0(U) + \boldsymbol{\beta}_0^T\mathbf{Z}\right\}(\kappa_2, 0, \kappa_2\mathbf{Z}^T)^T|U=u\right] \\
-c_n^{-1}\left\{O(\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_0\|) + o(h^2)\right\} \xrightarrow{D} N(0, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}).
\end{aligned}$$

This establishes Theorem 1.

We now verify (31). We write $(\mathbf{A}_n)_{ij} = (E\mathbf{A}_n)_{ij} + O_P\left(\{\mathrm{var}(\mathbf{A}_n)_{ij}\}^{1/2}\right)$ and deal with the first two moments of A_n. Using the expression for $q_2(x, y)$ given in (29) we find for the first moment

$$\begin{aligned}
E\mathbf{A}_n &= E\left[E\left\{q_2(\bar{\eta}, Y)\mathbf{X}^*\mathbf{X}^{*T}K_h(U-u)|\mathbf{X}, \mathbf{Z}\right\}\right] \\
&= E\left(\left[g^{-1}\left\{\eta_0(\boldsymbol{\alpha}_0^T\mathbf{X}) + \boldsymbol{\beta}_0^T\mathbf{Z}\right\} - g^{-1}(\bar{\eta})\right]\rho_1'(\bar{\eta})\mathbf{X}^*\mathbf{X}^{*T}K_h(U-u)\right) \\
&\quad -E\left\{\rho_2(\bar{\eta})\mathbf{X}^*\mathbf{X}^{*T}K_h(U-u)\right\}.
\end{aligned} \tag{37}$$

Now it is easily seen from the arguments in deriving (34) that the first term in (37) is of order $O(\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_0\|) + O(h^2)$. Hence from (37) we get

$$\begin{aligned}
E\mathbf{A}_n &= -E\left\{\rho_2(\bar{\eta})\mathbf{X}^*\mathbf{X}^{*T}K_h(U-u)\right\} + O(\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_0\|) + O(h^2) \\
&= -E\left(E\left[\rho_2\left\{\eta_0(u) + \boldsymbol{\beta}_0^T\mathbf{Z} + O(h)\right\}\mathbf{X}^*\mathbf{X}^{*T}|U\right]K_h(U-u)\right) + O(\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_0\|) + O(h^2) \\
&= -f^*(u)E\left[\rho_2\left\{\eta_0(U) + \boldsymbol{\beta}_0^T\mathbf{Z}\right\}\mathbf{A}(\mathbf{Z})|U=u\right] + o(1),
\end{aligned} \tag{38}$$

where

$$\mathbf{A}(\mathbf{Z}) = \begin{pmatrix} 1 & 0 & \mathbf{Z}^T \\ 0 & \kappa_2 & 0 \\ \mathbf{Z} & 0 & \mathbf{Z}\mathbf{Z}^T \end{pmatrix}. \tag{39}$$

Hence,

$$E\mathbf{A}_n = -\mathbf{A} + o(1). \tag{40}$$

Analogously we can show that

$$\mathrm{var}(\mathbf{A}_n)_{ij} = O(c_n^2). \tag{41}$$

Combining (40) and (41), we obtain (31). This completes the proof.

9.2 Proof of Theorem 2

We need the following two lemmas in order to prove Theorem 2.

Lemma A.1: Let C and D be respectively compact sets in $\mathbb{R}^d$ and $\mathbb{R}^p$ and $f(\mathbf{x}, \boldsymbol{\theta})$ is a continuous function in $\boldsymbol{\theta} \in C$ and $\mathbf{x} \in D$. Assume that $\widehat{\boldsymbol{\theta}}(\mathbf{x}) \in C$ is continuous in $\mathbf{x} \in D$, and is the unique maximizer of $f(\mathbf{x}, \boldsymbol{\theta})$. Let $\widehat{\boldsymbol{\theta}}_n(\mathbf{x}) \in C$ be a maximizer of $f_n(\mathbf{x}, \boldsymbol{\theta})$. If

$$\sup_{\theta \in C, \mathbf{x} \in D} |f_n(\mathbf{x}, \boldsymbol{\theta}) - f(\mathbf{x}, \boldsymbol{\theta})| \longrightarrow 0,$$

then

$$\sup_{\mathbf{x} \in D} |\widehat{\boldsymbol{\theta}}_n(\mathbf{x}) - \widehat{\boldsymbol{\theta}}(\mathbf{x})| \longrightarrow 0, \quad \text{as } n \to \infty. \tag{42}$$

Proof: If (42) is false, then there exists a sequence $\mathbf{x}_n \in D$ such that

$$\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n) - \widehat{\boldsymbol{\theta}}(\mathbf{x}_n) \longrightarrow \mathbf{c} \neq \mathbf{0}.$$

Since $\{\mathbf{x}_n\}$ is in the compact set D, there exists a subsequence $\{\mathbf{x}_{n_j}\}$ such that $\mathbf{x}_{n_j} \to \mathbf{x}_0 \in D$. Without loss of generality, we assume that $\mathbf{x}_n \to \mathbf{x}_0 \in D$. Let ε_n be a generic positive sequence tending to zero. Evidently,

$$\begin{aligned}
f\left\{\mathbf{x}_0, \widehat{\boldsymbol{\theta}}(\mathbf{x}_0) + c\right\} &\geq f\left\{\mathbf{x}_n, \widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n)\right\} - \varepsilon_n && \text{(continuity of } f\text{)} \\
&\geq f_n\left\{\mathbf{x}_n, \widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n)\right\} - \varepsilon_n && \text{(uniform convergence of } f_n\text{)} \\
&\geq f_n\left\{\mathbf{x}_n, \widehat{\boldsymbol{\theta}}(\mathbf{x}_0)\right\} - \varepsilon_n && (\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n) \text{ is a maximizer}) \\
&\geq f\left\{\mathbf{x}_n, \widehat{\boldsymbol{\theta}}(\mathbf{x}_0)\right\} - \varepsilon_n && \text{(uniform convergence of } f_n\text{)} \\
&\geq f\left\{x_0, \widehat{\boldsymbol{\theta}}(\mathbf{x}_0)\right\} - \varepsilon_n && \text{(continuity of } f\text{).}
\end{aligned}$$

Thus, $f\left\{\mathbf{x}_0, \widehat{\boldsymbol{\theta}}(\mathbf{x}_0) + c\right\} \geq f\left\{\mathbf{x}_0, \widehat{\boldsymbol{\theta}}(\mathbf{x}_0)\right\}$, which is in contradiction with the unique maximizer $\widehat{\boldsymbol{\theta}}(\mathbf{x}_0)$ of $f(\mathbf{x}_0, \boldsymbol{\theta})$. This proves the lemma.

Lemma A.2: Let $(\mathbf{X}_1, Y_1), \cdots, (\mathbf{X}_n, Y_n)$ be i.i.d. random vectors, where the Y_i's are scalar random variables. Assume further that $E|Y|^s < \infty$ and $\sup_{\mathbf{x}} \int |y|^s f(\mathbf{x}, y) dy < \infty$ where f denotes the joint density of $(\mathbf{X}, Y)$. Let K be a bounded positive function with a bounded support, satisfying a Lipschitz condition. Then,

$$\sup_{\mathbf{x} \in D} \left| n^{-1} \sum_{i=1}^{n} \{K_h(\mathbf{X}_i - \mathbf{x}) Y_i - E[K_h(\mathbf{X}_i - \mathbf{x}) Y_i]\} \right| = O_P\left[\{nh/\log(1/h)\}^{-1/2}\right],$$

provided that $n^{2\varepsilon - 1} h_n \to \infty$ for some $\varepsilon < 1 - s^{-1}$.

Proof: This follows immediately from the result obtained by Mack & Silverman (1982). In that paper a weaker condition on K is imposed.

Proof of Theorem 2: First of all, we note that under Condition 2, (32) holds uniformly in $u \in D$ since in the Taylor expansion arguments we only use the boundedness of random variables and the smoothness of the functions involved. Note that each element in $\mathbf{A}_n$ is a sum of i.i.d. random variables of kernel form, and hence, by Lemma A.2, it converges uniformly to its corresponding element in $\mathbf{A}$. Consequently, expression (32) holds uniformly in $u \in D$. By the Convexity Lemma, it also holds uniformly in $\boldsymbol{\beta}^* \in C$ and $u \in D$ for any compact set C. Lemma A.1 then yields

$$\sup_{u \in D} |\widehat{\boldsymbol{\beta}}^*(u) - \mathbf{A}^{-1} \mathbf{W}_n(u)| \xrightarrow{P} 0, \tag{43}$$

where $\widehat{\boldsymbol{\beta}}^*(u)$ and $\mathbf{W}_n(u)$ are defined in the proof of Theorem 1 except that we here stress the dependence on u. So, we have, by considering the first element of the vectors in (43),

$$\sup_{u\in D}\left|\widehat{\eta}(u)-\eta_0(u)-\frac{1}{nf(u)}\sum_{i=1}^n W_iK_h(X_i-u)\right|=o_P(c_n),$$

where $f(u)$ is the density of X_i and W_i is the first element of the vector $q_1(\bar{\eta}_i, Y_i)\Sigma^{-1}(u)(1,\mathbf{Z}_i^T)^T$, with $\bar{\eta}_i=\bar{\eta}_i(u)=\eta_0(u)+\boldsymbol{\beta}_0^T\mathbf{Z}_i+\eta_0'(u)(U_i-u)$. Moreover, we will show that the following stronger result holds:

$$\sup_{u\in D}\left|\widehat{\eta}(u)-\eta_0(u)-\frac{1}{nf(u)}\sum_{i=1}^n W_iK_h(X_i-u)\right|=O_P\left\{h^2c_n+c_n^2\log^{1/2}(1/h)\right\}. \tag{44}$$

Let $\widehat{\theta}=n^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)$. Denote by

$$\widehat{m}_i=\widehat{\eta}(X_i)+\boldsymbol{\beta}_0^T\mathbf{Z}_i, \qquad \text{and} \qquad m_i=\eta_0(X_i)+\boldsymbol{\beta}_0^T\mathbf{Z}_i.$$

Then, $\widehat{\boldsymbol{\theta}}$ maximizes

$$\ell_n(\boldsymbol{\theta})=\sum_{i=1}^n\left[Q\left\{g^{-1}(\widehat{m}_i+n^{-1/2}\boldsymbol{\theta}^T\mathbf{Z}_i),Y_i\right\}-Q\left\{g^{-1}(\widehat{m}_i),Y_i\right\}\right]. \tag{45}$$

By Taylor's expansion, we have

$$\ell_n(\boldsymbol{\theta})=n^{-1/2}\sum_{i=1}^n q_1(\widehat{m}_i,Y_i)\boldsymbol{\theta}^T\mathbf{Z}_i+\tfrac{1}{2}\boldsymbol{\theta}^T\mathbf{B}_n\boldsymbol{\theta}, \tag{46}$$

where

$$\mathbf{B}_n=\frac{1}{n}\sum_{i=1}^n\left[Y_i\rho_1'\left\{g^{-1}(\widehat{m}_i+\xi_{ni})\right\}-\rho_3\left\{g^{-1}(\widehat{m}_i+\xi_{ni}')\right\}\right]\mathbf{Z}_i\mathbf{Z}_i^T,$$

with ξ_{ni} and ξ_{ni}' between 0 and $n^{-1/2}\boldsymbol{\theta}^T\mathbf{Z}_i$, independent of Y_i, and with $\rho_3(x)=-g^{-1}(x)\rho_1'(x)-\rho_2(x)$.

It will be shown that

$$\mathbf{B}_n=-E\rho_2\left\{\eta_0(X)+\boldsymbol{\beta}_0^T\mathbf{Z}\right\}\mathbf{Z}\mathbf{Z}^T+o_P(1)\equiv-\mathbf{B}+o_P(1). \tag{47}$$

Using similar arguments as for obtaining (47), we get

$$\begin{aligned}n^{-1/2}\sum_{i=1}^n q_1(\widehat{m}_i,Y_i)\mathbf{Z}_i &= n^{-1/2}\sum_{i=1}^n q_1(m_i,Y_i)\mathbf{Z}_i\\ &\quad+n^{-1/2}\sum_{i=1}^n q_2(m_i,Y_i)\left\{\widehat{\eta}(X_i)-\eta_0(X_i)\right\}\mathbf{Z}_i+O_P(n^{1/2}\|\widehat{\eta}-\eta_0\|_\infty^2).\end{aligned}$$

By (44), the second term in the above expression can be expressed as

$$\begin{aligned}&n^{-3/2}\sum_{i=1}^n q_2(m_i,Y_i)f(X_i)^{-1}\sum_{j=1}^n W_jK_h(X_j-X_i)\mathbf{Z}_i+O_P\left\{n^{1/2}c_n^2\log^{1/2}(1/h)\right\}\\ &\equiv T_{n1}+O_P\left\{n^{1/2}c_n^2\log^{1/2}(1/h)\right\}.\end{aligned}$$

Now define $v_j=v(X_j,Y_j,\mathbf{Z}_j)$ as the first element of $q_1(m_j,Y_j)\boldsymbol{\Sigma}^{-1}(1,\mathbf{Z}_j^T)^T$. Using the definition of $\bar{\eta}_j(X_i)$, we obtain $\bar{\eta}_j(X_i)-m_j=O((X_j-X_i)^2)$ and therefore

$$\begin{aligned}T_{n1} &= n^{-3/2}\sum_{i=1}^n\sum_{j=1}^n q_2(m_i,Y_i)f(X_i)^{-1}v_jK_h(X_j-X_i)\mathbf{Z}_i+O_P(n^{1/2}h^2)\\ &\equiv T_{n2}+O_P(n^{1/2}h^2).\end{aligned}$$

It will be shown that

$$T_{n2} - T_{n3} \xrightarrow{P} 0, \tag{48}$$

where

$$T_{n3} = -n^{-1/2} \sum_{j=1}^{n} \boldsymbol{\gamma}(X_j) v_j$$

with

$$\boldsymbol{\gamma}(u) = E\left[\rho_2\left\{\eta_0(u) + \boldsymbol{\beta}_0^T \mathbf{Z}\right\} \mathbf{Z} | X = u\right].$$

Combining (45) — (48) we obtain that

$$\ell_n(\boldsymbol{\theta}) = n^{-1/2} \sum_{i=1}^{n} \Omega(X_i, Y_i, \mathbf{Z}_i) - \tfrac{1}{2}\boldsymbol{\theta}^T \mathbf{B} \boldsymbol{\theta} + o_P(1),$$

where

$$\Omega(X_i, Y_i, \mathbf{Z}_i) = q_1(m_i, Y_i)\mathbf{Z}_i - \boldsymbol{\gamma}(X_i) v_i.$$

By the Convexity Lemma we find that

$$\widehat{\boldsymbol{\theta}} = \mathbf{B}^{-1} n^{-1/2} \sum_{i=1}^{n} \Omega(X_i, Y_i, \mathbf{Z}_i) + o_P(1),$$

from which it follows that

$$n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{D} N(0, \mathbf{B}^{-1} \boldsymbol{\Sigma}_1 \mathbf{B}^{-1}).$$

This establishes the result in Theorem 2, and now it remains to prove (44), (47) and (48).
Proof of (44). Throughout this proof, terms of the form $\widehat{G}(u) = O_P(a_n)$ are always meant as $\sup_{u \in D} |\widehat{G}(u)| = O_P(a_n)$. Recall that $\boldsymbol{\alpha} = 1$ and $\widehat{\boldsymbol{\beta}}^*$ maximizes $\ell_n(\boldsymbol{\beta}^*)$ in the proof of Theorem 1, namely

$$\frac{\partial}{\partial \boldsymbol{\beta}^*} \ell_n(\boldsymbol{\beta}^*)|_{\boldsymbol{\beta}^* = \widehat{\boldsymbol{\beta}}(u)} = c_n h \sum_{i=1}^{n} q_1\left\{\bar{\eta}_i + c_n \widehat{\boldsymbol{\beta}}^*(u)^T \mathbf{X}_i^*, Y_i\right\} \mathbf{X}_i^* \mathbf{X}_i^{*T} K_h(X_i - u) \widehat{\boldsymbol{\beta}}^*(u) = 0.$$

By using (43) and a Taylor expansion, we have

$$\mathbf{W}_n(u) + \mathbf{A}_n(u)\widehat{\boldsymbol{\beta}}^*(u) + \frac{c_n^3 h}{2} \sum_{i=1}^{n} q_3(\bar{\eta}_i + \widehat{\xi}_i, Y_i) \mathbf{X}_i^* \left\{\widehat{\boldsymbol{\beta}}(u)^T \mathbf{X}_i^*\right\}^2 K_h(X_i - u) = 0, \tag{49}$$

where $\widehat{\xi}_i$ is between 0 and $c_n \widehat{\boldsymbol{\beta}}^{*T} \mathbf{X}_i^*(u)$, and with $\mathbf{W}_n(u)$ and $\mathbf{A}_n(u)$ as in (30). The last term in the above expression is of order $O_P(c_n \|\widehat{\boldsymbol{\beta}}(u)\|^2)$, by using the same arguments as in the proof of statement (47). Since each element in $\mathbf{A}_n(u)$ is of a kernel form, we can deduce from Lemma A.2 that

$$\mathbf{A}_n(u) = E\mathbf{A}_n(u) + O_P\left\{c_n \log^{1/2}(1/h)\right\} = -\mathbf{A}(u) + O_P\left\{h^2 + c_n \log^{1/2}(1/h)\right\},$$

where $\mathbf{A}(u)$ is given in (31) except that we express here its dependence on u. Consequently, by (49) we obtain that

$$\mathbf{W}_n(u) - \mathbf{A}(u)\widehat{\boldsymbol{\beta}}^*(u)\left[1 + O_P\left\{h^2 + c_n \log^{1/2}(1/h)\right\}\right] + O_P(c_n \|\widehat{\boldsymbol{\beta}}(u)\|^2) = 0.$$

Hence,

$$\widehat{\boldsymbol{\beta}}(u) = \mathbf{A}^{-1} \mathbf{W}_n(u) + O_P\left\{h^2 + c_n \log^{1/2}(1/h)\right\}.$$

By taking the first element of the above expression we obtain the result in (44).

Proof of (47). From the structure of $\mathbf{B}_n$, it can be seen that there exist large enough constants a and b such that

$$\|\mathbf{B}_n - \frac{1}{n}\sum_{i=1}^{n} q_2(m_i, Y_i)\mathbf{Z}_i\mathbf{Z}_i^T\| \le o_P\left\{\frac{1}{n}\sum_{i=1}^{n}(a|Y_i| + b)\right\} = o_P(1),$$

since X and $\mathbf{Z}$ are bounded random variables. Applying the law of large numbers we obtain that

$$\frac{1}{n}\sum_{i=1}^{n} q_2(m_i, Y_i)\mathbf{Z}_i\mathbf{Z}_i^T = Eq_2(m_1, Y_1)\mathbf{Z}_1\mathbf{Z}_1^T + o_P(1) = -E\rho_2\left\{\eta_0(X_1) + \boldsymbol{\beta}_0^T\mathbf{Z}_1\right\}\mathbf{Z}_1\mathbf{Z}_1^T + o_P(1).$$

This proves statement (47).

Proof of (48). Denoting

$$R(X_i, Y_i, \mathbf{Z}_i) = q_2(m_i, Y_i)f(X_i)^{-1}\mathbf{Z}_i$$

it is easy to see that for $i \neq j$

$$\begin{aligned} E\left\{R(X_i, Y_i, \mathbf{Z}_i)K_h(X_j - X_i)|X_j\right\} &= -E\left\{\rho_2(m_i)K_h(X_j - X_i)f(X_i)^{-1}\mathbf{Z}_i|X_j\right\} \\ &= -\boldsymbol{\gamma}(X_j) + O_P(h^2). \end{aligned} \tag{50}$$

Assume without loss of generality that $\mathbf{Z}_j$ is a scalar. Otherwise, we would treat each of its components separately. Evidently,

$$\begin{aligned} E(T_{n2} - T_{n3})^2 &= n^{-3}\sum_{i,j,k,\ell} E\left[v_j v_\ell \left\{R(X_i, Y_i, Z_i)K_h(X_j - X_i) + \boldsymbol{\gamma}(X_j)\right\}\right. \\ &\qquad \left.\left\{R(X_k, Y_k, Z_k)K_h(X_\ell - X_k) + \boldsymbol{\gamma}(X_\ell)\right\}\right] \\ &= n^{-3}\sum_{\text{at most two indices equal}} \quad + \quad O(n^{-1}). \end{aligned} \tag{51}$$

We now deal with the different cases of summands appearing in the above expression.

Case 1. All four indices are different. Then, all terms involved in the summation are zero, since $E(v_j|X_j, Y_j) = 0$.

Case 2. $i \neq k$, and precisely two indices are equal. All these cases are dealt with in a similar way. For example in the case $j = \ell$, conditioning on all random variables indexed by j and k and using (50) yield that all these summands are of order $O(h^2)$.

Case 3. $i = k$ and all other indices unequal. In this case

$$E\left\{R^2(X_i, Y_i, Z_i)K_h(X_j - X_i)K_h(X_\ell - X_i)|X_j, X_\ell\right\} = O(h),$$

by using the conditions on the kernel K. The other terms involved are dealt with as in the previous two cases. Thus the summands of this type are of order $O(h)$. This establishes (48).

9.3 Proof of Theorem 4

We only outline the key ideas of the proof. Details can be filled in with techniques similar to those used in the proofs of Theorems 1 and 2.

In our analysis, it is important to be careful to note that $\widehat{\eta}$ depends on $\boldsymbol{\alpha}$, so we will write this as $\widehat{\eta}(u, \boldsymbol{\alpha})$. Using the same argument as in Theorem 1, one can show that

$$\widehat{\eta}(u, \widehat{\boldsymbol{\alpha}}) - \widehat{\eta}(u, \boldsymbol{\alpha}_0) = -\eta_0'(u)\mathcal{G}(u)^T\left(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\right) + o_P(n^{-1/2}).$$

This means that

$$\widehat{\eta}(\widehat{\boldsymbol\alpha}^T\mathbf{X},\widehat{\boldsymbol\alpha})-\eta_0(\boldsymbol\alpha_0^T\mathbf{X})=\widehat{\eta}(\boldsymbol\alpha_0^T\mathbf{X},\boldsymbol\alpha_0)-\eta_0(\boldsymbol\alpha_0^T\mathbf{X})+\eta_0'(\boldsymbol\alpha_0^T\mathbf{X})\left\{\mathbf{X}-\mathcal{G}(\boldsymbol\alpha_0^T\mathbf{X})\right\}^T(\widehat{\boldsymbol\alpha}-\boldsymbol\alpha_0)+o_P(n^{-1/2}).$$

Using these facts, we now consider Theorem 4. For a given $\widehat{\eta}(\cdot)$ we have to solve the following equations (see (21))

$$\left\{\begin{array}{l} n^{-1}\sum_{i=1}^n q_1\left\{\widehat{\eta}(\widehat{\boldsymbol\alpha}^T\mathbf{X}_i)+\widehat{\boldsymbol\beta}^T\mathbf{Z}_i,Y_i\right\}\widehat{\eta}'(\widehat{\boldsymbol\alpha}^T\mathbf{X}_i)\mathbf{X}_i+\theta\widehat{\boldsymbol\alpha}=0,\\ n^{-1}\sum_{i=1}^n q_1\left\{\widehat{\eta}(\widehat{\boldsymbol\alpha}^T\mathbf{X}_i)+\widehat{\boldsymbol\beta}^T\mathbf{Z}_i,Y_i\right\}\mathbf{Z}_i=0,\end{array}\right. \tag{52}$$

where θ is a Lagrange multiplier associated with the constraint $\boldsymbol\alpha^T\boldsymbol\alpha=1$. Denote

$$D_n=n^{-1}\sum_{i=1}^n q_1\left\{\widehat{\eta}(\widehat{\boldsymbol\alpha}^T\mathbf{X}_i)+\widehat{\boldsymbol\beta}^T\mathbf{Z}_i,Y_i\right\}\widehat{\eta}'(\widehat{\boldsymbol\alpha}^T\mathbf{X}_i)\mathbf{X}_i.$$

With a Taylor expansion we get

$$\begin{aligned} D_n &= n^{-1}\sum_{i=1}^n\Big[q_1(m_i,Y_i)+q_2(m_i,Y_i)\left\{\widehat{\eta}(\widehat{\boldsymbol\alpha}^T\mathbf{X}_i,\widehat{\boldsymbol\alpha})-\eta_0(\boldsymbol\alpha_0^T\mathbf{X}_i)+\mathbf{Z}_i^T(\widehat{\boldsymbol\beta}-\boldsymbol\beta_0)\right\}\\ &\quad+O_P\left(\|\widehat{\eta}-\eta_0\|_\infty^2+\|\widehat{\boldsymbol\beta}-\boldsymbol\beta_0\|^2\right)\Big]\left\{\widehat{\eta}'(\boldsymbol\alpha_0^T\mathbf{X}_i,\boldsymbol\alpha_0)+\widehat{\eta}''(\xi_{ni})(\widehat{\boldsymbol\alpha}^T\mathbf{X}_i-\boldsymbol\alpha_0^T\mathbf{X}_i)\right\}\mathbf{X}_i,\end{aligned}$$

with ξ_{ni} between $\boldsymbol\alpha_0^T\mathbf{X}_i$ and $\widehat{\boldsymbol\alpha}^T\mathbf{X}_i$.

From now on we will drop the second argument to $\widehat{\eta}$, since it is $\boldsymbol\alpha_0$. We next find that

$$\begin{aligned} D_n &= n^{-1}\sum_{i=1}^n q_1(m_i,Y_i)\widehat{\eta}'(\boldsymbol\alpha_0^T\mathbf{X}_i)\mathbf{X}_i\\ &\quad+n^{-1}\sum_{i=1}^n q_2(m_i,Y_i)\left\{\widehat{\eta}(\boldsymbol\alpha_0^T\mathbf{X}_i)-\eta_0(\boldsymbol\alpha_0^T\mathbf{X}_i)\right\}\widehat{\eta}'(\boldsymbol\alpha_0^T\mathbf{X}_i)\mathbf{X}_i\\ &\quad+n^{-1}\sum_{i=1}^n\Big[q_2(m_i,Y_i)\left\{\widehat{\eta}'(\boldsymbol\alpha_0^T\mathbf{X}_i)+o_P(1)\right\}^2+q_1(m_i,Y_i)\widehat{\eta}''(\xi_{ni})\\ &\qquad+q_2(m_i,Y_i)\left\{\widehat{\eta}(\boldsymbol\alpha_0^T\mathbf{X}_i)-\eta_0(\boldsymbol\alpha_0^T\mathbf{X}_i)\right\}\widehat{\eta}''(\xi_{ni})\Big]\mathbf{X}_i\left\{\mathbf{X}_i-\mathcal{G}(\boldsymbol\alpha_0^T\mathbf{X}_i)\right\}^T(\widehat{\boldsymbol\alpha}-\boldsymbol\alpha_0)\\ &\quad+n^{-1}\sum_{i=1}^n q_2(m_i,Y_i)\widehat{\eta}'(\boldsymbol\alpha_0^T\mathbf{X}_i)\mathbf{X}_i(\widehat{\boldsymbol\beta}-\boldsymbol\beta_0)^T\mathbf{Z}_i\\ &\quad+O_P\left(\|\widehat{\eta}-\eta_0\|_\infty^2+\|\widehat{\boldsymbol\alpha}-\boldsymbol\alpha_0\|^2+\|\widehat{\boldsymbol\beta}-\boldsymbol\beta_0\|^2\right).\end{aligned} \tag{53}$$

For the first term we get

$$n^{-1}\sum_{i=1}^n q_1(m_i,Y_i)\eta_0'(\boldsymbol\alpha_0^T\mathbf{X}_i)\mathbf{X}_i+n^{-1}\sum_{i=1}^n q_1(m_i,Y_i)\left\{\widehat{\eta}'(\boldsymbol\alpha_0^T\mathbf{X}_i)-\eta_0'(\boldsymbol\alpha_0^T\mathbf{X}_i)\right\}\mathbf{X}_i.$$

Using a projection argument similar as in the derivation of (48) we can show that the second term in the above expression is negligible. We can deal with the second term in (53) in an analogous way, namely

$$\begin{aligned} &n^{-1}\sum_{i=1}^n q_2(m_i,Y_i)\left\{\widehat{\eta}(\boldsymbol\alpha_0^T\mathbf{X}_i)-\eta_0(\boldsymbol\alpha_0^T\mathbf{X}_i)\right\}\widehat{\eta}'(\boldsymbol\alpha_0^T\mathbf{X}_i)\mathbf{X}_i\\ &=n^{-1}\sum_{i=1}^n q_2(m_i,Y_i)\left\{\widehat{\eta}(\boldsymbol\alpha_0^T\mathbf{X}_i)-\eta_0(\boldsymbol\alpha_0^T\mathbf{X}_i)\right\}\eta_0'(\boldsymbol\alpha_0^T\mathbf{X}_i)\mathbf{X}_i+o_P(n^{-1/2})\\ &=-n^{-1}\sum_{i=1}^n\boldsymbol\gamma_1(\boldsymbol\alpha_0^T\mathbf{X}_i)v_i+o_P(n^{-1/2}),\end{aligned}$$

with $\boldsymbol{\gamma}_1(u) = E\left[\rho_2(\eta_0(u) + \boldsymbol{\beta}_0^T\mathbf{Z})\eta_0'(u)\mathbf{X}|\boldsymbol{\alpha}_0^T\mathbf{X} = u\right]$, and with $v_i = \mathbf{e}_1^T q_1(m_i, Y_i)\boldsymbol{\Sigma}^{-1}(1, \mathbf{Z}_i^T)^T$. Therefore, D_n can be written as

$$\begin{aligned}
D_n &= n^{-1}\sum_{i=1}^{n}\left\{q_1(m_i, Y_i)\eta_0'(\boldsymbol{\alpha}_0^T\mathbf{X}_i)\mathbf{X}_i - \boldsymbol{\gamma}_1(\boldsymbol{\alpha}_0^T\mathbf{X}_i)v_i\right\} \\
&\quad -\mathbf{C}_{11}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) - \mathbf{C}_{12}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\
&\quad +O_P\left(||\widehat{\eta} - \eta_0||_\infty^2 + ||\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0||^2 + ||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||^2\right) + o_P(n^{-1/2}), \qquad (54)
\end{aligned}$$

where

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix}.$$

We deal with the second equation in (52) in an analogous fashion and find that

$$\begin{aligned}
& n^{-1}\sum_{i=1}^{n}\left\{q_1(m_i, Y_i)\mathbf{Z}_i - \boldsymbol{\gamma}(\boldsymbol{\alpha}_0^T\mathbf{X}_i)v_i\right\} \\
& -\mathbf{C}_{21}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) - \mathbf{C}_{22}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\
& +O_P\left(||\widehat{\eta} - \eta_0||_\infty^2 + ||\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0||^2 + ||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||^2\right) + o_P(n^{-1/2}). \qquad (55)
\end{aligned}$$

We remove the Lagrangian by multiplying by

$$\mathbf{P}_{\widehat{\boldsymbol{\alpha}}} = \begin{bmatrix} \mathbf{I} - \widehat{\boldsymbol{\alpha}}\widehat{\boldsymbol{\alpha}}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$$

and noticing that $\widehat{\boldsymbol{\alpha}}^T\widehat{\boldsymbol{\alpha}} = 1$ and that $\mathbf{I} - \widehat{\boldsymbol{\alpha}}\widehat{\boldsymbol{\alpha}}^T = \mathbf{I} - \boldsymbol{\alpha}_0\boldsymbol{\alpha}_0^T + O_P(||\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0||)$. With this, from equations (54) and (55) it follows that

$$n^{1/2}\begin{pmatrix} \widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0 \\ \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \end{pmatrix} \xrightarrow{D} \text{Normal}\left[\mathbf{0}, (\mathbf{P}_{\boldsymbol{\alpha}}\mathbf{C})^{-}\mathbf{P}_{\boldsymbol{\alpha}}\boldsymbol{\Sigma}_2\mathbf{P}_{\boldsymbol{\alpha}}\left\{(\mathbf{P}_{\boldsymbol{\alpha}}\mathbf{C})^{-}\right\}^T\right].$$

This completes the proof.

9.4 Bandwidths of Order $n^{-1/5}$

Here we indicate that for the fully iterated estimate defined by (11), bandwidths of the order $h \sim n^{-1/5}$ yield the results of Theorems 3 and 5. We consider the latter case only. The calculations are similar to those of Härdle, et al. (1993), see also Severini & Staniswallis (1994) and Hunsberger (1994). Here we only give a very brief sketch of the details. Unlike in the more formal argument of Theorem 4, for the sake of convenience here we ignore the constraint that $||\boldsymbol{\alpha}|| = 1$.

For any $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, the limit of $\widehat{\eta}(\boldsymbol{\alpha}^T\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ is denoted by $\eta(\boldsymbol{\alpha}^T\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta})$. It can be shown that

$$\begin{aligned}
&(\partial/\partial\boldsymbol{\alpha})\eta(\boldsymbol{\alpha}^T\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta})|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_0, \boldsymbol{\beta}=\boldsymbol{\beta}_0} = \eta_0'(\boldsymbol{\alpha}^T\mathbf{X})(\mathbf{X} - R_x); \qquad (56) \\
&\mathbf{Z} + (\partial/\partial\boldsymbol{\beta})\eta(\boldsymbol{\alpha}^T\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta})|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_0, \boldsymbol{\beta}=\boldsymbol{\beta}_0} = \mathbf{Z} - R_z; \qquad (57) \\
&R_x = E\left[\rho_2\left\{\eta_0(\boldsymbol{\alpha}_0^T\mathbf{X}) + \boldsymbol{\beta}_0^T\mathbf{Z}\right\}\mathbf{X}|\boldsymbol{\alpha}_0^T\mathbf{X}\right] / E\left\{\rho_2(\cdot)|\boldsymbol{\alpha}_0^T\mathbf{X}\right\}; \\
&R_z = E\left\{\rho_2(\cdot)\mathbf{Z}|\boldsymbol{\alpha}_0^T\mathbf{X}\right\} / E\left\{\rho_2(\cdot)|\boldsymbol{\alpha}_0^T\mathbf{X}\right\}.
\end{aligned}$$

Equation (56) is the generalization of a result of Härdle, et al. (1993), and follows the same line of argument as the easier (57), which we now sketch. By definition, for any $(x, \boldsymbol{\beta})$, $\eta(\boldsymbol{\alpha}_0^T x, \boldsymbol{\alpha}_0, \boldsymbol{\beta})$ is the solution to

$$0 = E\left(\left[Y - \mu\left\{\eta(\boldsymbol{\alpha}_0^T x, \boldsymbol{\alpha}_0, \boldsymbol{\beta}) + \boldsymbol{\beta}^T\mathbf{Z}\right\}\right]\rho_1\left\{\eta(\boldsymbol{\alpha}_0^T x, \boldsymbol{\alpha}_0, \boldsymbol{\beta}) + \boldsymbol{\beta}^T\mathbf{Z}\right\}|U = \boldsymbol{\alpha}_0^T x\right),$$

where $U = \boldsymbol{\alpha}_0^T \mathbf{X}$ and $\mu(\cdot) = g^{-1}(\cdot)$. Differentiating with respect to $\boldsymbol{\beta}$ and evaluating at $\boldsymbol{\beta}_0$, we find that

$$\begin{aligned} 0 &= E\left(\left[Y - \mu\left\{\eta_0(\cdot) + \boldsymbol{\beta}_0^T \mathbf{Z}\right\}\right] \rho_1'\left\{\eta(\cdot) + \boldsymbol{\beta}_0^T \mathbf{Z}\right\} \{(\partial/\partial\boldsymbol{\beta})\eta(\cdot) + \mathbf{Z}\} | U = \boldsymbol{\alpha}_0^T x\right) \\ &\quad - E\left[\rho_2\left\{\eta(\cdot) + \boldsymbol{\beta}_0^T \mathbf{Z}\right\} \{(\partial/\partial\boldsymbol{\beta})\eta(\cdot) + \mathbf{Z}\} | U = \boldsymbol{\alpha}_0^T x\right]. \end{aligned}$$

The first term on the right hand side equals zero, while the second is

$$-(\partial/\partial\boldsymbol{\beta})\eta(\cdot) E\left\{\rho_2(\cdot) | U = \boldsymbol{\alpha}_0^T x\right\} - E\left\{\rho_2(\cdot)\mathbf{Z} | U = \boldsymbol{\alpha}_0^T x\right\}.$$

This yields (57).

By differentiating (11) with respect to $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, the fully iterated estimator solves

$$0 = \sum_{i=1}^n (\mathbf{Y}_i - \widehat{\mu}_i)\widehat{\rho}_{1i} \begin{bmatrix} (\partial/\partial\boldsymbol{\alpha})\widehat{\eta}(\widehat{\boldsymbol{\alpha}}^T \mathbf{X}_i, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}) \\ \mathbf{Z}_i + (\partial/\partial\boldsymbol{\beta})\widehat{\eta}(\widehat{\boldsymbol{\alpha}}^T \mathbf{X}_i, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}) \end{bmatrix},$$

where $\widehat{\mu}_i = \mu\left\{\widehat{\eta}(\widehat{\boldsymbol{\alpha}}^T \mathbf{X}_i, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}) + \widehat{\boldsymbol{\beta}}^T \mathbf{Z}_i\right\}$ and similarly for $\widehat{\rho}_{1i}$. Let $\widehat{\mu}_{i*} = \mu\left\{\widehat{\eta}(\boldsymbol{\alpha}_0^T \mathbf{X}_i, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) + \boldsymbol{\beta}_0^T \mathbf{Z}_i\right\}$ and similarly for $\widehat{\rho}_{1i*}$. By a Taylor series expansion, for the obvious cross-products matrix A_n,

$$0 = n^{-1/2} \sum_{i=1}^n (\mathbf{Y}_i - \widehat{\mu}_{i*})\widehat{\rho}_{1i*} \begin{bmatrix} (\partial/\partial\boldsymbol{\alpha})\widehat{\eta}(\boldsymbol{\alpha}_0^T \mathbf{X}_i, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) \\ \mathbf{Z}_i + (\partial/\partial\boldsymbol{\beta})\widehat{\eta}(\boldsymbol{\alpha}_0^T \mathbf{X}_i, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) \end{bmatrix} - A_n n^{1/2} \begin{pmatrix} \widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0 \\ \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \end{pmatrix}. \tag{58}$$

It is easily shown that A_n converges to the matrix Q defined in Theorem 5. If we ignore the bias in $\widehat{\eta}(\boldsymbol{\alpha}_0^T \mathbf{X}, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$, it can also be shown by the same arguments as in Theorems 2-4 that the first term on the right hand side of (58) is asymptotically normally distributed with mean zero and covariance matrix also given by Q. Adjusting as in Theorem 4 for the constraint that $\|\boldsymbol{\alpha}\| = 1$ yields Theorem 5.

Hence, we next consider the bias. As noted in Section 3.3, the bias of $\widehat{\eta}(\boldsymbol{\alpha}_0^T \mathbf{X}, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ can be written as $h^2 \mathcal{G}(\boldsymbol{\alpha}_0^T \mathbf{X}) = h^2 \mathcal{G}(U)$. The first term on the right hand side of (58) can then be shown to equal an asymptotically mean zero, covariance matrix Q normal random variable plus the bias terms and terms of order $o_P(1)$, with the bias terms easily seen to be

$$n^{-1/2} h^2 \sum_{i=1}^n \mathcal{G}(\boldsymbol{\alpha}_0^T \mathbf{X}_i)\rho_2\left\{\eta_0(\boldsymbol{\alpha}_0^T \mathbf{X}) + \boldsymbol{\beta}_0^T \mathbf{Z}\right\} \begin{bmatrix} (\partial/\partial\boldsymbol{\alpha})\eta(\boldsymbol{\alpha}_0^T \mathbf{X}_i, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) \\ \mathbf{Z}_i + (\partial/\partial\boldsymbol{\beta})\eta(\boldsymbol{\alpha}_0^T \mathbf{X}_i, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) \end{bmatrix} + O_P(n^{1/2} h^4). \tag{59}$$

By (56)–(57), the first term is of order $O_P(h^2)$. Thus, (59) is of order $O_P(h^2 + n^{1/2} h^4)$. The choice $h = O(n^{-1/5})$ is small enough to make the bias tend to zero.

	`age`	`trblood`	`logchol`	`smoker`
Ordinary logistic	0.43	0.57	0.69	0.57
std. err.	0.10	0.13	0.11	0.25
GPLSIM	0.37	0.65	0.66	0.59
std. err.	0.086	0.11	0.12	0.24

Table 1: Framingham Heart Study. "`trblood`" is transformed systolic blood pressure, "`logchol`" is the log of serum cholesterol and "`smoker`" is smoking status. The ordinary logistic coefficients for age, trblood and logchol have been normalized to have Euclidean norm equal to 1.0, and the standard errors have been adjusted appropriately.

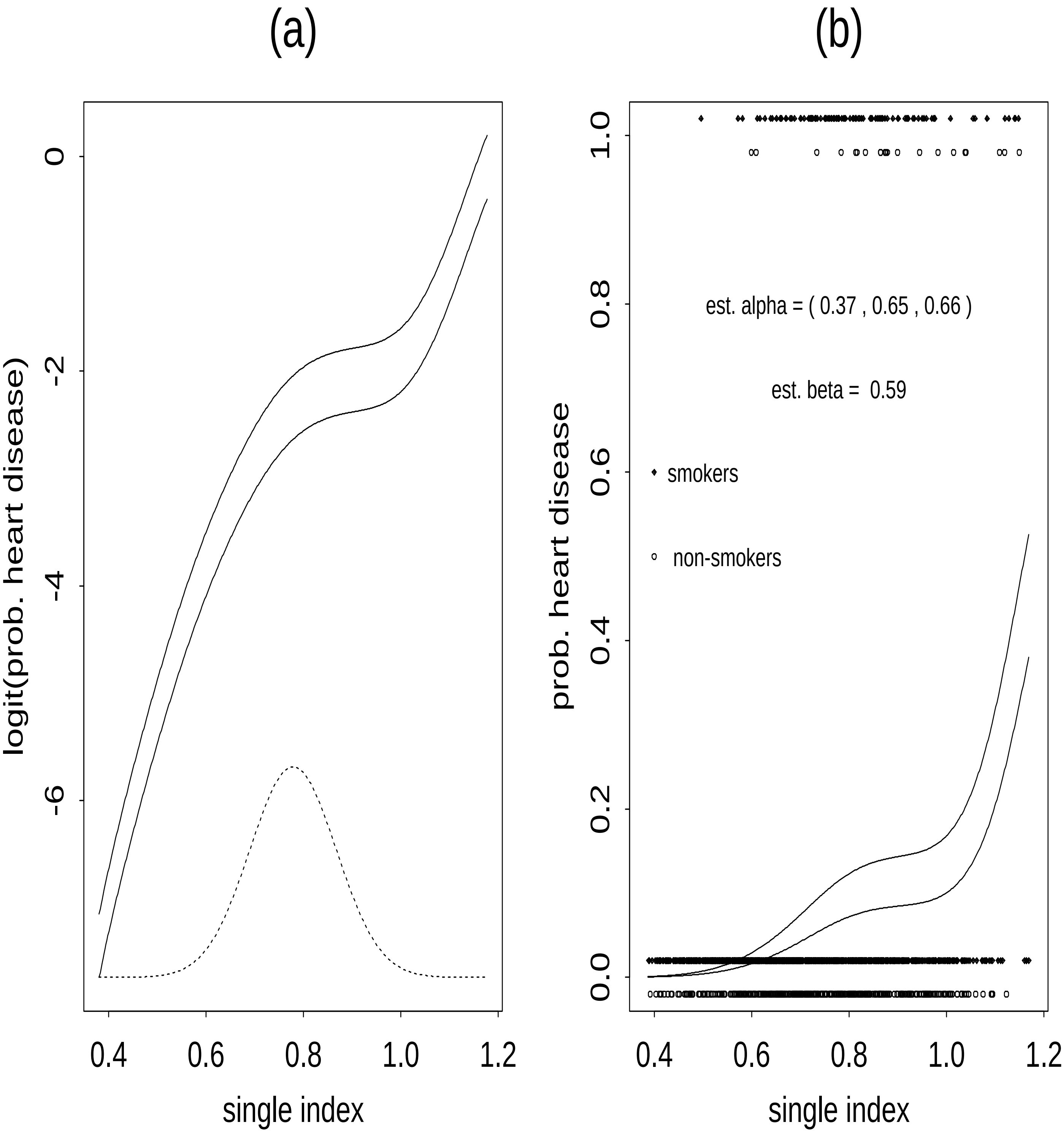

Figure 1: Curve estimates for the Framingham Heart Study data. (a) Solid curves correspond to estimates of logitP(heart disease) for smokers (upper curve) and non-smokers (lower curve) against the estimated single-index described in the text. The dotted curve is the kernel weight used in the local linear fitting process. (b) Estimates of P(heart disease) for smokers (upper curve) and non-smokers (lower curve) against the single index described in the text.